Advances in Computational Linguistics

Research in Computing Science

Series Editorial Board

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 130,** diciembre 2016. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. : 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 130**, December 2016. Printing 500. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research. Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Advances in Computational Linguistics

Sabino Miranda Jiménez (ed.)

SEP
SECRETARÍA DE
EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
“La Técnica al Servicio de la Patria”

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2016

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

Indexed in LATINDEX, DBLP and Periodica

Printing: 500

Printed in Mexico

ISSN 1870-4069

Editorial

This volume of the journal “Research in Computing Science” contains selected papers related to computational linguistics and its applications. The papers were carefully chosen by the editorial board on the basis of the at least two reviews by the members of the reviewing committee or additional reviewers. The reviewers took into account the originality, scientific contribution to the field, soundness and technical quality of the papers. It is worth noting that various papers for this special issue were rejected.

The volume contains 11 papers about various aspects of computational linguistics. Sentiment analysis or opinion mining is an important topic today, because of its potential applications in a wide variety of business. Three papers are about this topic dealing with different applications. The first one considers opinions in Twitter in order to analyze how verbs influence in gender opinions in various domains such as sports, movies, politics, entertainment, etc. The second one deals with aspect-based polarity detection using an n-gram approach, i.e., the positive or negative sentiment over different aspects such as food, service, ambience, among other, for restaurant and laptop domains in a well-known SemEval evaluation campaign. The third one considers gas station aspects to extract main features related to emotions (security, happiness, etc.), that make consumers prefer a gas station over others; the approach rely on machine learning techniques and statistical analysis to identify the relevant features such as customer's age and gender, average gasoline consumption, coffee shop, closeness, quick service, light place, etc.

Classification task is another fashion topic. The papers in this direction are from a point of view of data preprocessing facet to applications. In the case of data preprocessing, the authors show how sets of feature configurations influence in the performance of classification algorithms for plagiarism detection task, in the context of popular PAN competition. In the case of applications, two papers are about classifying an event in medical and personal assistant domain, respectively. In the medical domain, the authors show the performance of six classification algorithms (SVM, Naive Bayes, KNN, etc.) to classify two classes of cancer (colon or brain) based on textual medical records of patients; they use different text units (unigrams, bigrams, and 3-grams of characters) as textual representations, and TF-IDF as weighting scheme. In the personal assistant application, the authors analyze the user's email contents to identify events, and extract details of the event based on rule-based patterns to schedule automatically into the user’s agenda.

Also, developments on learning object, recommender systems, and information retrieval system are topics of current interest. Here, a metric for comparing designs of learning objects is proposed; it is based on information of a fixation measure using eye tracking data, and entropy-based approach. The recommender system gives an analysis of the best cloud services according to the user's needs. The cloud services are tested according to the user's selection criteria such as security, privacy, data storage, etc. In information retrieval systems, typically, a user's query (group of concepts) is searched in the document collections, here, it is proposed an expansion of user's queries in order to increase the accuracy of systems; the approach is based on ontologies of four domains (oil, tourism, e-learning, and artificial intelligence), and semantic relations of WordNet in order to expand the original queries.

Finally, written language is of current interest as well as spoken and sign language. According to the topic of spoken language, a broad survey of Spoken English learner corpora is given. The corpora described include recordings of speeches for various first languages (L1) such as Chinese, Spanish, Arabic, French, etc. For instance, there are texts spoken by English learners (L2, second language) whose first language is Spanish (L1). Also, a system for translating sing language to text is presented; in this work, the authors use image analysis and pattern recognition to interpret the sign and map it to the corresponding text according to the Mexican Sign Language.

I would like to thank Mexican Society for Artificial Intelligence (Sociedad Mexicana de Inteligencia Artificial), and MICAI 2016.

The entire submission, reviewing, and selection process, as well as preparation of the proceedings, were supported for free by the EasyChair system (www.easychair.org).

Sabino Miranda Jiménez
Guest Editor
Conacyt Catedra at Infotec, Mexico

December 2016

ISSN 1870-4069

Table of Contents

ISSN 1870-4069

Emotion Analysis in Gasoline Consumption in Mexico using Machine Learning

Dafne Rosso Pelayo, Joel Armando Colín Pacheco, Luis Miralles-Pechuán

Universidad Panamericana, Facultad de Ingeniería,
Mexico

drosso@up.edu.mx, jacolinp@yahoo.com, lmiralles@up.edu.mx

Abstract. Emotions and Sentiment Analysis has had an important role in increasing business benefits on commerce sector. Emotions Analysis as well as Sentiment Analysis is a common machine learning technique used to analyze opinions of people about certain company aspects such as products image, product consumption, marketing campaigns, client's preferences and social or political movements. The relevance of Emotions Analysis research lies in the enormous economic impact that it provides to enterprises. In this work, we present an emotion analysis to obtain the principal feature set related to the emotions that make consumers prefer a gas station over others. Our approach, to understand gasoline consumption behavior in Mexico, is based on machine learning and statistical analysis. We use a conventional statistical approach to analyze the characteristics of gas stations preferred by customers based on their emotions. Finally, supervised Machine Learning classification methods are applied in order to predict the probability that a gas station is selected on the basis of customers' emotions.

Keywords: Gasoline consumption, machine learning, emotions.

1 Introduction

Emotions are sensations generated by feelings and perceptions. Emotions are accompanied with thinking and actions. Emotions never exist alone; they are always accompanied by thoughts [1]. Emotions are the result of brain stimulation; they can be originated by a remembrance, an action, an observation, a comment or any other stimulus that impacts feelings or perceptions of people.

Emotions reflect an important impact in client’s decisions [2, 3, 4, 5, 6]. When emotions are positive, the likelihood of finding positive action from clients is higher; otherwise when the emotion is negative the likelihood of finding positive action is lower.

Antonio Damasio, neurologist and director of the Brain and Creativity Institute (BCI), formulated a theory about how emotions and their biological underpinnings are involved in decision-making. His main field of study is the neurological system based on memory, language, emotions and decision-making process. Damasio in [7] shows that "basic emotions like happiness, sadness, shame and empathy, are a set of complex chemical responses as neurological both forming a distinctive pattern" [8, 9].

In this work we present an emotion analysis to obtain the principal features and characteristics of a gas station that produce an emotion and make impact over the selection process of the clients.

To this end, we analyze customer's preferences on gas stations from two different approaches. First, a classical statistical analysis is developed in order to explore the collected information and to observe how the gas station selection process can be affected by perceptions. Perceptions are originated from customer experiences when they consume gasoline.

Secondly, we perform a supervised classification ML model to identify patterns that make people select a particular gas station based in their emotions. Emotions are generated when people consume the gas station's services. Several ML models were applied; their results and comparative accuracies are presented in results section.

The paper is organized in 6 sections. Section 1 is for this introduction. In section 2 we present related work about emotion analysis. Section 3 describes the conventions used in the research. In section 4 we describe acquisition and preprocessing of data; we include the experiments and results: we achieve a classical statistical analysis to model client's behavior based on clients' perception of gas stations, and we present a supervised ML approach to discover the emotions and the factors that most influence on the gas station selection process. In section 5 conclusions and future work are presented. Finally section 6 lists the references.

2 Related Works

There is a lot of research on how emotions affect customer preferences [10, 11]. However, few investigations focus on gasoline consumption.

What drive consumers to prefer a gas station over another? At first sight one may think that it is the price of gasoline at that station. In [4] Turrentine, Kurani and Heffer conducted interviews to people who had recently bought a vehicle, giving more attention to the context and validity of the information than to its statistical reliability.

They found that consumers do not budget, manage or track fuel costs. They also discovered that fuel economy decisions are based more by emotions than by critical analysis, and that are more influenced by social awareness than by its monetary value. Buyers of hybrid vehicles did not think of a payback when they made the purchase but they paid a lot of attention to fuel economy [4].

This social awareness is also related to social responsibility on environmental issues. In [5] it is concluded that consumer behavior regarding gasoline products is affected substantially within environmentally sensitive target groups. In their research they used a survey method to collect data regarding three gasoline brands (one green brand, one non-green brand and one partially green brand) and analyzed them with multiple regression analysis. Their results show a pattern that favors gasoline brands perceived as environmental friendly and that even people are willing to drive far in order to locate their favorite station. They did not find any gender consumer behavior difference.

Emotional experience with product's brands is also an important element that affects customer choice. Experience, if fully processed, is stored in different parts of our memory; including the emotion associated with it. In [6] Hansen, Christensen and Lundsteen state that when a memory is recalled, all of its components get together and the emotional association with the brand comes up too; these emotional responses are the frame of conscious cognitive process.

According to them, if emotional associations with a brand recalled in memory are understood then marketing communication that triggers emotional responses consistent with the brand should help to reinforce positive brand attitudes as well as purchase intentions. Toward that end a study was undertaken where they looked at 16 categories (gasoline included) covering 64 brands and then they recruited a random sample of consumers who answered a questionnaire. To measure the emotions associated with brands they developed a set of scales and then they used factor analysis to reduce them and make them reliable and easily used in advertising and brand research. This resulted in 2-factor solutions that reflect a strong positive – negative distinction. Scores were then calculated for each subject based upon the intensity of felt emotion (on a 6-point scale) and the factor loadings to produce a positive and negative score for each brand and category. Their results showed important insights into how people perceive brands and the emotions associated with that perception.

ML techniques are being widely applied to discover emotions people reflect in their text work, their speech and music, their facial expression or their body movement and so on. Those efforts are mainly focused on discovering what the subject felt or intended to transmit when creating his work [12].

Here we focus on finding and predicting the emotions that elicit the actual features of gas stations in Mexico that make a difference in gas station selection.

3 Conventions

The following conventions are used in our research. We can express the customer behavior in the following way: the independent object's features $\boldsymbol{x_i}$ derives in a particular emotion $\boldsymbol{e_i}$. Even more, the combination of dependent or independent features of objects $\boldsymbol{x_i}$ can lead to other particular emotions, represented by $\boldsymbol{e_j\ U \ldots U\ e_m}, \forall \boldsymbol{e_i} \in \boldsymbol{E}$ where $\boldsymbol{E}$ is the set of possible emotions.

In this way client's feelings and perception over any object $\boldsymbol{x_i}$ becomes an input to produce an emotion $\boldsymbol{e_i}$. The emotions $\boldsymbol{e_i}$ influence clients' actions and decisions. Emotions have been effectively used to analyze customers' behavior and to help business units to develop marketing strategies [5, 6].

Each gas station $\boldsymbol{G_i}$ possesses a feature set $\boldsymbol{X}$ of size $\boldsymbol{m}$, these features are independent, and each feature $\boldsymbol{x_i}$ has been rated for each client in an online questionnaire. We have defined a set of personal client's features $\boldsymbol{Y}$, where each $\boldsymbol{y_i} \in \boldsymbol{Y}$ represents a client feature i.e. $\boldsymbol{y_i}$ = {*age, gender, average gasoline consumption*, etc.}.

Each feature $\boldsymbol{x_i} \in$ X represents a feature that clients can get from a gas station $\boldsymbol{G_i}$ during the service time, i.e. $\boldsymbol{x_i}$ ={clean bathrooms; store; card of rewards, windshield wiper, etc.}.

The surveyed people rated the gas station feature $\boldsymbol{x_i}$ according to their preferences. These features $\boldsymbol{x}_{\mathrm{i}}$ were associated to a particular emotion $\boldsymbol{e}_{\mathbf{i}}$ in an unconsciously way to the responder.

4 Experiments and Results

4.1 Dataset Description

The dataset used in this research was obtained by conducting an online survey. The survey was answered only by Mexican residents that consume gasoline.

The survey was distributed over the internet and many people were invited to answer it through several social networks.

According to the "National Statistics and Geographic Institute" (INEGI), in 2014 Mexico had 35,353,077 vehicles. We represent this population by ***MV***.

Through our survey we obtained from ***MV*** a representative sample of the population ***S*** of size ***n,*** where ***S*** ∈ ***MV***, of enough size to obtain statistically significant results, we select a constant value for standard deviation ***σ,*** a 95% confidence level (***Z***=1.96), and an error ***e*** of 0.056%. The analysis of this article is based on the set ***S***. The sample size responding to a minimal selected 95% of confidence level was estimate according statistical principles with the following formula:

$$n = \frac{Z^2\sigma^2N}{e^2(N-1)+Z^2\sigma^2}. \tag{1}$$

For every client $\boldsymbol{t_j}$ that answered our survey, we obtain a set of personal features ***Y*** and a vector of ranked features ***R***, where **r** is a ranked feature, **r** ∈ **R**. And ***r(i,j)*** is the rating given from a client t_j to a gas station feature $\boldsymbol{x_i}$ with $\boldsymbol{x} \in \boldsymbol{X}$, where ***X*** represents the set of gas station's features evaluated. The ratings ***r(i,j)*** varies in an interval [1,10] where "1" is the minimum rating to evaluate a gas station's feature and "10" represents the maximum rating.

Personal features ***Y*** of set ***S*** were correlated to clients. This correlation analysis helps to understand gasoline consumption behavior and clients' preferences.

Segmentation and statistical analyses bring us as a result a set of segments and statistician's patterns about clients' consumption. In the set ***S,*** each gasoline feature $\boldsymbol{x_i}$ is related in a supervised way to a particular emotion $\boldsymbol{e}_{\mathbf{i}}$. In this way, our data set ***S*** can be used to analyze features $\boldsymbol{x_i}$, which had been rated by clients, and to discover the most relevant emotions that drive clients to make consumption in a specific gas station.

4.2 Statistical Analysis

Our statistical analysis involves an independent variable analysis of clients' personal characteristics over their consumption using ranked gasoline features $\boldsymbol{x_i}$**.** Table 1 represents an example of set $\boldsymbol{Y}$. This set includes clients' personal features, where each $\boldsymbol{y_i} \in \boldsymbol{Y}$ represents a client characteristic.

Table 1. Personal client features.

Client t_i	**Y_i**
y_1	Age
y_2	Gender
⋮	⋮
y_n	Average gasoline consumption

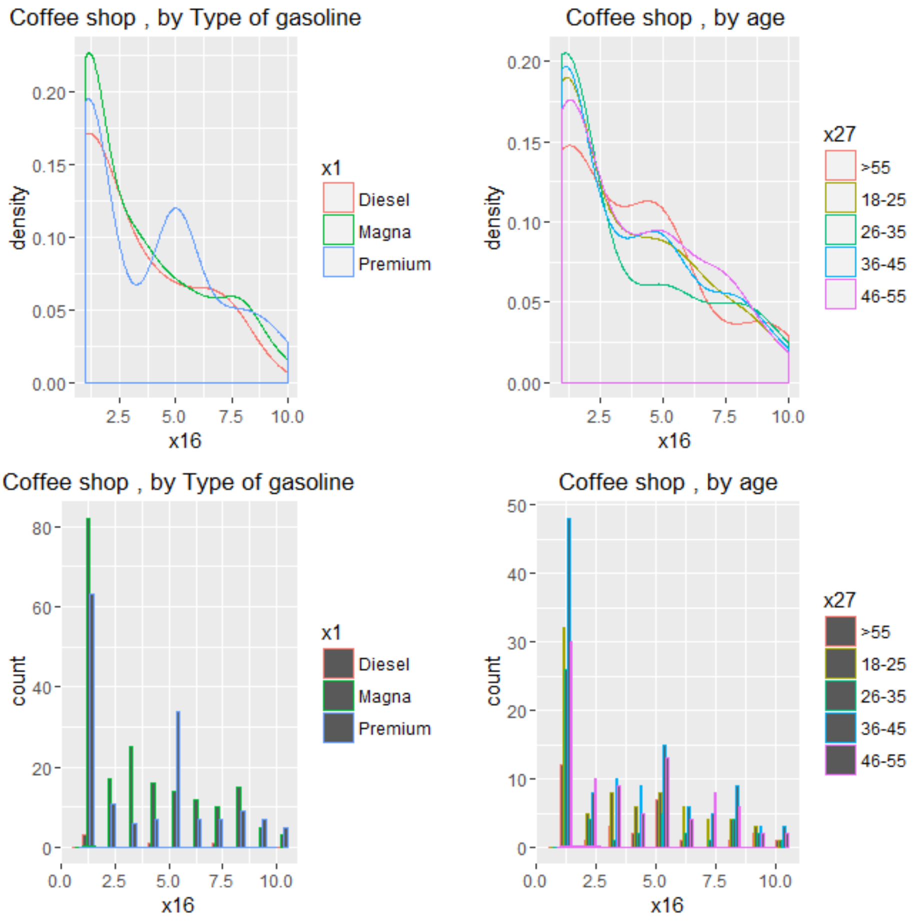


Fig 1. Distinct preferences distribution related to Coffee shops.

The statistical analysis was performed to every gasoline characteristic and personal feature as well as their combinations. Along this article we mention some of them. One of the first relevant discovered patterns is that local stores or coffee shops are not an important factor for gas station selection. Figure 1 represents the distribution of the

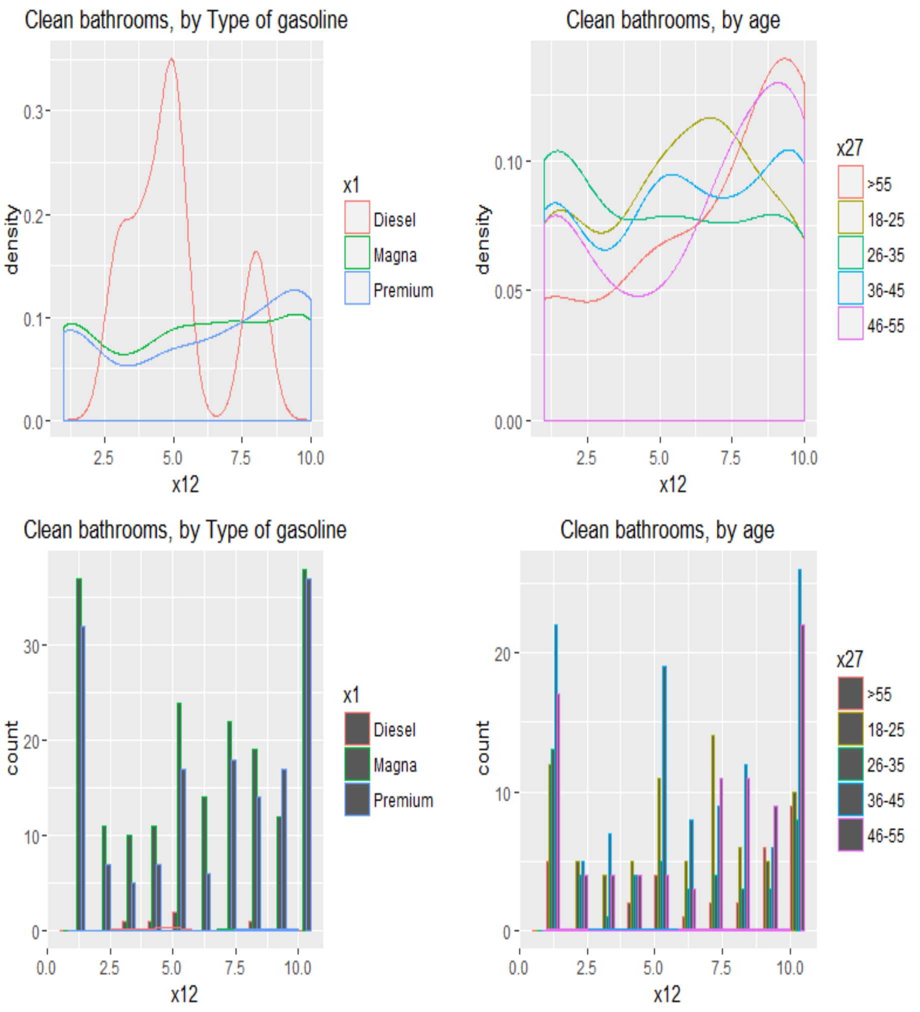


Fig 2. Distinct preferences distribution related to clean bathrooms.

different consumption preferences related to coffee shops.

This kind of analysis shows how clients' preference varies according to their profile. And more important, it identifies those gas station characteristics that are not relevant in any client segment or that are relevant only in a few. Figure 1 shows density distribution of 'coffee shop' feature. This distribution changes for different customer ages and the type of products consumed.

However, we can conclude that regardless clients' age, they do not select a particular gas station based on the 'coffee shop' feature. Statistical analysis of the gas station characteristics shows that clients prefer other kind of intrinsic characteristics related to gas station instead of additional services as it is shown in Figure 2.

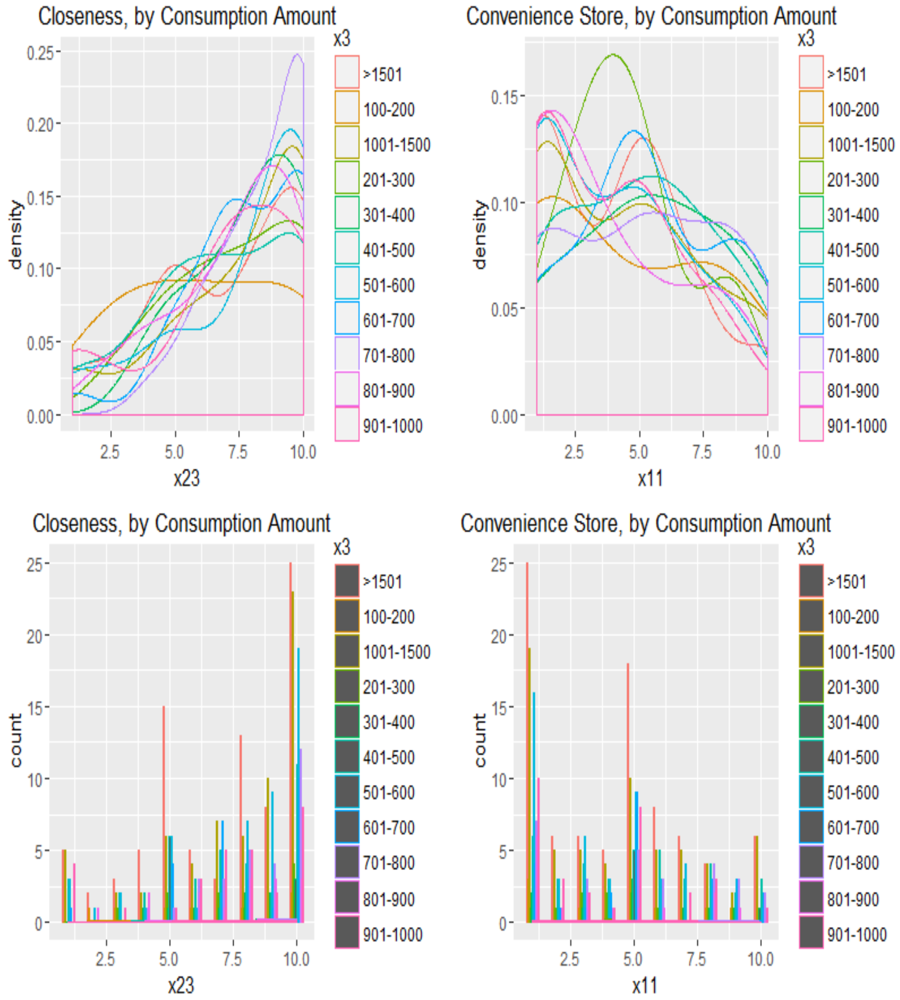


Fig 3. Distinct preferences distribution over gas stations characteristics.

In Figure 2, the analysis of the distribution of the 'Clean bathrooms' feature is very different from the distribution analysis of 'Coffee shop' feature as shown in Figure 1. Although this is not a defining feature to select a gas station, it is a very influential one. As we can see many customers prefer a station with clean bathrooms so the tendency is to select a gas station with this feature.

However, there are some variables that are critical in the selection of the gas station. Figure 3 shows the behavior of these detected variables related to the amount of consumption.

This statistical analysis shows that there are much more important features than others in the gas station selection process. In the second part of our analysis, the emotion variables added in association with the characteristics of gas stations will show not only the features that customers prefer when they select a gas station, but also a set of features that trigger one or more emotions on the client which influences the selection of a particular gas station.

4.3 Data Transformation

Data transformation is a fundamental stage in ML analysis. In this phase, we perform a transformation process over the data in order to be able to mine our data set ***S***. We alter data dimensionality in the following way, for every gasoline characteristic $\boldsymbol{x_i}$ we add an associated emotion $\boldsymbol{e_i}$, in such a way that personal characteristics are preserved.

The emotion $\boldsymbol{e_i}$, was mapped to characteristics $\boldsymbol{x}_{\mathrm{i}}$ based in studies that shows that emotions are produced by external stimulus and they produce an instantaneous impact causing an immediate response over people that they can't realize in a complete conscientious way about what is happening at that moment [8]. To distinct between positive and negative emotions we take into account that people exposed to constant external stimulus, incorporate in their neurological patterns those that are beneficial to their life and discard those that are not [10,13]. In our research, best ranking characteristics were taken as accepted for people and were mapped with positive emotions, in other way, bad ranking characteristic were associated with negative emotions, nevertheless all characteristics about gas station services focused to bring a commodity to clients like convenience stores or automatic cashiers were mapped to positive emotions.

Our supervised analysis also requires a class ***C*** to classify the items set into selected and not selected. Thus, we add a binary class called “Selection” with two possible values {selected, not selected}. We classified every characteristic in a supervised way based in its ranking, if it is higher than 5.0 then it is a characteristic that will most probably be accepted by clients; in other way it is rejected (not selected).

After the transformation process we call our new data set as ***S’***.

Table 2. Example of client Emotions in S’.

Client t_i	x_i	e_i
e1	Automatic charge	Security
e2	Rapid service	Happiness
⁞	⁞	⁞
en	Illuminated station	Security

It is important to mention that the mapping between emotion $\boldsymbol{e}_{\mathrm{i}}$ and characteristic $\boldsymbol{x_i}$ is constant and has a unique value thorough the analysis. The emotion $\boldsymbol{e_i}$ was assigned according to the type of characteristic, nevertheless the set of emotions of a single client has a particular combination and proportion of emotions and characteristics that make him to select or not select a specific gas station; thus an ML analysis can tell us with accuracy the probability of a gas station of being selected or not selected.

4.4 Machine Learning Model

The objective of supervised ML model is to predict the selection probability of a gasoline station based on the emotions triggered in clients through the gas station features. The contribution of our model is the novelty of associating the selection probability with the analysis of emotions rather than basing the analysis only on the gas stations feature set.

We accentuate the requirement of an ML model because it emphasizes prediction and can deal with the uncertainty founded in the missing data; it also provides the advantage that its predictions are based on historical data and will remain valid if the population increases, while statistical analysis emphasizes inference and the confidence intervals will not be valid under other parameters of the distribution.

The greater the probability of selection, the greater the acceptance of the gas station according to customer preferences and the greater the realization of future consumption.

From an economic point of view, this type of analysis can bring huge profits on future investments in the construction of new gas stations in Mexico. Building gas stations that trigger emotions in customers that ensure greater acceptance will get preference over others and thus higher economic gains.

Figure 4 shows the correlation between emotions and the class ***C*** where blue color represents “Selected” and red color represents “Not selected”. It can be observed how emotions are related to gas station features in the selection or in the rejection of a gas station.

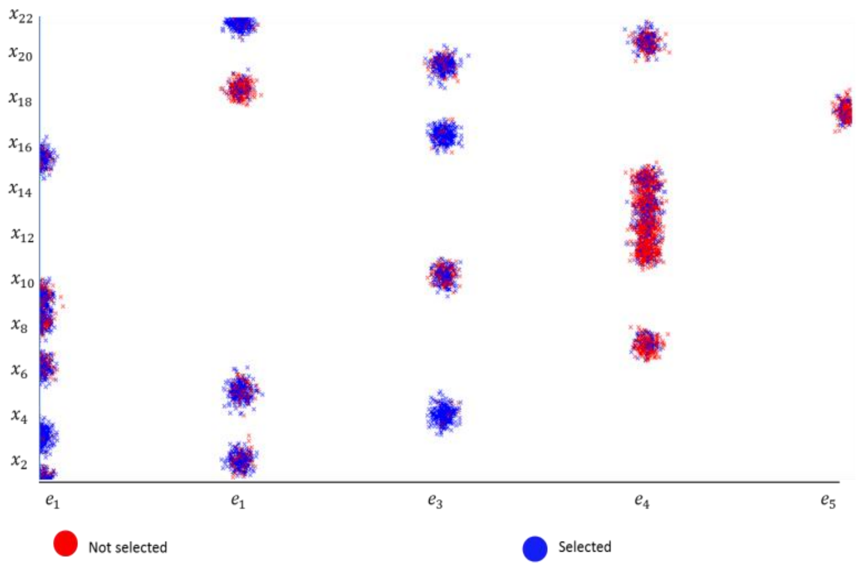


Fig. 4. Correlation between features and emotions.

The classification model can be represented in the following way. Probability ***P*** that has a gas station $\boldsymbol{g_i}$ to be selected is given by $\mathrm{P(x_i|selected)}$. This is expressed in equation 2:

$$P(g_i|selected) = \alpha \text{ and } P(g_i|not\ selected) = 1 - P(g_i|selected) = 1 - \alpha. \quad (2)$$

Our supervised model estimates the likelihood of a gas station to be selected or not, which observes a Bernoulli distribution with class C, where C = {selected, not selected} and where the set of features X is discrete. For this module we built a discriminant model also known as a binary classifier $\mathrm{r_i^t}$ where $\mathrm{X = \{x^t, r^t\}_{t=1}^N}$ and:

$$r_i^t = \begin{cases} 1\ si\ x^t \in C_i \\ 0\ si\ x^t \in C_j \end{cases} for\ j\ \neq i.$$

The selection probability is given by:

$$\mathrm{P(x^t|selected)} = \frac{\mathrm{e}^{\alpha+\beta\mathrm{x}}}{1 + \mathrm{e}^{\alpha+\beta\mathrm{x}}}. \quad (3)$$

Thus, the probability of being not selected is:

$$P(x^t|not\ selected) = 1 - P(x^t|selected). \quad (4)$$

Linearizing (1) by transforming $Y = ln\ \frac{y}{1-y}$ we have:

$$\frac{\log P(x^t)}{1-P(x^t)} = \alpha + \beta x \quad (5)$$

and then

$$y = \alpha + \beta x. \quad (6)$$

In the next section we present the results obtained for this classification model.

4.5 Results Comparisons

In this section are presented the results of the analysis performed by varying the training set at 80% and taking the rest of the set for testing. The following table shows the best results obtained with the classification ML methods. For the experiments, we used version 3.1.2 (2014-10-31) of R Studio Software. This software has been executed on an Intel R Core TM i5-2400 CPU @ 3.10 GHz with 16 Gb RAM computer, over Windows 7 Pro operating system, Service Pack 1 64 bit.

For the experiments we used the following classification methods: k-Nearest Neighbors (KNN), Random Forest (RF), C4.5-like Trees (C4.5-Trees), Rule-Based

Classifier (RBC), AdaBoost.M1 (AB), Stochastic Gradient Boosting (SGB), Neural Network (NN), Model Averaged Neural Network (ANN), C5.0, Single Rule Classification (SRC), Neural Network (NN), SVM with Radial Basis Function Kernel (SVM), Multivariate Adaptive Regression Splines (MARS), Mixture Discriminant Analysis (MDA), Naive Bayes (NB), Shrinkage Discriminant Analysis (SDA), Penalized Discriminant Analysis (PDA).

Table 4. Results using 80% of the dataset for training and 20% for testing.

Nº	Method name	Acc	Sens	Spec	Prec	F1-score	Bal. Acc.	Train time*	Test time*
1	KNN	87.54	81.15	91.91	87.27	89.53	86.53	2.05	0.29
2	RF	84.63	76.79	89.99	83.99	86.88	83.39	5.32	0.12
3	C4.5-Trees	80.77	74.77	84.88	77.17	80.84	79.82	0.45	0.07
4	RBC	78.12	81	76.14	69.89	72.88	78.57	1.06	0.04
5	AB	76.85	69	82.22	72.62	77.12	75.61	9.51	0.55
6	SGB	76.28	66.2	83.17	72.9	77.7	74.69	0.19	0.02
7	NN	75.84	65.89	82.64	72.18	77.06	74.26	0.72	0.01
8	ANN	75.71	63.4	84.13	73.2	78.29	73.76	3.72	0.01
9	C5.0	75.4	61.37	84.98	73.64	78.91	73.18	0.18	0.05
10	SRC	75.08	60.28	85.2	73.57	78.96	72.74	0.23	0.04
11	NN	74.95	62.77	83.28	71.96	77.21	73.03	0.71	0.01
12	SVM	74.45	53.43	88.82	76.56	82.24	71.12	6.36	0.77
13	MARS	73.88	55.45	86.47	73.71	79.58	70.96	0.33	0.02
14	MDA	73.75	56.07	85.84	73.02	78.91	70.96	0.29	0.01
15	NB	73.06	56.85	84.13	71.01	77.02	70.49	0.89	2.49
16	SDA	71.92	52.02	85.52	71.06	77.62	68.77	0.04	0.01
17	PDA	71.41	57.79	80.72	67.21	73.35	69.26	0.05	0.02
	Average	**75.36**	**62.97**	**83.83**	**72.68**	**77.75**	**73.4**	**1.7**	**0.25**

* The training time is expressed in seconds and the test time in milliseconds.

Table 5. Supervised classification methods parameter configuration.

Nº	Method name	Parameters name	Parameter values	Config.
1	AB	(mfinal, maxdepth, coeflearn)	(150, 3, 3)	27
2	BLR	(nIter)	(21)	3
3	C4.5 Trees	(C)	(0.25)	1
4	C5.0	(trials, model,winnow)	(10, 2, FALSE)	12
5	KNN	(kmax, distance, kernel)	(9, 2, 1)	3
6	MDA	(subclasses)	(2)	3
7	ANN	(size, decay,bag)	(5, 0.1, FALSE)	9

Nº	Method name	Parameters name	Parameter values	Config.
8	MARS	(degree)	(1)	1
9	NB	(fL, usekernel)	(0, TRUE)	2
10	NSC	(threshold)	(0.6496)	3
11	NN	(size, decay)	(5, 0.1)	9
12	PDA	(lambda)	(2)	3
13	RF	(mtry)	(2)	3
14	RBC	(threshold, pruned)	(0.25, 1)	1
15	SDA	(diagonal, lambda)	(FALSE, 1)	3
16	SRC	(-)	(-)	1
17	SGB	(n.trees, interaction.depth, shrinkage)	(150, 3, 0.1)	9
18	SVM	(C)	(0.25)	3

5 Conclusions and Future Work

In this work we show how client decisions are influenced by the emotions produced from the gas station characteristics.

We show that an ML approach can obtain the probability of selecting a gas station over another based on the emotions it produces in customers and it will remain valid if the population grows.

We would like to stress the importance of analytical decision-making. In this study it is found that ML approaches can improve the economical profit of investments made in Mexico in generating marketing strategies that make the clients prefer a national gas station over competency, emphasizing the gas station characteristics in adequate proportions according to the preferences of clients and their emotions.

References

1. Ratner, C.: A cultural-psychological analysis of emotions. Culture & Psychology, 6(1), pp. 5–39 (2000)
2. Frijda, N. H., Manstead, A. S., Bem, S.: Emotions and beliefs: How feelings influence thoughts. Cambridge University Press (2000)
3. Cambria, E., Melfi, G.: Semantic Outlier Detection for Affective Common-Sense Reasoning and Concept-Level Sentiment Analysis. In: FLAIRS Conference, pp. 276–281 (2015)
4. Turrentine, T., Kurani, K. S., Heffner, R. R.: Fuel economy: what drives consumer choice? Institute of Transportation Studies (2008)
5. Singh, S., Vrontis, D., Thrassou, A.: Green marketing and consumer behavior: The case of gasoline products. Journal of Transnational Management, 16(2), pp. 84–106 (2011)
6. Hansen, F., Christensen, S. R., Lundsteen, S.: Measuring emotions in a marketing context.Innovative Marketing, 2 (2), pp. 68–75 (2006)
7. Damasio, A. R.: En busca de Spinoza: neurobiología de la emoción y los sentimientos. Ed. Crítica, S.L. (2005)
8. Goleman, D.: Inteligencia emocional. Editorial Kairós (2012)

9. Ahn, H., Picard, R. W.: Affective-cognitive learning and decision making: A motivational reward framework for affective agents. In: International Conference on Affective Computing and Intelligent Interaction, Springer Berlin Heidelberg, pp. 866–873 (2005)
10. Ferrer, A.: Neuromarketing, la tangibilización de las emociones. Universidad Abat Oliba CEU, Francia (2009)
11. Minsky, M.: The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind. Simon and Schuster (2007)
12. Ponce, H., Martínez-Villaseñor, L., Miralles-Pechuán, L.: A Novel Wearable Sensor-Based Human Activity Recognition Approach Using Artificial Hydrocarbon Networks. Sensors, 16(7), pp. 1033 (2016)
13. Aurier, P., Guintcheva, G.: The Dynamics of Emotions in Movie Consumption: A Spectator-Centred Approach. International Journal of Arts Management, 17(2), pp. 5 (2015)

ISSN 1870-4069

Ye shall Know them by their Verbs: How Gender Express their Opinion in Twitter

Madai Ramírez, Octavio Sánchez

UNAM, Grupo de Ingeniería Lingüística,
Mexico

MRamirezCi@iingen.unam.mx, oct_sanc@unam.mx

Abstract. It is commonly said that adjectives are a kind of word that people use for emitting their opinion. This is because this lexical category designates the qualities of an entity. However, there are different ways to express an explicit opinion besides the use of adjectives. People also can choose verbs, nouns, adverbs and even groups of words that have complex meaning, like idioms or other multi-word expressions. The main purpose of this work was to discover which gender prefers which lexical category to express their opinion. In this paper, we show how sentiment analysis can help to identify a gender. A corpus of tweets was compiled for this research. The tweets were classified into 'opinions' and 'no opinions'. Within the corpus, we looked at which grammatical category of word was bearing the holder's subjectivity. We found that women used 27.98% verbs while men 16.91%, specifically psych verbs, in order to express their assessment.

Keywords: Author profiling, gender classification, opinion mining.

1 Introduction

Within social networks and various websites, million users express their opinion daily. Many of the opinions in these media remain anonymous. Knowing who wrote them might help for business intelligence or marketing as well as in security issues.

This paper is framed within forensic linguistics and opinion mining. The first discipline makes use of linguistic knowledge to solve legal problems such as plagiarism and authorship attribution. When the experts do not know who is the author of a text, they do author profiling, it means, they try to predict the age, gender and level of studies from the writer's style. In this work, we try to provide linguistic knowledge that could be used, in a future, to train a system for doing this prediction. We try to contribute to gender characterization by using an opinion mining analysis.

Opinion mining deals with subjective information, this is utterances that express appraisals, sentiments, or believes of people. It tries to find, extract and classify these expressions in text. For this work, we use the definition of opinion given by Liu [11]; this allowed us to tag them manually.

pp. 23–34; rec. 2016-05-24; acc. 2016-09-15

In this case, the opinion mining analysis was used to find out the differences of how a gender express their assessment. The hypothesis was that gender of a person can be identified by analysing the words they use to emit their opinion. We used a corpus of tweets that was made for this research, we looked at which kind of word or phrase was bearing the holder's subjectivity and we found that women used 27.98% verbs while men 16.91%, specifically psych verbs, in order to express their assessment.

In this work we provide information that can be used in automatic profiling. Instead of use all sentiment words, which imply high computational costs, we study which kind of subjective lexical category should be used and which other functional words has to be taken into consideration when profiling author by their emotions.

The rest of the paper is organised as follows. Section 2 will discuss past work related to this research. Section 3 is going to explain the methodology, which includes three subsections: how the corpus was made, how the corpus was labelled and the analysis. Section 4 talks about the results and finally, in section 5 we will present our conclusions and discuss future work.

2 Related Work

Since the middle of the last century, identifying the linguistic markers dependent on gender is a theme that has been exploited a lot. There has been papers like the one by Robin Lakoff [10] where she affirmed that women used more question tags and they asked questions when they really wanted to say a statement (rhetoric questions). Women also used more colours (e.g.: lavender), empty adjectives (e.g.: cute, divine, sweet, adorable) and weak insults when expressing themselves.

Subsequent studies [1, 17] found that women prefer using pronouns (I, you, he, she, her, their, myself, yourself, herself), while men prefer the use of determinants (a, the, that, these) and quantifiers (one, two, more, some). Furthermore, Scheler [17] identified affirmation and negation words as female features and prepositions, articles and hyperlinks as male features. The use of hyperlinks means that men share more news, videos, pictures and so on.

Ana Janssen [7] notices that women use more third person pronouns and men make more use of first person pronouns. Those results contrast with Schartz [18] who states that women use more the first person pronouns and emotive words and psychological processes.

In Mexico, Rivera Vidal's dissertation [15] studies how genders express their opinion with adjectives. However, we believe that this is not the only kind of words which can express subjectivity, there are also adverbs, interjections, nouns, verbs and many idioms.

In opinion mining, Mohammad and Yang [12] researched about toward which emotion each gender orient their words within love letters, hate mail and suicide notes. Women lean more their words toward happy and sadness sentiments and men to trust and fear.

In automatic author profiling, Patra et al.[13], Weren et al.[20] and Pimas et al.[14] used sentiments as a feature in order to classify gender and age. Besides, Patra et al. took into consideration pronouns, functional words and topics. With these features they got an accuracy of 56.83% for gender and 28.95% for age. Weren et al. considered the length of sentence, word and paragraph, and text readability. This means, how well written was the text, (repetition of vowels, punctuation, and misspelled words). Neither Patra nor Weren found that affective words can help to identify gender. Pimas et al. also used concreteness and syntactic information. For concreteness, they understand not abstract concepts. For sentimental features, they used *SentiWordNet* [4] in order to find the polarity of a word and they estimated the polarity of each tweet. As syntactic features they considered the word lenght and hashtags. They train their classifier with one type of text from social media, as tweets, blogs and others. They did the experiment twice and, the first time they got an accuracy of 0.5769 in the first set, and 0.0201 in the second one.

As it can be seen, the problem of profiling author from the subjective utterances they use, is far from being solved. We think there is a lack of linguistic knowledge on the area, therefore we try to offer this study in order to pave the way for automatic twitter author profiling based on the way they express their opinions.

3 Methodology

3.1 Corpus

We manually compiled a corpus of Twitter for this research because we wanted to control our demographic variables and it was the only way to guarantee that our data satisfy our requirements. There were a total of 40 participants: 20 men and 20 women; each one between the ages of 18 to 28, with a college degree or that were studying university and everyone was born and had been living in Mexico City.

The last 50 tweets of each account were recollected and they had to be written in Spanish, and had to be original tweets, not just quotations or retweets.

3.2 Labelling Corpus

Each instance of the corpus was tagged with 'opinion' or 'non opinion'. "An explicit opinion is a subjective statement that gives a regular or comparative opinion" [11]. Therefore, we only considered a tweet like 'opinion' if it had the elements for a 'regular' or a 'comparative' opinion.

According to Wiebe [21], a regular opinion must have 4 elements:

1. **Target:** The target or topic of the private state, i.e., what speech event or private state is about.
2. **Source:** The person or entity that is expressing the private state, possibly the writer.

3. **Text anchor:** a pointer of the span of texts that represents the speech event or explicit mention of a private state.
4. **Polarity:** The possible values are positive, negative, other or none.

With regard to the polarity, in this research, it was considered either positive or negative and in some cases we found that the message entails at the same time a positive and negative orientation, so we call those bipolar. E.g.:

(1) Me asusta lo mucho que me gusta Gossip Girl. ¡Maldita sea!
(It scares me how much I like Gossip Girl. Dammit!)

(2) ¿Qué tan culposo es el gusto por Salón Victoria! Sí, los tengo en mi iTunes.
(How guilty is the pleasure for Salón Victoria! Yes, I have it in my iTunes.)

Examples appear to be paradoxical; but ”they involve a dissociation of enunciating subject into an evaluator subject (with more or less objectified criteria) and a taster subject” [9]. That means, the people who wrote those tweets really like gossip girl and Salon Victoria, but the series and the band had some aspects that might be considered as dislikeable by the society, or even themselves.

In table 1 we show an example of a regular opinion extracted from the corpus:

(3) Breaking bad es una joya de serie … desde la forma hasta el fondo.
(Breaking bad is a jewel of series...from the form to the bottom).

Table 1. Example of a regular labelled opinion.

Target	**Source**	**Text anchor**	**Polarity**
Breaking bad	Hombre(man)	una joya de serie(a jewel of serie)	Positivo(Positive)

A comparative opinion “expresses a relation of similarities and differences between two or more entities and/or a preference of the opinion holder based on some shared aspects of entities”[8]. According to [11] comparative opinion must have:

- **E1:** entity 1
- **E2:** entity 2
- **A:** aspects
- **PE:** which entity the opinion holder prefers
- **H:** opinion holder.

As you can see in this schema, the polarity is not a feature because “this type of opinions are not directly positive or negative. Instead, the entities are being compared and ordered according to the aspects they share between them. This is, they express a preference for one or more entities” [11]. However, we designated one polarity to comparative opinions with respect to E1. In 2 there is an example of comparative opinions and how we labelled them:

(4) Honestamente no hay mejor voz femenina mexicana que la de ANA GABRIEL para cantar el Cielito Lindo
(Honestly there is no better feminine voice than Ana Gabriel for singing Cielito Lindo)

Table 2. Example of a comparative labelled opinion.

E1	**E2**	**holder**	**aspect**	**polarity**
Ana Gabriel	el resto de las cantantes femeninas mexicanas (the other mexican singers)	hombre (man)	no hay mejor voz femenina mexicana (there is no better mexican feminine voice)	positivo (positive)

In addition to the above labels, we tagged comparative opinions into the different kinds. It was considered that there are two types of comparative opinions: gradable and non gradable. Nevertheless, non gradable opinions in spite of comparing two or more objects, they do not rank them like gradable opinions do [11], therefore it could be difficult to identify an opinion and we did not classify them. Gradable opinions are subclassified [8] into:

- **Non equal gradable** relations of the type greater or less than that express a total ordering of some entities with regard to certain features. This type also includes user preferences.
- **Equative:** relations of type equal to that state to two entities as equal with respect to some features.
- **Superlative:** Relations of the type greater or less than all others that rank an entity over all others.

On the other hand, “non opinion” does not have the requirements previously mentioned for regular or comparative opinion like target, entity, sentiment, polarity, and so on.

It is important to say that some tweets had more than one opinion and they had to be separated, although it is the same subject but it is considered a different aspect of the target and was given another polarity. Then, every opinion was classified in themes.

3.3 Analysis

Once we had the corpus compiled and labelled, we made a linguistic analysis of the elements. In order to achieve this, the ‘text anchor’ was manually classified into one lexical category such as noun, verb, adjective, interjection and adverb.

We labeled axiological and affective adjectives. The former do an evaluation between ‘good’ and ‘bad’, while the latter take into consideration an object’s property and a subject’s emotional reaction [9]. Within this adjectives are evaluative (good, bad, ugly, beautiful) and adjectives of skills and human predisposition, which include emotional attitudes (sensitive, friendly, cordial), intellectual (intelligent, capable, sabio), and passions and primary disposal (nervous, aggressive) [2].

According to [9], there are occasionally subjective verbs and inherently subjective verbs. The first one evaluates the object of the process and the second ones evaluate the process and one of the agents. Among the first, there are psych verbs, which one express an appraisal and an emotional reaction at the same time (e.g. like, love, hate, appreciate), and say verbs where "the emotional state of x is explicit in a verbal behavior" (e.g. regret, blame, deplore, praise). Among the inherently subjective verbs, there are *stink*, *offend*, *infringe*, *deserve*, *failed*.

Sometimes it was found participles and gerunds outside verbal periphrases (which were tagged into verbs), in that case, participles were labelled as adjectives and gerund as verbs. Many nouns that we labelled are derived from verbs and adjectives, such as love, beauty, etc.

Interjections, like *¡Ay!*, *¡bah!*, *¡bravo!*, *¡guau!*, *¡Aj!*, *¡puaj!*, communicates feelings and impressions, they show various emotional reactions. This means that they can express surprise, assent or rejection, among other moods [3].

Moreover, it seemed appropriate to label different kind of multi-word expressions such as idioms and collocations like *¿eso qué?* (So what!)that reflected sentiments. Most MWE found in the corpus appeared as so in the online DRAE (Diccionario de la Real Academia Española), for example: *¡Maldita sea!* (Dammit!), *Valer la pena* (worth it). The rest has a similar structure and meaning.

The modifiers and syntax function were also analysed. The modifiers can increase or decrease the intensity of an opinion. We considered modifiers adverbs as *muy* (very), *demasiado* (extremely), *poco* (few/little), *tremendamente* (monstrously), *increíblemente* (unbelievably) [2]. Their syntax function can be attribution or predication. Attributive function happens when the adjective influence directly on the noun [16].E.g.:

(5) La **hermosa** casa donde solía vivir.
(The **beautiful** house where I used to live).

With regard to predication, the adjective is attached with the noun by and explicit or implied verbal copulation [16]. E.g.:

(6) La casa era **hermosa**.
(The house was **beautiful**)

Sometimes the adjective can function as a noun, this is called nominalization and it is made with an article (el, la, los, las, lo) plus and adjective [16].E.g.:

(7) Hay que resolver **lo difícil**.
(You have to solve **the difficult** [things]).

There is another adjective function which is not syntactical but it is discursive: vocative. It is like "interjections but, they are isolated words from the rest of the sentence by commas, reinforcement of intensity and special intonation in spoken language and written language, by commas. They belong to the appellate function in language" [5]. E.g.:

(8) La neta me dan muchísimo oso las personas que aprovechan los disfraces de jalouin para mostrar su cuerpo y verse "súper sexies". **¡Ridículos!**. (It is shameful when the people take advantage of Halloween costumes so that they can show their body and be 'super sexy'. **Ridiculous!**)

Finally, a frequency count of any single label was done, thus a statistical test called chi-squared was executed so we could evaluate the dependence of the variables. When you got p-value under 0.05, you can accept that there are dependence between the variables [6].

4 Results

Table 3 shows the quantity of 'opinions' and 'no opinions'. As it can be seen, the sum of both it is not 1000 because, as it was said before, some tweets had more than one opinion. In this case, women shared 4.08% more opinions, but the p-value of chi-square was 0.5098, which means there is no dependency between the variables gender and opinion.

Table 3. Opinions.

	men		**women**	
opinion	686	61.19%	720	65.27%
no opinion	435	38.80%	383	34.72%

Figures 1 and 2 present the most common topics of the opinions, those that exceed 2%. In both cases, the most common theme was 'person', with 16% and 19% respectively. Men talked more about 'politics', 'sports' and 'films'. 'Other' encompasses topics with percentage of occurrence less than 2% like 'climate ',' places', 'galaxy', etc. In this case, women used more and varied topics.

We also compared the polarity of the opinion, table 4 show the differences. In the corpus, women's sentiments were more positive than men's sentiments, 3.69%. Meanwhile, men's sentiments were more negative than women's sentiments, 3.93%.The p-value result was 0.3327, that means the variables, gender and polarity are independent.

The type of opinions: 'regular' and 'comparative', are shown in table 5. The results were not very different. For 'regular', 89.16% were feminine and 88.19%, masculine. For 'comparative', 10.83% were feminine and 11.80% masculine.

Table 6 shows the amount of type of comparative opinions. We did not find a big difference between genders.

Table 7 shows the lexical categories and phrases that was bearing the holder's subjectivity and which is their distribution inside our corpus. You can note that

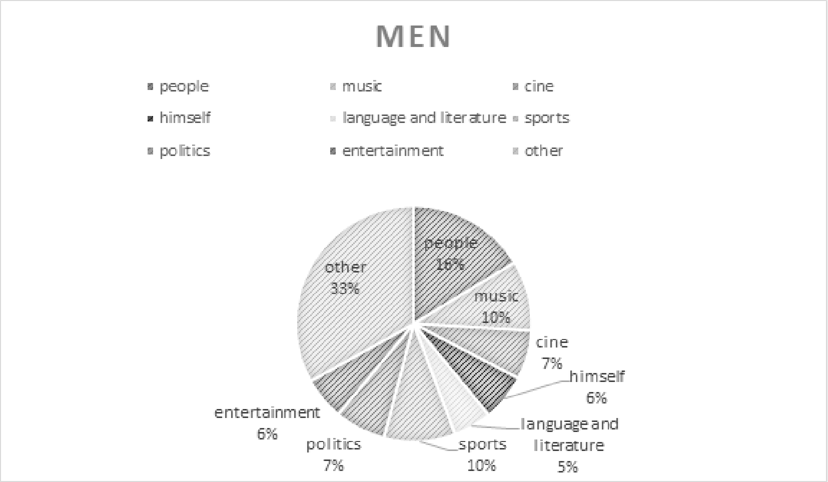


Fig. 1. Men's topic.

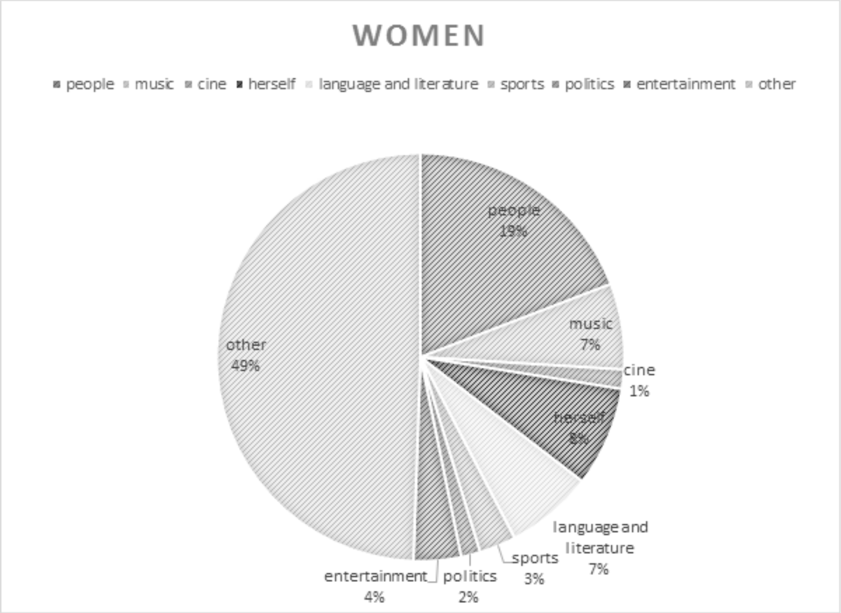


Fig. 2. Women's topics.

adjectives are the most useful lexical category, those represents almost the fifty percent of opinion words in the corpus. And men use more than women, almost 5% more. Nouns are more used by men (3.61% more) and verbs are more used by women (11.07%).The rest of categories do not exceed 1% of difference.

Table 4. Opinion polarity.

	men		**women**	
positive	331	48.25%	374	51.94%
negative	347	50.58%	336	46.65%
bipolar	8	1.16%	10	1.38%
total	686		720	

Table 5. Types of opinion.

	men		**women**	
regular opinions	605	88.19%	642	89.16%
comparative opinions	81	11.80%	78	10.83%
total	686		720	

Table 6. Types of comparative opinions.

	men		**women**	
superlative	56	69.13%	60	76.92%
non equal gradable	22	27.15%	17	21.79%
equative	3	3.70%	1	1.28%
total	81		78	

Table 7. Lexical category.

	men		**women**	
nouns	147	18.01%	122	14.40%
adjectives	433	53.06%	407	48.05%
verbs	138	16.91%	237	27.98%
interjections	16	1.96%	23	2.71%
adverbs	21	2.57%	12	1.41%
idioms	32	3.92%	41	4.84%
other phrases	29	3.55%	28	3.30%
total	816		847	

Because verbs are the best candidates in order to differentiate gender, we decided to look into it. We observed that psych verbs are the most common kind of verbs. Psych verbs are verbs that “denote emotional states as fear, liking and annoyance. And that implicate two arguments: an experimenter, which in this case refers to the person that experiences the emotion indicated by the verb, and a theme, which refers to the entity that relates with the emotion” [19]. We separated the verbs into 'psych' and 'no psych'. Figure 3 illustrates that. The darker colour shows that women used the double psych verbs than men.

Table 8 shows most common psych verbs. Note that women use more than double than men. We did a chi-squared test and we got 4.061e-06, it means the variables are dependent one from other.

Adjectives are the most common category for expressing opinion. Because of this, we decided to analyse the function, which is in the table 9, in order to see the

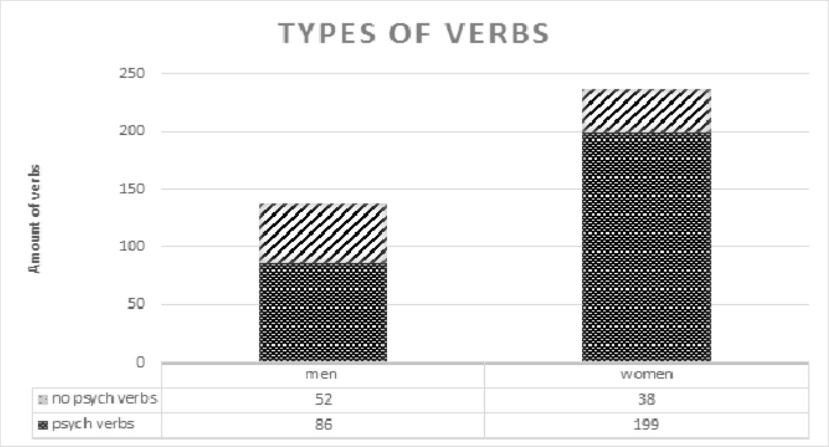


Fig. 3. Type of verbs.

Table 8. The most common psych verbs.

verbs	**men**	**women**
amar(to love)	6	39
gustar(to like)	12	26
querer(to want)	11	22
adorar(to adore)	0	12
extrañar(to miss)	8	16
odiar(to hate)	4	20

differences in use between genders. Men employed more adjectives in attributive function (50.57%) while women, more in predicative function (45.94%). We did chi-squared. The p-value was 0.05241. That means, there is no dependence.

Table 9. Adjectives function.

	men		women	
attributive	219	50.57%	172	42.26%
predicative	159	36.72%	187	45.94%
nominalization	42	9.69%	38	9.33%
vocative	13	3.00%	10	2.45%
total	433		407	

With regard to adjectives modifiers, 122 were masculine and 147, feminine. In other words, girls modified their adjectives 33.16%, when boys did it just 26.09%.

5 Conclusion and Future Work

In this work we produced a small Twitter corpus in order to be manually analised. It was done manually because we were interested in getting precise features of how men and women utter opinions. We produced linguistic knowledge that can be used later by forensic experts as well as automatic author profiling systems.

The results showed that men and women utter opinions in similar proportions in Twitter. It was found some differences in the issues that they review; however the polarity or the types of opinions that they used is not a distinctive feature.

Adjectives are the kind of word most productive for expressing opinions, but the meaningful difference consists in women preferring predicative function, while men preferred attributive function.

In addition, verbs are the kind of words that women had rather for expressing their opinion, specifically with psych verbs.

For future work, it will be researched if the results shown can also help to identify gender but with different age and scholar level. Besides it will be explored gender subjectivity and age group in another kind of private states like desires and beliefs.

Acknowledgements. This research was funded by a PAPIIT project called "Distributional semantics approaches for opinion mining" with code IN403016 and to the Consejo Nacional de Ciencia y Tecnología (CoNaCyT) for the support given through grant 387405 and project CB2012/178248. We also want to thank Red Temática de Tecnologías del Lenguaje de CoNaCyT

References

1. Argomon, S., Koppel, M., Fine, J., Shimoni, A.: Gender, genre, and writing style in formal written text. Text 23, 321–345 (2003)
2. Demonte, V., Bosque, I.: Gramática descriptiva de la lengua española. Espasa Calpe (1999)
3. Española, R.A.: Manual de la nueva gramática de la lengua española. Madrid. Asociación de Academias de la Lengua Española (2010)
4. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC. vol. 6, pp. 417–422. Citeseer (2006)
5. Gili Gaya, S.: Curso superior de sintaxis española. Bibliograf (1943)
6. Gries, S.T.: Statistics for linguistics with R: a practical introduction. Walter de Gruyter (2013)
7. Janssen, A., Murachver, T.: The role of gender in new zealand literature comparisons across periods and styles of writing. Journal of Language and Social Psychology 23(2), 180–203 (2004)
8. Jindal, N., Liu, B.: Mining comparative sentences and relations. In: AAAI. vol. 22, pp. 1331–1336 (2006)
9. Kerbrat-Orecchioni, C.: Los subjetivemas "afectivo" y "evaluativo"; axiologización y modalización. In: La enunciación de la subjetividad en el lenguaje (1086)
10. Lakoff, R.: Language and woman's place. Language in society 2(01), 45–79 (1973)

11. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies 5(1), 1–167 (2012)
12. Mohammad, S.M., Yang, T.W.: Tracking sentiment in mail: how genders differ on emotional axes. In: Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis (acl-hlt 2011. pp. 70–79 (2011)
13. Patra, B.G., Banerjee, S., Das, D., Saikh, T., Bandyopadhyay, S.: Automatic author profiling based on linguistic and stylistic features. Notebook for PAN at CLEF (2013)
14. Pimas, O., Rexha, A., Kröll, M., Kern, R.: Profiling microblog authors using concreteness and sentiment (2016)
15. Rivera Vidal, M.A.: Sociolingüística de los adjetivos calificativos en un corpus de español mexicano. Master's thesis, Universidad Nacional Autónoma de México (UNAM) (2015)
16. Romero Gualda, M.V.: El nombre: sustantivo y adjetivo. Arco (1989)
17. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. vol. 6, pp. 199–205 (2006)
18. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)
19. Vroon, S.: ¿ le está gustando la música o lo molesta el ruido? una investigación sobre el aspecto semántico de los verbos psicológicos & su uso con las formas de los pronombres átonos y en las perífrasis de gerundio. (2006)
20. Weren, E.R., Kauer, A.U., Mizusaki, L., Moreira, V.P., de Oliveira, J.P.M., Wives, L.K.: Examining multiple features for author profiling. Journal of Information and Data Management 5(3), 266 (2014)
21. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language resources and evaluation 39(2-3), 165–210 (2005)

ISSN 1870-4069

Automatic Detection and Registration of Events by Analyzing Email Content

Omar Juárez Gambino[1], José-David Ortega-Pacheco[1],
Consuelo-Varinia García-Mendoza[1], Miguel Felix-Mata[2]

[1] Instituto Politécnico Nacional, ESCOM, Mexico City,
Mexico

[2] Instituto Politécnico Nacional, UPIITA, Mexico City,
Mexico

omarjg82@gmail.com, david82d@hotmail.com, cvgarcia@ipn.mx, mmatar@ipn.mx

Abstract. In this paper we describe a system for automatic event detection and registration. This system monitors the received emails by the user and determines which of them are events based on their content. For analyzing the email content, Natural Language Processing techniques were used. Once an email is classified as an appointment, the details about this event are extracted and registered in the calendar of the user. Besides a mobile application notifies the user about this event and allows to view the specifics. 300 emails written in Spanish were collected to evaluate the performance of the system. A Naïve Bayes classifier was used to identify emails containing events and the specifics were extracted using pattern matching showing an overall system accuracy of 87%.

Keywords: Event detection, text classification, information retrieval.

1 Introduction

Electronic media has changed the way people communicate. Nowadays it is possible to send and receive a message from a person who is on the other side of the world in seconds or even have a conversation in real time. One of the most popular means of communication is email and it remains as the most pervasive form of communication in the business world. According to [3] in 2014 a business person receives and sends an average of 121 email per day. Due to the large amount of information users receive through email, the time invested in reviewing this information is increasing. In addition much of this information is impersonal, spam or little relevant to the user.

Text classification is an Artificial Intelligence task for automatically assigning predefined categories based on text content [8]. A comprehensive study of methods and applications of text classification is found in [1]. One interesting application is email classification, this problem has been approached from different perspectives. In [12] a Bayesian network was used to classify emails between

pp. 35–43; rec. 2016-05-28; acc. 2016-09-15

spam and non-spam. The authors show that some specific domain attributes like mail subject, email domain and typical grammatical structure of spam mails help to improve the precision of the classification. Assigning messages to user-created folders is another interesting problem, in [7] the authors used the Enron corpus [6], which is a large set of email messages made public, to prove the SVM classifier for this task.

Event detection is another application that analyzes text contents, but the objective is to determine if there is an event specified into it [4]. This application is specialy useful for automatic email analysis. Nowadays some email services can detect and even register (in the user calendar) some events as flight departures and concerts attendance but emails must have a specific structure. For unstructured information [2] proposed algorithms for isolating events emails from incoming messages written in English and algorithms for extracting specific information. We have not found yet an integrated system for automatically detect and register events in unstructured emails written in Spanish.

In this paper we describe the development of a system that classifies emails written in Spanish as an event or non-event and extracts the specifics for later register it in the user calendar. Natural Language Processing techniques were used to process the text and a Naïve Bayes algorithm was used for classification. In the following section we present an overview of the system and a description of each implemented phases (Section 2); the experimentation and results (Section 3); and finally our conclusions and future work (Section 4).

2 System Description

The developed system is composed by several subsystems and modules. In Figure 1 we show an overview. The whole system includes the following:

- Interaction with web email services. Gmail® and Outlook® are two of the most popular web email services, from these services we got the emails received by the user.
- Server side application. The received emails are processed by this application in order to determine if their content is related to an event.
- Interaction with a web application. If an event is detected the information related is registered in Google Calendar®.
- Mobile application. A mobile application receives a notification when a new event is registered and the user can display the specific details of it.

One of the main objectives in this work was to integrate the full process in a system; from the arrival of a new email, to the user's notification of the registered event. Despite the system has different modules, the user interacts only with the mobile application making the rest of the components invisible. All the components are described next.

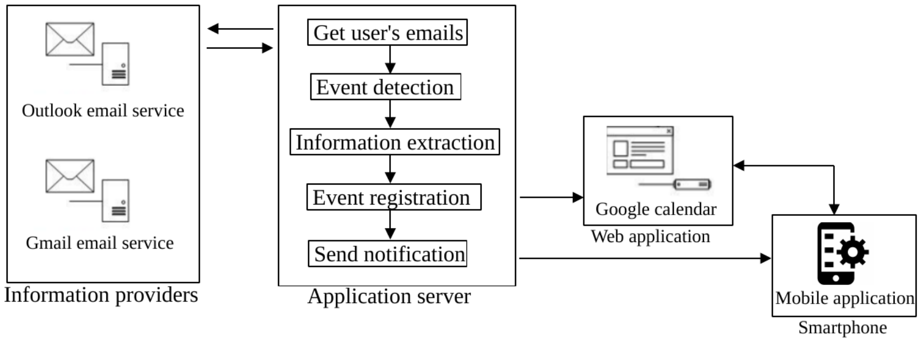


Fig. 1. System overview.

2.1 Email Client

The first module of the system is used for obtaining the received emails by the user, for this purpose an email client was developed. The client is constantly monitoring the user inbox and when a new one arrives the content is sent to the event detection module. To get the received emails, the address and the password of the account is required, this information is provided by the user during the configuration of the mobile application. Due to the sensitivity of the information, this is encrypted using the MD5 algorithm [11].

2.2 Text Processing

Before the system can determine if an email contains an event description, it needs to be processed. The email received by the client is in HTML format therefore a parser is used to extract the content. Once the content is obtained several Natural Language Processing techniques were applied using the Freeling [10] language suite:

1. Tokenization. Separates the content into words.
2. Lemmatization. Transforms the original words into their base form (lemma)
3. Named Entity Recognition (NER). Identifies entities like proper names and places. For our purpose dates and email addresses were also recognised.
4. Stopwords removal. A list of not useful words were eliminated, like pronouns and articles.

As an example of text processing consider the following email:

- "Hola buen día. Por medio de la presente se le informa que deberá asistir a la reunión de trabajo con el propósito de firmar su contrato laboral. La cita será el día miércoles 27 de Agosto de 2016 a las diez de la mañana en el World Trade Center de la Ciudad de México. Se le pide puntualidad. Cualquier aclaración comunicarse con nosotros. Atte. Luis Antonio Hernández De La Luz. Director General de COMPUMEX S.A. de C.V." (*Hello good morning.*

Through this you are advised to attend a working meeting with the purpose of signing the employment contract. The appointment will be on Wednesday, August 27, 2016 at ten o'clock in the World Trade Center of Mexico City. You are asked to be punctual. Contact us for any further information. Sincerely Luis Antonio Hernández De La Luz. COMPUMEX CEO).

After the text processing is applied the generated output is:

- "hola bueno día por_medio_de presente informar deber asistir reunión trabajo propósito firmar contrato laboral cita ser día date_pattern np_pattern se pedir puntualidad cualquiera aclaración comunicar np_pattern np_pattern" (*Hello good day through_this inform ask asist work meet purpose sign employment contract appointment be day date_pattern np_pattern you be ask punctual contact any further information np_pattern np_pattern*).

The generated output shows some named entities marked as "date_pattern" and "np_pattern". Some of these entities will be used to extract the details of the appointment. It is important to mention that all these transformations are used only during the appointment detection phase, so the original version of the email is preserved for the information extraction phase.

2.3 Event Detection

The processed text is sent to a Naïve Bayes classifier to determine if the email is related to an event or not. The classifier was previously trained with examples of both classes and provides the probability of belonging to classes.

2.4 Information Extraction

Once an email has been classified as an event it is important to extract its details. For this objective we follow a pattern matching approach. To correctly extract the details, the email content needs to have the following characteristics:

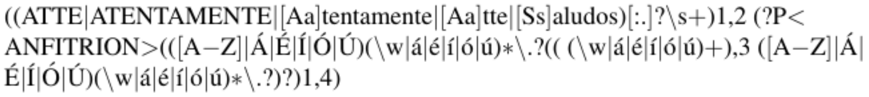
1. Written in formal style. Using appropiate language, avoiding the use of slang, abbreviations, etc.
2. Lexically and grammatically correct. Free of lexical and grammatical errors.
3. No nested references. The details of event must be contained in the current email, not in a previously referred.

The patterns were defined experimentally using the most common grammatical structures in formal emails. The content of the email is compared with the patterns, looking for matches. As an example, by using the patterns with the email shown in the text processing section, the extracted information will be:

- Purpose: la reunión de trabajo con el propósito de firmar su contrato laboral (*a working meeting with the purpose of signing the employment contract*)
- Starting date: 2016-08-27 10:00 (YYYY-MM-DD HH:MM)
- Ending date : 2016-08-27 10:00 (YYYY-MM-DD HH:MM)

- Place: el Word Trade Center de la Ciudad de México (*the World Trade Center of Mexico City*)
- Host: Luis Antonio Hernández De La Luz

In Figure 2 we show the pattern used for host extraction, considering the most common variations according to our corpus. It is likely that not all the emails classified as events will contain the full details. To handle this the system takes some considerations, for instance, if the purpose of the event is not found in the email content, the subject of the email is used instead. If the ending date is not found the starting date is used (like in the previous example). In the worst case some details could be blank, but the system would try to extract as much information as possible.



```
((ATTE|ATENTAMENTE|[Aa]tentamente|[Aa]tte|[Ss]aludos)[:.]?\s+)1,2 (?P<
ANFITRION>(([A−Z]|Á|É|Í|Ó|Ú)(\w|á|é|í|ó|ú)∗\.?(( (\w|á|é|í|ó|ú)+),3 ([A−Z]|Á|
É|Í|Ó|Ú)(\w|á|é|í|ó|ú)∗\.?)?)1,4)
```

Fig. 2. Pattern used for host extraction.

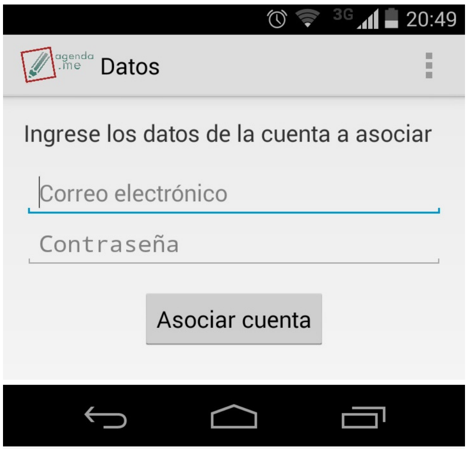
2.5 Event Registration

Calendar applications allow users to register events, thus they can remember them later. Nowadays there are a lot of calendar application available, and one of the most popular is Google Calendar. This application provided by Google associates a Gmail account to a calendar service, therefore every Gmail user has a Google Calendar. All the information related to an event that the system was able to extract it is sent to the Google Calendar web application in order to register it.

2.6 Mobile Application

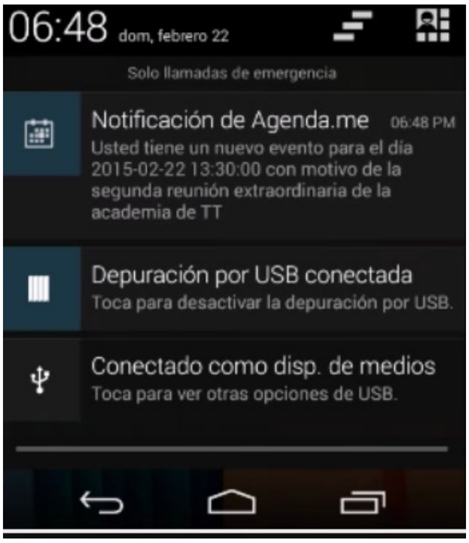
All the previous processes are invisible to the user because they are running in the server side. To allow the interaction of the user with the system an Android mobile application was developed. The user must have an active Gmail account and also an optional Outlook account. After installing the mobile application, it is necessary a configuration step. During configuration the user needs to register their email accounts (or account) by providing the address and password (see Figure 3).

When the server side application registers a new event in the Google Calendar account of the user a notification is sent to the mobile application. The user can display the details of the registered event by selecting it using the calendar of the mobile application. In Figure 4 a notification message is shown, while Figure 5 shows the calendar with the registered appointments.



Ingrese los datos de la cuenta a asociar

Fig. 3. The registration screen asking for user's email address and password.



Notificación de Agenda.me

Usted tiene un nuevo evento para el día 2015-02-22 13:30:00 con motivo de la segunda reunión extraordinaria de la academia de TT

Depuración por USB conectada

Toca para desactivar la depuración por USB.

Conectado como disp. de medios

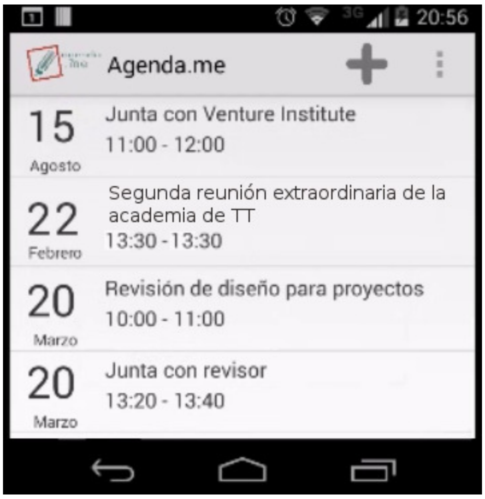
Toca para ver otras opciones de USB.

Fig. 4. The first element of the list shows a notification of a new event (*You have a new event in 2015-02-22 13:30 00 with the purpose of the second extraordinary meeting of the TT academy*).

3 Experiments and Results

3.1 Corpus

A corpus of Spanish emails of events and non-events was used to test the system. To our knowledge there is not a publicly available corpus with these characteristics, therefore it was created by ourselves. During a two months period



15 Agosto
Junta con Venture Institute
11:00 - 12:00

22 Febrero
Segunda reunión extraordinaria de la academia de TT
13:30 - 13:30

20 Marzo
Revisión de diseño para proyectos
10:00 - 11:00

20 Marzo
Junta con revisor
13:20 - 13:40

Fig. 5. The calendar shows some registered appointments, the second element is the previously notified event.

500 emails were collected from 30 different accounts. After that, two annotators manually annotated the emails as event and non-event. We discarded 26 emails because annotators could not reach consensus. Finally, we manually selected 150 event emails that met all the characteristics described in the information extraction section and 150 non-event email in order to have a balanced corpus.

3.2 Classification Test

For classifying the emails, the system used a Naïve Bayes algorithm. Following a vector space model all the words in the processed emails are counted and represented as a vector of frequencies. These vectors are the features that the Naïve Bayes classifier will use during training and testing. To evaluate the accuracy of the system a 10-fold cross validation was used. In Table 1 we show the average accuracy obtained for both classes.

Table 1. Average accuracy for event and non-event classes.

Class	**Accuracy**
Event	94%
Non-Event	97%

As we can see the system obtained more than 90% of correctly classified emails for both classes. This results show that it is possible to identify 9 of every 10 emails correctly using only their content.

3.3 Information Extraction Test

The system applies an extra phase to the emails classified as event for extracting the important information. If this information is not correctly extracted the event registered on the user's calendar will not be useful. For evaluating this module, we performed 3 different tests over 20 random selected event emails. The extracted information of every test email was compared with the manually extracted information described in the information extraction section. In Table 2 the obtained accuracy for every test is shown.

Table 2. Results for the information extraction module.

Test	Accuracy
1	78%
2	85%
3	81%

The results in Table 2 show that the automatically extracted information has an average of 81% of accuracy. Finally, in Table 3 we show the overall system accuracy by averaging the results obtained in both modules.

Table 3. Overall system accuracy.

Module	Accuracy
Classification (only event emails)	94%
Information extraction (3 test average)	81%
Overall system	87%

4 Conclusions and Future Work

In this work a system for automatically identify emails containing information regarding to an event was developed. The system can also extracts the details of the event and registers this information in a calendar associated with the user. Besides, a mobile application was created to notify the user when a new event is detected in the inbox and by using this application the user can view the specific details of the registered event. The evaluations performed over the classification module and the information extraction module got an overall system accuracy of 87%. The results show that it is possible to integrate different modules in a system to fulfill this task with good accuracy. As a future work, we would try other classifiers like SVM and MaxEnt which has proven effective in text classification task [5, 9]. It would be useful to increase the size of the corpus and also include a spell correction module to handle some common writing errors. For the information extraction module, more experiments are required in order to improve the accuracy, a richer lexical and syntactical representation could help.

Acknowledgments. We thank the support of Insituto Politécnico Nacional (IPN), ESCOM-IPN, UPIITA-IPN, SIP-IPN projects numbers: 20160815, 20160819, COFAA-IPN and EDI-IPN.

References

1. Aggarwal, C., Zhai, C.: A Survey of Text Classification Algorithms. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 163–222. Springer US (2012)
2. Black, J.A., Ranjan, N.: Automated event extraction from email (2004)
3. Group, R.: Email Statistics Report, 2014-2018. http://www.radicati.com/wp/wp-content/uploads/2014/01/Email-Statistics-Report-2014-2018-Executive-Summary.pdf (Jun 2014)
4. Hogenboom, F., Frasincar, F., Kaymak, U., Jong, F.D.: An overview of event extraction from text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011). vol. 779, pp. 48–57 (2011)
5. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: ECML-98, 10th European Conference on Machine Learning. pp. 137–142 (1998)
6. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Boulicaut, J.F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML. Lecture Notes in Computer Science, vol. 3201, pp. 217–226. Springer (2004)
7. Klimt, B., Yang, Y.: Introducing the Enron corpus. In: CEAS (2004)
8. Korde, V., Mahender, C.N.: Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications 3(2), 85 (2012)
9. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI-99 Workshop on Machine Learning for Information Filtering (1999)
10. Padró, L., Stanilovsky, E.: FreeLing 3.0: Towards Wider Multilinguality. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2012). ELRA, Istanbul, Turkey (May 2012)
11. Rivest, R.L.: The MD5 Message-Digest Algorithm. Request for Comments (RFC) 1321 (April 1992)
12. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: AAAI-98 Workshop on Learning for Text Categorization. pp. 55–62 (1998)

ISSN 1870-4069

A Lexical Supervised Approach for Opinion Mining in the Domain of Laptops and Restaurants

Karen Vazquez[1], Mireya Tovar[1], David Pinto[1], José A. Reyes-Ortiz[2]

[1] Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla,
Mexico

[2] Autonomous Metropolitan University, Systems Department, Azcapotzalco,
Mexico City, Mexico

krnlet@gmail.com, {mtovar,dpinto}@cs.buap.mx, jaro@correo.azc.uam.mx

Abstract. This paper presents a study of opinion mining or sentiment analysis for detection of polarity in a set of users opinion about restaurants written in Spanish and English. The research work is performed with the aim of solving a task proposed in SemEval 2016, thus we employed the same dataset proposed in that evaluation conference. The proposed approach uses a vector model for representing the information, including lexical features such as the following ones: word unigrams, bigrams and trigrams. The obtained results show a performance up to 71% when using word unigrams for representing the opinions written in English in the domain of restaurants.

Keywords: Opinion mining, vector space model, natural language processing.

1 Introduction

Nowadays, the major of people which is connected to Internet do it through social networks. Social communication media are used as a space for consuming and producing information. Thus, there is a great opportunity for studying, among other things, public opinions of consumers with the aim of providing information to final users and business owners about the quality of the service of, for example, restaurants and computer selling shops. In this manner, it would be possible to know the quality of a given restaurant according to the consumer rankings, or the best place to buy computers (for example, laptops) according to the consumer opinions.

This task has been proposed by SemEval 2016[1], a semantic evaluation forum, among other 13 tasks associated with semantic issues of natural language under-

[1] http://alt.qcri.org/semeval2016/

pp. 45–52; rec. 2016-05-29; acc. 2016-09-19

standing. The task number 5 (aspect-based sentiment analysis), in particular, its subtask 2 (text-level aspect-based sentiment analysis) is the one that has been considered for the experiments carried out in this paper [7]. The final aim is to automatically obtain the "category" (determined by the tuple: aspect-polarity) for a number of opinions about a entity or domain (in this case, restaurants or laptops) given by consumers (users or clients). Thus, the idea is to automatically detect the polarity of those opinions as positive, negative, neutral or conflict.

As contribution to solving this particular task, we propose to employ a vector space model with a number of lexical features based on word unigrams, bigrams and trigrams. The text representation schema considers the use of term frequency (TF) and inverse document frequency (IDF) for obtaining the representative vectors (TF-IDF) taking into account a training dataset provided by SemEval. The obtained results show a good performance when this model is employed.

The remaining of this paper is structured as follows. In Section 2 we present the related work. Section 3 describes the algorithm or model proposed. In Section 4.1 we describe the dataset employed in the experiments. The obtained results are given in Section 4.2. Finally, in Section 5 the conclusions are given.

2 Related Work

In recent years, various approaches have been proposed tackling the task of sentiment analysis through Natural Language Processing. Even if most of the works reported in literature deal with documents written in English, other languages such as Spanish have also been reported. In this section we present some research works related with the topic of this paper.

In [8] it is described the opinion mining system named "sentiue", which claims to determine the polarity of the sentiment expressed about a certain aspect of a target entity. This system participated in the Task 12 of SemEval-2015 obtaining a 79% of accuracy when determining the sentiment polarity of a given text.

In [2] it is presented a contribution to the Task 5 of SemEval 2016, working with documents written in English and French for user opinions for the domain of Restaurants. This system is based on composite models, combining linguistic features with machine learning algorithms. According to the reported results, they obtained 88% of accuracy for determining polarity in the restaurant domain (English language), and 78% of accuracy for determining polarity in the restaurant domain (French language).

In [4], authors describe the system they used in the task 5 of SemEval 2016. Their system is based on supervised machine learning, using a Maximum Entropy classifier, conditional random fields, and a large number of features such as global vectors, Latent Dirichlet Allocation, bag of words, emoticons, and others. They obtained very competitive results in the SemEval competition by using this system.

In [3], it is proposed a supervised term weighting scheme based on two factors: importance of a term in a document (ITD) and importance of a term for expressing sentiment (ITS). For ITD, they explore three definitions based on term frequency, and seven statistical functions are employed to learn the ITS of each term from training documents with manually annotated categories. The experimental results show that their method produces the best accuracy on two of three data sets.

The main objectives of the approach proposed by [6] are two-fold, first to improve feature-based opinion mining by employing ontologies in the selection of features, and second, to provide a method for sentiment analysis based on vector analysis. Their approach achieved an accuracy of 89.6% for the sentiment classification of the opinions in one of the following classes: positive, negative and neutral.

In the research work conducted by [1], they present an approach based on ontologies matching for opinion analysis. The aim of their work is to allow two enterprises to share and merge the results of opinion analyses on their own products and services.

Martínez Camara et al. [5] tested two classification algorithms (SVM, Naïve Bayes) and several weighting schemes and linguistic preprocessing (stopwords removing and stemmer) to determinate the opinion polarity in the domain of movies using Spanish language. The authors conclude that SVM works better than Naïve Bayes.

As we mentioned before, there are many works reported in literature associated with opinion mining, so we will avoid to be exhaustive on mentioning all of these works and we will proceed to describe the approach employed in our experiments.

3 Description of the approach employed

In this paper we propose an approach for determining the polarity of user opinions provided by organizers of Task 5 of SemEval-2016 [7]. First of all, we apply a preprocessing step to the training data in order to obtain the representative vectors for each tuple {aspect, polarity}. We do exactly the same process for the test dataset so that we can be able to apply the cosine similarity between these two datasets in order to determine the polarity of each element of the test dataset. The performance of the approach is obtained by comparing the results with those reported in the gold standard. Fig. 1 shows, graphically, the algorithm proposed.

PHASE 1

- Preprocessing:
  - Extraction of opinions from the XML document: To filter in order to obtain only the opinions from the XML document.
  - Cleaning of opinions: To remove stop words, punctuation symbols, isolated character and sorting of terms.
  - Tokenization: Tokenize opinions by words in order to obtain the vocabulary of the dataset.
  - Stemming: The aim is to reduce the vocabulary by stemming each word in order to reduce them to a common base form.

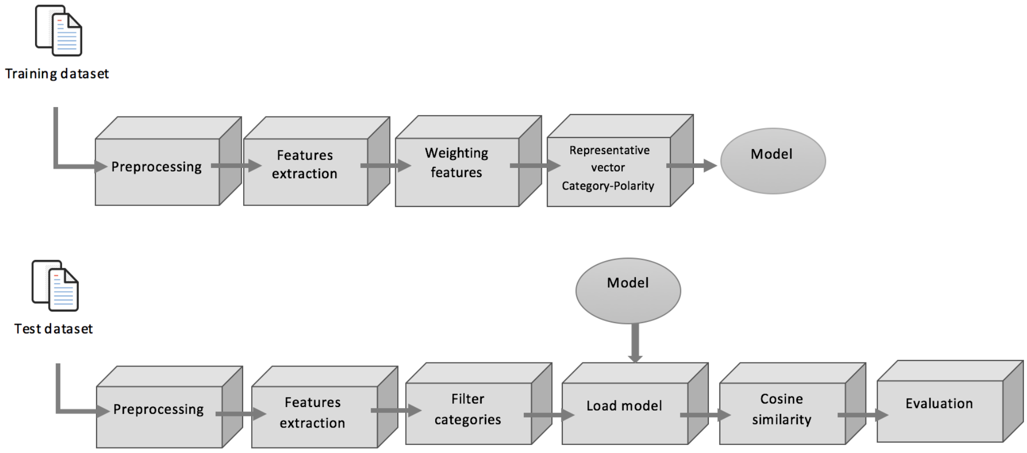


Fig. 1. Proposed algorithm.

  - Filtering categories: To filter opinions of the training dataset by entity and attribute, for example, RESTAURANT#GENERAL, FOOD#PRICES, etc.
- Weighting features extraction:
  - *Term Frequency* (TF): The number of times that a given term appears in a document or dataset, which allows to represent it.
  - *Inverse Document Frequency* (IDF): The number of documents in which a given term appear is calculated. This measure allows to determine how discriminative is a given term. Rare terms are more discriminative than common terms.
  - n-grams: For each weighting matrix, both TF and IDF are calculated for different sequences of words named n-grams. These text strings are the result of grouping together a sequence of words from a given text, previous preprocessing step. In this approach, we consider $n = 1, 2, 3$, i.e., word unigrams, bigrams and trigrams.
- Representative vector for each training data category-polarity:
  based on the weighting matrices generated with the training dataset, we proceed to create a representative vector for each category (Entity-Attribute) considering its associated polarity, for example, for the Entity ambience, the general attribute and its corresponding types of polarity (positive, negative, neutral and conflict), we obtain four different representative vectors:

$$\{AMBIENCE\#GENERAL, positive\}$$
$$\{AMBIENCE\#GENERAL, negative\}$$
$$\{AMBIENCE\#GENERAL, neutral\}$$
$$\{AMBIENCE\#GENERAL, conflict\}$$

- Detection by means of the cosine similarity measure:

We apply the cosine similarity measure, see Eq.(1), to determine the similarity between two weighting vectors, one of the training set and the other one from the test set. The target opinion will be assigned with the polarity according to the value obtained by the cosine measure (the highest one):

$$sim(\boldsymbol{d}_j, \boldsymbol{q}) = \frac{\boldsymbol{d}_j \cdot \boldsymbol{q}}{|\boldsymbol{d}_j| \times |\boldsymbol{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt[2]{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}. \tag{1}$$

– Polarity evaluation. In order to determine the performance of the approach, we employ accuracy as the evaluation measure.

4 Obtained results

In this section we describe the results obtained with the proposed approach.

4.1 Dataset

In the experiments carried out, we use the training and test datasets provided by SemEval 2016, task 5, subtask 2. Test dataset includes the gold standard evaluations, so that it is possible to measure the quality of the approach proposed. User opinions are given for two domains: Restaurants (written in English and Spanish), and Laptops (written only in English). In Table 1 we show the number of texts (opinions) provided by SemEval 2016.

It is very important to mention that opinions in both, training and test datasets, may be assigned with more than one category and polarity. In Table 2, it is shown the number of tuples that opinions may have associated for each domain, i.e., the different categories and polarities by opinion.

Table 1. Number of texts given for each domain in subtask 2 of task 5.

Domain	TRAINING	TEST	GOLD
Restaurant (Spanish)	627	268	268
Restaurant (English)	335	90	90
Laptops (English)	395	80	80

Table 2. Number of tuples by domain.

Domain	TRAINING	TEST	GOLD
Restaurant (Spanish)	2,121	881	881
Restaurant (English)	1,435	404	404
Laptops (English)	2,082	545	545

The domain Restaurants-Spanish presents 12 categories with four possible polarities for each one. In Table 3 it is shown the corresponding information for each type of category and polarity in the domain of Restaurants (Spanish and English), see also Table 4.

Table 3. Distribution of polarity and category (tuples) for the domain Restaurants-Spanish.

Category	Positive		Negative		Neutral		Conflict	
	Gold	Train	Gold	Train	Gold	Train	Gold	Train
AMBIENCE#GENERAL	61	150	27	50	3	10	5	11
FOOD#QUALITY	148	383	17	51	11	17	7	10
FOOD#STYLE_OPTIONS	30	79	14	47	4	5	2	3
FOOD#PRICES	23	61	13	47	1	5	1	0
RESTAURANT#GENERAL	175	436	45	108	19	36	15	23
RESTAURANT#PRICES	17	50	19	44	3	13	0	1
RESTAURANT#MISCELLANEOUS	6	8	5	4	1	1	0	0
SERVICE#GENERAL	123	301	30	64	5	4	4	10
DRINKS#PRICES	3	4	5	6	0	0	0	0
DRINKS#QUALITY	9	20	0	9	0	0	0	0
DRINKS_STYLE#OPTIONS	5	11	5	8	1	0	0	0
LOCATION#GENERAL	18	13	0	2	0	0	0	0
TOTAL	618	1516	180	440	48	91	34	58

4.2 Experimental results

Taking into account the algorithm aforementioned, tuples are first evaluated by category and thereafter by polarity. In Table 5 it is presented the domain, the total of tuples per domain, the amount of samples classified by employing the TF text representation with unigrams (1-gram), bigrams (2-grams) y trigrams (3-grams), and the same when using TF-IDF. In Table 6, the results are reported with average accuracy for each domain.

The obtained results allow to determine that TF reports accuracies greater than 50% for domain Laptops, whereas TF-IDF obtains acceptable results when word unigrams are used.

5 Conclusions

In this paper it is presented an algorithm for automatic classification for identification of polarity and category for a given set of tuples provided by task 5 of SemEval 2016, in particular, by subtask 2. The domains considered for the tests are Restaurants (Spanish), Restaurants (English), and Laptops (English). According to the obtained results, the proposed algorithm obtained a performance

Table 4. Distribution of polarity and category (tuples) for the domain Restaurants-English.

Category	Positive		Negative		Neutral		Conflict	
	Gold	Train	Gold	Train	Gold	Train	Gold	Train
AMBIENCE#GENERAL	34	130	1	21	2	10	1	4
FOOD#QUALITY	69	235	9	47	4	13	4	19
FOOD#STYLE_OPTIONS	15	62	7	28	6	3	4	2
FOOD#PRICES	6	33	9	34	3	0	0	2
RESTAURANT#GENERAL	68	249	20	76	1	4	1	6
RESTAURANT#PRICES	6	33	9	25	1	5	0	0
RESTAURANT#MISCELLANEOUS	13	38	8	22	3	6	0	2
SERVICE#GENERAL	43	140	19	61	1	6	1	6
DRINKS#PRICES	0	14	1	6	0	0	0	0
DRINKS#QUALITY	15	33	0	4	0	2	0	0
DRINKS_STYLE#OPTIONS	10	27	1	2	0	0	0	0
LOCATION#GENERAL	7	18	0	1	2	6	0	0
TOTAL	286	1012	84	301	23	55	11	41

Table 5. Results of opinions classified correctly according to polarity by domain.

Domain	Test	TF			TF-IDF		
		1-gram	*2-grams*	*3-grams*	*1-gram*	*2-grams*	*3-grams*
Restaurant (Spanish)	881	589	506	459	535	570	466
Restaurant (English)	404	288	248	249	268	248	248
Laptops (English)	545	290	276	295	324	302	296

of 50% of accuracy. Unigrams is the text representation schema that obtained the best results for the Restaurants-English domain with an accuracy of 71%. TF obtained the best results with unigrams for the domain Restaurants-Spanish. Finally, unigrams with TF-IDF obtained the best results for Laptops-English with an accuracy of 59%.

Important is to mention that we have not employed any additional resource, such as dictionaries or lexicons for this classification process. We have only employed information provided by SemEval. As future work we plan to employ other linguistic resources as SentiWordnet as well as other text features for improving the performance of the approach.

Table 6. Accuracy results by domain.

Domain	TF			TF-IDF		
	1-gram	*2-grams*	*3-grams*	*1-gram*	*2-grams*	*3-grams*
Restaurant (Spanish)	**66.85**	57.43	52.09	60.72	64.69	52.89
Restaurant (English)	**71.28**	61.38	61.63	66.33	61.38	61.38
Laptops (English)	53.21	50.64	54.12	**59.44**	55.49	54.31

Acknowledgements. This research work has been partially supported by PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, by VIEP-BUAP project 00570 and CONACyT project 257357.

References

1. Balaguer, E.V., Rosso, P., Locoro, A., Mascardi, V.: Análisis de opiniones con ontologıas. Polibits (41) pp. 29–36 (2010)
2. Brun, C., Perez, J., Roux, C.: Xrce at semeval-2016 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 282–286. Association for Computational Linguistics, San Diego, California (June 2016), TOBEFILLED-http://www.aclweb.org/anthology/W/W05/W05-0245
3. Deng, Z.H., Luo, K.H., Yu, H.L.: A study of supervised term weighting scheme for sentiment analysis. Expert Systems with Applications 41(7), 3506–3513 (2014)
4. Hercig, T., Brychcín, T., Svoboda, L., Konkol, M.: Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 354–361. Association for Computational Linguistics, San Diego, California (June 2016), TOBEFILLED-http://www.aclweb.org/anthology/W/W05/W05-0257
5. Martínez Cámara, E., Martín Valdivia, M.T., Perea Ortega, J.M., Ureña López, L.A.: Técnicas de clasificación de opiniones aplicadas a un corpus en español. Procesamiento del Lenguaje Natural 47(0), 163–170 (2011)
6. Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M.Á., Moreno, V., Fraga, A., Sánchez-Cervantes, J.L.: Feature-based opinion mining through ontologies. Expert Systems with Applications 41(13), 5995–6008 (2014)
7. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S.M., Eryiğit, G.: Semeval-2016 task 5: Aspect based sentiment analysis. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 19–30. Association for Computational Linguistics, San Diego, California (June 2016), http://www.aclweb.org/anthology/S16-1002
8. Saias, J.: Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 767–771. Association for Computational Linguistics, Denver, Colorado (June 2015), http://www.aclweb.org/anthology/S15-2130

ISSN 1870-4069

Automatic Translation System from Mexican Sign Language to Text

Luis Alberto Flores Montaño, Rosa María Rodríguez-Aguilar

Universidad Autónoma del Estado de México, Unidad Académica Profesional Nezahualcóyotl,
Departamento de Ingeniería en Sistemas Inteligentes, Estado de México,
Mexico

United_fenrir@hotmail.com, rmrodrigueza@uaemex.mx

Abstract. The Mexican Sign Language (*Lenguaje de Señas Mexicano*, *LSM*) consists of movements of the human body and hands to communicate, this is used by the deaf-mute to articulate thoughts and emotions. Thus emerges the initiative to device an Automatic Computer System for the Translation of Sign Language to Text, in order to facilitate communication to deaf and non-deaf people. The system is based on pattern recognition. This recognition is applied continuously using a capture device, in this case a Web camera using the Visual Studio IDE specifically with the OpenCV library. The recognition of the signs is made by comparing a matrix, previously obtained by means of several algorithms for the extraction of the object of interest, in this case the hand. In tests for the recognition of numbers, an average reliability percentage of 80%was obtained for the patterns analyzed.

Keywords: Pattern recognition, Mexican sign language, image processing, segmentation.

1 Introduction

Human communication is the transmission of information through various languages, among which, oral expression is the ideal way to share ideas, thoughts, emotions, etc. Human communication has as its main purpose that the individual achieves a satisfactory social integration that (in turn) allows the scientific and technological development that society requires.

However, hearing and visual disabilities of people is a problem that continues to demand immediate attention because such disabilities limit the opportunities for social, emotional, educational, professional and cultural integration that every human being deserves. People with visual impairment find support in a system of engraved symbols (Braille language) registered by means of touch for reading and writing. The hearing impaired people use sign language (signs articulated with movements of the hands and fingers) to communicate. In particular, sign language uses various movements generated with the body (corporal movement) and the face (gestural movement) to emphasize the expressions established by the hands and fingers, thus allowing to establish a communication that, despite being non-verbal, is significant.

pp. 53–67; rec. 2016-06-08; acc. 2016-09-15

Currently, the rapid development of intelligent computer systems has begun to impact positively on the social integration of the auditory disabled with applications that facilitate their non-verbal communication. For the recognition of a gestural movement, there have been a series of works based on the digital processing of signs, [which are] captured in video-images in order to be translated into text with the possibility of voice assigning for mobile devices. Such applications recognize body and gestural movements in real time by means of techniques such as neural networks for non-supervised learning [1].

1.1 Mexican Sign Language

Sign language is considered one of the earliest forms of communication used by the deaf community, [and it] consists of a series of gestural signs articulated with the hands, accompanied by facial expressions, intentional look and body movements. In 1620, Juan Pablo de Bonet in Spain, wrote the first book to teach sign language for deaf-mute people. In 1755 Charles Michele L'Epee in Paris, France founded the first school for people with hearing disabilities in the world. In 1778 Samuel Heinicke of Leipzig, founded the first school for deaf-mutes in Germany. In the United States the natives had a communication system similar to that of deaf-mute people [2].

In Mexico, the Mexican Sign Language (*Lenguaje de Señas Mexicana, LSM*), which is composed of the dactylology and ideograms, is used. Where dactylology is sign language generated by the gestures of the hand, which seek to interpret each letter of the alphabet through different shapes or forms, as shown in Figure 1. On the other hand, the ideograms are representations of words with one or more hand configurations [2].

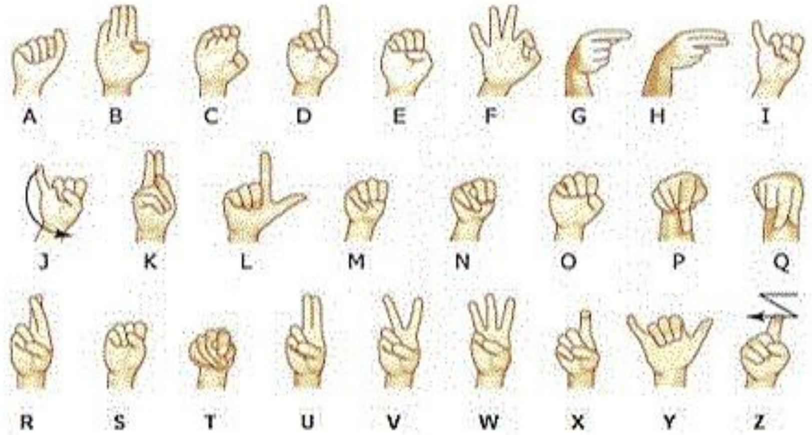


Fig. 1. Mexican Sign Language (LSM). [3]

El Colegio de México [The College of Mexico] developed the Mexican Spanish-Sign Language Dictionary (*DIELSEME*) to support the purpose of contributing to socialization with the *LSM*, since it is a useful tool for teaching and learning the language.

The *DIELSEME* is considered bilingual because it is an answer to the basic need to teach the *LSM*, with reference to written Spanish, to deaf people as well as to a variety of potential users: [*i.e.*] hearing parents with deaf children, teachers of all educational levels teaching deaf students, interpreters in training, interpreter schools and society in general [3].

2 Used Algorithms

The recognition of the signs is made by comparing a previously obtained matrix, — by means of the Pavlidis algorithms (for contour), the three coins algorithm (for convex hull), the Skyum algorithm (for minimum enclosing circle), and lastly the Wen and Niu algorithm (for the detection of fingertips) — for the extraction of the object of interest, in this case the hand.

2.1 Theo Pavlidis Algorithm

The Theo Pavlidis Algorithm is used to find the outline of an image. [4] This algorithm begins with a starting point; having located this starting point, the scanning of each row of pixels can be executed, starting from the lower left-hand corner. The scanning of each row starts from left to right. When the white pixel is found, it is declared as the initial pixel. After identifying the initial pixel, three more pixels, P1, P2 and P3 — which are of interest for the algorithm [5] — are located.

P1 =Upper left pixel
P2 =Upper pixel
P3 =Upper right pixel

All points are relative to the start pixel, as shown in Figure 2.

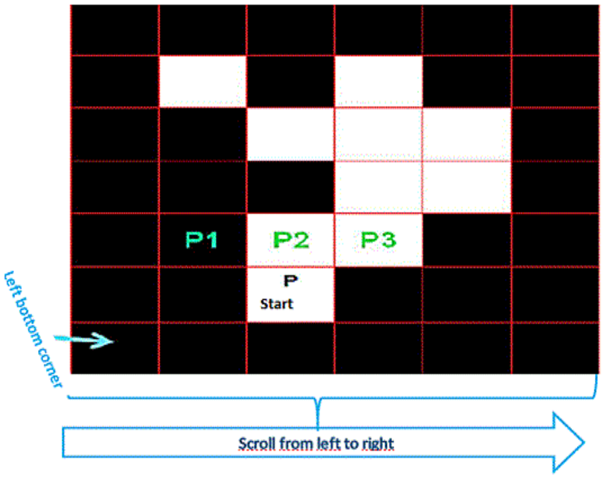


Fig. 2. Algorithm of Theo Pavlidis, position of P1, P2 and P3.

When the start pixel is located, there are 4 [possible] cases to consider for the recognition of the contour of the image, they are enumerated in detail in the following algorithm:

```
Input: A binary image that contains a group of white pixels.
Output: A sequence with B (b1, b2… bk) of the limit pixels as contour.

Start
Scan each pixel in rows from the lower left part until a white pixel is found s
Insert s in B and adjust so that it is a START pixel.
while (s has not been visited 2 times)

If pixel P1 is white
—Insert P1 in B
—Adjust s = P1
—Move one position forward followed by a new position in the current [one] to
  the left
else if P2 is white
—Insert P2 = B
—Adjust s = P2
—Move one position forward
else if P3 is white
—Insert P3 in B
—Adjust s = P3
—Move one position to the right and actualize its position, move one position in
  the current [one] to the left
else if s has been rotated 3 times
—End process and declare s as an isolated pixel.
else
—Rotate 90° clockwise while P is positioned over the current pixel
End
```

Fig. 3. Formal description of Pavlidis Algorithm [5].

2.2 Convex Hull and Convexity Defects

The convex hull method or three-coin algorithm is an efficient and robust procedure to obtain the structural parts of the hand, which consists in calculating a convex envelope of the contour of the hand and [then] comparing it. The calculation of a convex polygon envelope consists in obtaining a polygon, the sides of which connect the outermost points of the contour, thus eliminating any concavities or convexity defects that might arise [5].

Obtaining the convexity defects allows the identification of the interdigital spaces, in addition to other concavities generated by the shape of the hand or errors in calculating the contour. The defects are represented with their starting point, their end point and their point of maximum depth [5]. See figure 4 for details of the convex hull algorithm.

Input: *A set of N-points that form a polygon on a plane.*
Output: *A set of points that form the convex hull of the input.*

Start:

- *Find the point located furthest to the left as a start point. Label it as p0.*
- *Label the rest of N-1 points in a clockwise direction.*
- *Place the three coins on p0, p1, p2 and label them respectively as end, middle and front coin.*

Make up to: "the front coin", is the start point, and from there rotate to the right.
if the three coins form a right hand rotation

- *Take "end coin", place it on the next point following "front point".*
- *"End coin" is now "front coin", "front coin " is now "middle coin", la "middle coin" is now "end coin"*

else coins make a left hand rotation.

- *Discard the point (associated to the borders) that is on the "middle coin".*
- *Take the "middle coin", and place it on the former "end point".*
- *"Middle coin" becomes "end coin", "end coin" becomes "middle coin".*

End
The remaining points form a convex hull when connected in an order.

Fig. 4. Formal description of the three coin algorithm.

2.3 Minimum Enclosing Circle

The minimum enclosing circle is used in computational geometry, sometimes in planning to locate a shared facility in an easy way. A hospital or a post office are good examples of a shared facility. If one considers each household in the community as a point in the plane, the minimum center of the enclosing circle is a suitable location for [the] shared installation, as shown in figure 5. This idea is applied to find the position of the hand palm [6].

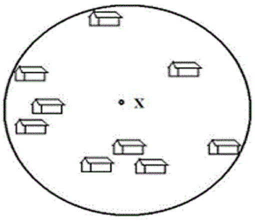


Fig. 5. Minimum enclosing circle [6].

2.4 Wen and Niu Algorithm

The algorithm, proposed in 2010 by Wen and Niu, marks up the fingertips on the contour calculating the angle of each point. The contour of the hand includes only the palm and fingers. The forearm is removed by covering it in swaddling. In order to identify the closest points on a contour, the angle of a point needs to be defined. This can be done by means of the following equation (1) [5].

$$\text{Fingertip angle} < X >= (X - X_1) \bullet (X - X_2)/|X - X_1||X - X_2|, \qquad (1)$$

where:

- X1 and X2 are separated the same distance away by X.
- X1 is the previous point with a certain number of points before X.
- X2 is the next point with a given number of points after X.
- The symbol (•) means that the scalar product.
- The value of the fingertip-Angle <X> is between 0 and 1, as it is the cosine value of X1 X2. Figure 6 illustrates the angle of a point [5].

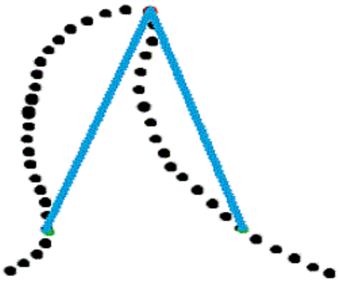


Fig. 6. Angle of a contour point [5].

3 Methodology

COMET (Concurrent Object Modeling and Architectural Design Method), is considered the first methodology for the development of our *LSM* system because it is a methodology that bases its operation on use cases for object modeling and architectural design of concurrent applications, distributed and real-time applications in particular, as is the case of our system.

During the investigation and under the criteria that the COMET methodology provides, it was identified that, during the integration and increment stage of the software, [it?] was not [robust] enough to solve the problem in order to perform the recognition and interpretation of the signs expressed with the hands. In this manner, based on the first tests carried out with the software in Microsoft Visual Studio 2010 with [the] OpenCV tool, which is a free library for artificial vision analysis; it was decided to select another methodology named Automated Object Recognition (*Sistema* *de Reconocimiento Automatizado de Objetos*, *SRAO*), complementary to the research by ensuring better results for the system on the part of the video image recognition.

3.1 COMET

It serves as a cycle of software development — based on use cases — working sequentially with the phases that comprise it: development of requirement modeling, analysis modeling, design modeling, and construction and incremental integration of the software in an iterative development cycle, until the validation phase of the system. [10].

3.1.1 Description of the Phases of the COMET Methodology

This methodology consists of the following phases:

- Requirement Modeling: We will have the support of a person who performs *LSM* signs, and a capture device will be used for the representation of the movements made by the user. [10].
- Analysis Modeling: The movements detected by the image capture device will be studied, based on user requirements (initial prototype), in order to then analyze each signal captured with an appropriate pattern recognition algorithm. (Medina, s. f.)
- Design Modeling: Once the image to be processed is captured, it will be analyzed and directed to the knowledge base where the respective sign will be compared. [10].
- Construction and incremental software integration: Through the image capture device, the input image of the human body will be taken, and [then] it will be processed in real time, by means of various algorithms, in order to obtain the skeleton of the image and thus build an approximation based on the results of the image analysis. Once having obtained the latter [results?], they [in turn] will serve to structure the new prototype (Incremental Prototype), this based on the initial characteristics and new results of the algorithm. [10].
- Validation of the system: The system should validate the entry of new data and compare that the results are true, in order to determine the corresponding translation of Mexican Sign Language to written form, returning the text that corresponds to the image on screen. If validation is not carried out successfully, the results are once more assigned to the incremental prototype phase, compared with established features and directed to a new construction in software integration. [10].

3.2 AORS Methodology

The problem that arises with automatic object recognition is to identify and label objects that the capture receives. More specifically, the problem can be described as follows: given an image or capture containing one or more objects of interest, including the background and a set of labels (one for each region of the image), the system must assign labels that correspond to known models or set of regions in the image [11].

3.2.1 Phases of the AORS Methodology

A System for Automatic Object Recognition (AORS) has the following stages:

- **Detection**
- **Acquisition of characteristics**
- **Tracing**
- **Interpretation**

a) **Detection** The purpose of this module is to obtain the image by means of the webcam in real time and retrieve regions that indicate the recognition of a pattern. In this case, our object of interest, a hand. This functionality is accomplished through a learning process focused on a collection of images scanning all along the image until the calculation of the model is obtained. The function is to divide the image into subgroups of pixels. Where each subgroup approximates, in shape and number of pixels, the region of each of the objects in the image. The instant that the detection is satisfactory, that is, that the region of interest has been obtained, then we proceed to the next stage "Getting the characteristics".

b) **Obtaining [of] Characteristics.** This step permits the calculation of the contour of the hand, then the calculation allows to obtain its structural points, and — based on these points — a number of features in common are sought in order to verify that the region actually corresponds to a hand. Another feature that this module includes is [the possibility of] improving the image quality for the processing and calculations to be obtained in the following stages and thus have better results. Operations on the image are performed in order to detect the presence of known objects in the image, these operations are applied to the subsets formed in the segmentation stage. As well as the usage of filters or other methods to reduce noise in images. Feature extractors or operators depend on the type of objects to be recognized and on the models stored for reference.

c) **Tracing.** Once the previous stage "Acquisition of Characteristics" has been executed successfully, one proceeds to the next step which is the tracing of structural points previously obtained along the image [in turn] acquired from the camera.

d) **Interpretation.** Lastly, and after having followed the structural points correctly, the interpretation is carried out, as to the position of structural points. This allows the recognition of gestures and signs of the hand, which are considered as displacements, rotations, finger flexions or the combination of all, to finally take action when they occur. It is worth mentioning that for each incoming image in the camera, the monitoring and interpretation stage will run continuously until a gestural movement is not recognized; in which case the algorithm will restart automatically returning to the execution stage.

4 Digital image Preprocessing Stages

The first step is the change of color space, from an RGB to an HSV color space, in order to make the image capture less sensitive to the brightness of the environment; this is achieved with a function of the OpenCV library that performs the procedure.

Binarization was subsequently implemented with different functions belonging to OpenCV within the HSV color space, as well as previous parameters for image segmentation.

4.1 Extraction of Characteristics

The characteristics extraction problem was solved as follows: first the input is a video capture in real time and the output was established by a set of patterns calculated over the video capture. When the image/video capture is obtained, a treatment is applied to it, with the color space (RGB to HSV), [with a] morphological application (erosion and dilatation) in order to eliminate noise and segmentation (binarization); with the latter we have the discriminated areas (black) and of interest (white); then the Pavlidis algorithm is applied using a function belonging to OpenCV in order to get the coordinates on the border between white and black pixels, and then have them stored in vectors (of points), as a consequence of all this the contours of the object of interest, in this case the hand, is obtained.

With vectors of points in coordinates, [and] in order to detect borders, two more functions are additionally used, both also belonging to OpenCV, with which the convex hull and the enclosing circle will be obtained, the latter two are described below:

1. For the convex hull, a counter-clockwise scan is executed, which takes care of enveloping the previously detected contour, subsequently the center of the object is sought by [means of] the function of the minimum enclosing circle. Due to the implementation of these functions, patterns of distances between the center and the contour of the hand can be obtained, the latter help detect convexity defects. Figure 7.

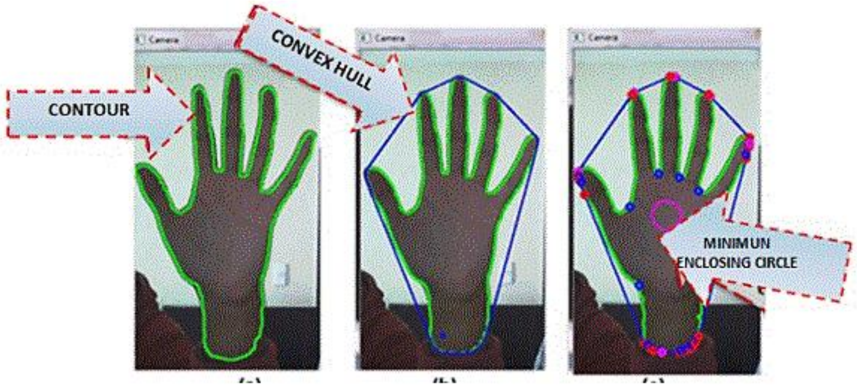


Fig. 7. (a) Contour. **(b)** Convex hull. **(c)** Minimum enclosing circle.

2. The convexity defects were tested as follows: for each defect of convexity there are 4 types of points, the first is the "startpoint", "endpoint", "farpoint ", and finally "depthpoint ", the latter determines the farthest distance between the contour and the convex hull; therefore, with this set of points will obtain specific characteristics, in this case patterns to identify each signal from the *LSM* made by the object of analysis, in this case the hand.
3. To obtain the pattern count (points) a list of these will be made, named F (n), so there will be from F1 to F24, this will be done heuristically, the number of comparisons of variables being 24 (depth, pt.Start.y, ptEnd.y, ptFar.y, ptStart.x, ptEnd.x and ptFar.x); against the constants (mycenter.x, mycenter.y and 11), using for this the operators "greater than" and "less than", so the hand zone will be divided into several areas about its center — as shown in figure 8 —, then the F (n) that are not useful — because they show a large number of patterns (points) — are discarded.

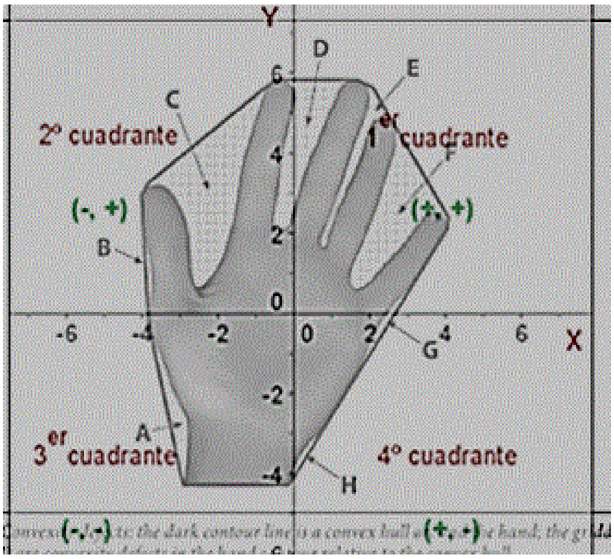


Fig. 8. 1st quadrant, 2nd quadrant, 3rd quadrant, 4th quadrant.

Below, some of the validations of F (n) that were functional for the signs to be detected are shown:

F1: If (depth > 11 && pt.Start.y < mycenter.y) (2)

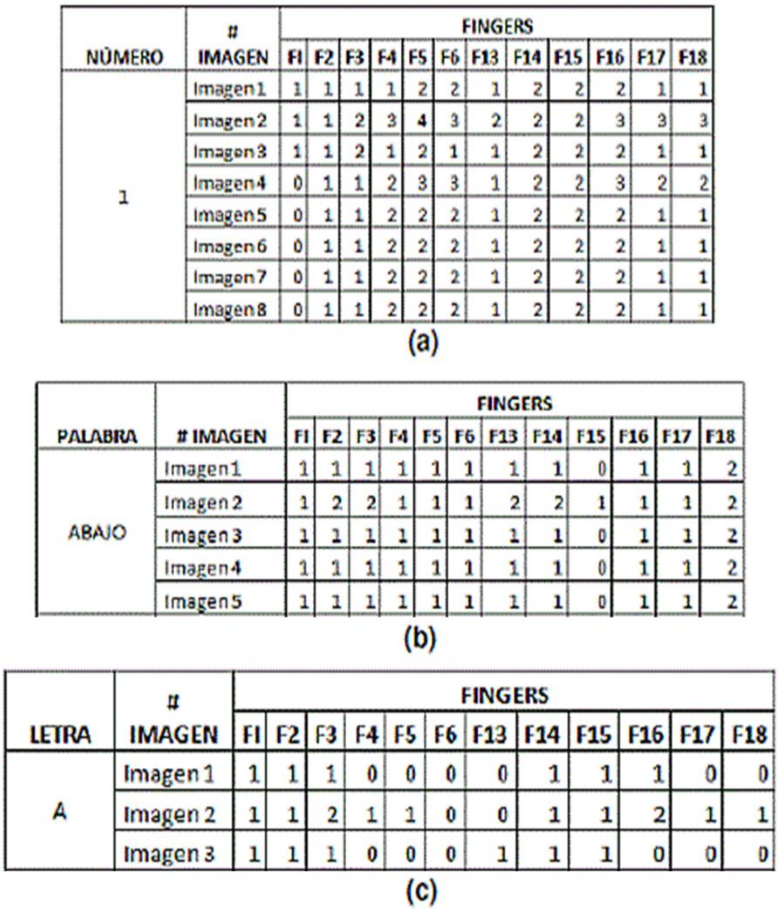
F2: If (depth > 11 && ptFar.y < mycenter.y) (3)

F3: If (depth > 11 && ptEnd.y < mycenter.y) (4)

F13: If (depth > 11 && pt.Start.x < mycenter.x) (5)

F14: If (depth > 11 && ptFar.x < mycenter.x) (6)

F15: If (depth > 11 && ptEnd.x < mycenter.x) (7)



NÚMERO	# IMAGEN	FINGERS											
		F1	F2	F3	F4	F5	F6	F13	F14	F15	F16	F17	F18
1	Imagen1	1	1	1	1	2	2	1	2	2	2	1	1
	Imagen2	1	1	2	3	4	3	2	2	2	3	3	3
	Imagen3	1	1	2	1	2	1	1	2	2	2	1	1
	Imagen4	0	1	1	2	3	3	1	2	2	3	2	2
	Imagen5	0	1	1	2	2	2	1	2	2	2	1	1
	Imagen6	0	1	1	2	2	2	1	2	2	2	1	1
	Imagen7	0	1	1	2	2	2	1	2	2	2	1	1
	Imagen8	0	1	1	2	2	2	1	2	2	2	1	1

(a)

PALABRA	# IMAGEN	FINGERS											
		F1	F2	F3	F4	F5	F6	F13	F14	F15	F16	F17	F18
ABAJO	Imagen1	1	1	1	1	1	1	1	1	0	1	1	2
	Imagen 2	1	2	2	1	1	1	2	2	1	1	1	2
	Imagen 3	1	1	1	1	1	1	1	1	0	1	1	2
	Imagen4	1	1	1	1	1	1	1	1	0	1	1	2
	Imagen5	1	1	1	1	1	1	1	1	0	1	1	2

(b)

LETRA	# IMAGEN	FINGERS											
		F1	F2	F3	F4	F5	F6	F13	F14	F15	F16	F17	F18
A	Imagen 1	1	1	1	0	0	0	0	1	1	1	0	0
	Imagen 2	1	1	2	1	1	0	0	1	1	2	1	1
	Imagen 3	1	1	1	0	0	0	1	1	1	0	0	0

(c)

Fig. 9. Matrices of structural values. **(a)** Number Matrix. **(b)** Word Matrix. **(c)** Letter Matrix.

5 Results

The results of the tests to verify that the contours that belong to a hand — by calculating structural points — are based on heuristic conditions where three types of points of convexity defects were considered: the starting point, (which is located where the outline of the hand starts), the convex hull, and the end point (which in this case is the last point of the series before the points once again converge). The *depth* variable is assigned to the end point — which determines the maximum distance between the contour and the convex-hull — to finally assign that point to the place of maximum distance iteratively, until concluding the recognition of the hand. It must be taken into account that different thresholds and logical operators were used. Another point to consider for the optimizing of the identification of the object is to calculate the minimum enveloping circle, which is used to determine the approximate center of the object of interest (hand). This based on coordinates of the Cartesian 'X', 'Y' Plane and numeric variables, in order to define the threshold conditions, [and in turn] in order to separate the "plane" by quadrants ('y' positive / negative and 'x' positive / negative). It is highly relevant to separate (by quadrants) the convexity points (patterns) by means of heuristics. All of the above based on the depth and the minimum enveloping circle, using the variables: *mycenter*, starting point (*ptStart*) and *depth*:

$$\text{if (depth >11 \&\& ptStart.y < mycenter.y).} \tag{8}$$

Where C, D, E and F indicate the depth, the reason for which the threshold is taken as higher than 11 (based on the image pixel count), the other depths being discarded, as to ptStart.y <mycenter. y, [it] will take the points higher than mycenter.y, in this case quadrants 1 and 2, after this other conditions will be used using the above variables, so that the convexity points can be used in all quadrants Furthermore — for better accuracy — we opted for the use of [either] a red or [an] orange glove, as these, in the HSV color space is easier to detect, additionally it can be distinguished from the environment where the video is being captured.

To achieve the recognition of the signs for letter, number and word, a set of values — which represent the structural points of the hand (F1, F2, F3, F4, F5, F6, F13, F14, F15, F16, F17 and F18) — were used, as shown in figure 10.

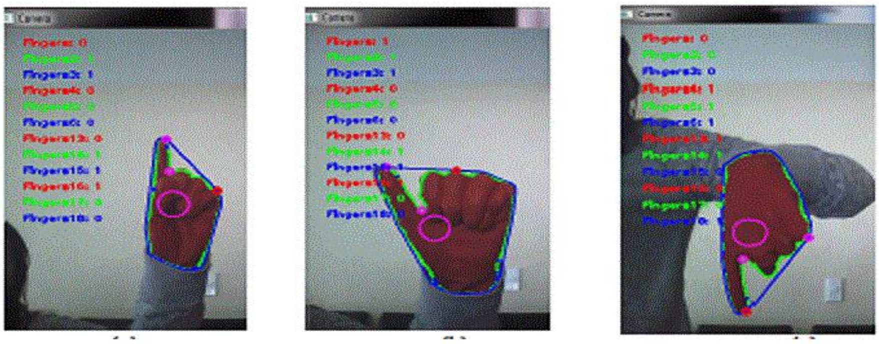


Fig. 10. Structural points. **(a)** Number. **(b)** Letter A. **(c)** Word Below (Abajo).

For the acquisition of the structural points a series of images is captured where the above mentioned, are distinguished (Figure 10)**,** such values are stored in an nX12 matrix, where n is determined when finding a repeatability or constant in the values, thus generating matrices with different numbers of rows, and 12 is the number of structural points used to limit the hand.

It should be mentioned that the functionality of the system will be taken as satisfactory by means of the percentage that its accuracy yields. Recognition is performed based on the counting of the fingers and the results report an accuracy percentage between 50 - 100%; percentages are based on executions that are carried out in the program, in this case 10 executions which were taken of which if the established sign was right in the 10 it is taken as 100%, if one of those 10 executions only 9 were right 90% taken and so on the accuracy percentage decreases according to the number of times it was right within these 10 runs. The results obtained for a total of 8 numbers, 20 letters of the alphabet and 25 words follow.

Table 1. Percentages of the test of numbers.

NUMBER	PERCENTAGE %
1	90
2	90
3	80
4	90
5	90
6	60
7	70
8	70

Average for the case of the analysis of the numbers

$P_N = \frac{640}{8} = 80\%$

Table 2. Average for the case of the analysis of letters.

LETTERS	PERCENTAGE %
A	90
B	80
C	80
D	70
E	60
F	70
G	60
H	60
I	60
L	70
M	80
N	70
O	70
P	60
R	50
S	60
T	60
U	70
V	70
W	80
Y	60

Average for the case of the analysis of letters:

$P_L = \frac{1430}{21} = 68\%$

Table 3. Average for the case of the analysis of words.

WORDS	PERCENTAGE
ABAJO	70
ADENTRO	60
CEREZA	70
CISNE	50
CONSUEGRO	50
COPA	60
HOGAR	50
JIRAFA	50
MADRE	50
MANO	50
MENOS	60
MESA	60
MUCHO	50
MUÑECA	50
OTOÑO	50
PESOS	50
PISTOLA	70
RESTA	70
SOLO	50
SUMA	50
TORTA	50
UÑA	50
VACA	70
YO	60

Average for the case of the analysis of words:

$$P_L = \frac{1350}{25} = 54\%$$

6 Conclusions

The automatic translation of sign language (*LMS*) is a tool capable of recognizing number, letters and words [expressed] in the *LSM* alphabet, at the end of the development [of the project] the system efficiency was found to have been limited by numerous difficulties that have been encountered during processing of the images or signs captured in real time, such as conducting a training [of the program] for the detection of all the words in sign language, which has forced us to define some parts of the human body that would more easily help identify those signs, but [in turn] would generate a greater degree of complexity in its implementation, [by] adding further convexity points of interest, or otherwise, [by] employing a data base, linked and synchronized to the capture of image data in real-time. Therefore, the use of both hands was restricted, and only the detection with one hand was considered.

As could be seen in the results of tables 1, 2 and 3, these values were very wide-ranging, because the patterns to be identified were becoming confusing by the increase in the signs to be recognized, for example in the analysis of the numbers, 8 cases were considered where additionally the contour patterns of these signs were well defined, [thus] allowing to obtain a percentage of reliability of approximately 80%; while in the [case of the] letters, in addition to having analyzed more cases (21), the percentage of reliability was lower, this because of several situations, an example is the case of the letter "R", where the middle and index fingers overlap, and the system confuses this pattern as a single point instead of two, because of the crossing of the fingers. In the case of the words "*consuegro*" (≈in law) and "*cisne*" (swan) the use of both hands is necessary, generating noise with the additional hand, for which reason the percentage of identification in those words was also low.

Another point to consider, is the case of the tracking of the frequency of the points, which was lost when performing the bending of a finger, since when straightening the finger there was no continuity of the tracking as in the case of the words "*muñeca*" (wrist/doll), "*suma*" (sum) and "*uña*" (finger/toe nail), to this a solution it was given [by] using convexity defects obtaining a more robust tracking in which point loss is less during the process. So the regions are divided by a convexity hull and a minimum enveloping circle, thus delimiting region by region of interest, from the fingertips to the lower part of the hand.

References

1. Kelly, D., McDonald, J., Markham, C.: A Person Independent System for Recognition of Hand Postures Used in Sign Language. Pattern Recognition Letters, pp. 31, 1359–1368 (2010)
2. Lopez, L. A., Rodriguez, R. M., Zamora, M. G., San Esteban, S.: My hands speak sign language for the deaf. 1st. Editing, editorial Trillas, Mexico, pp. 305 (2010)
3. Serafin, M. E., González, R.: Hands voice. Dictionary of Mexican Sign Language. Free access, National Council to Prevent Discrimination, Mexico, Available at: http://www.conapred.org.mx/documentos_cedoc/DiccioSenas_ManosVoz_ACCSS.pdf (2011)
4. Pavlidis, T.: Algorithms for Graphics and Image Processing. Computer Science Press, Rockville, Maryland (1982)
5. Chen, W. C.: Real-time palm tracking and hand gesture estimation based on fore-arm contour. Doctoral dissertation, PhD thesis, Master dissertation, Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan (2011)
6. Frank, N., Giraldo, M., González, Camargo, E.: Image Processing Algorithms for satellite with Hough Transform. Revista Visión Electrónica Año 5, No. 2, pp 26–41, July-December (2011)

ISSN 1870-4069

An Algorithm for Semantic Expansion of Queries in a Boolean Information Retrieval System

Ana Laura Lezama, Mireya Tovar, David Pinto, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla,
Mexico

yumita1102@gmail.com, {mtovar,dpinto,darnes}@cs.buap.mx

Abstract. The increased amount of information in different domains, complicating the quick access to a particular need or specific query of any person or user, so arises the need to expedite this need, where the initial query is sought within a set of a domain documents chosen by the user. To retrieve more documents is important to incorporate techniques that allow to get more documents with the desired information. In this case extend the original query with the synonyms of the words that compose it can help improve the accuracy of the retrieval system (RS). In this paper, we propose an algorithm for query expansion of a Boolean Information Retrieval System (BIRS), in which the queries are formed by the concepts of four domain ontologies. According to the experimental results, we observe an improvement in the accuracy of the BIRS with the query expansion.

Keywords: Information retrieval system, semantic query expansion, ontology.

1 Introduction

Information Retrieval (IR) is the area of the science and the technology that tries to acquire, represent, and store information [1]. IR is the discovery of documents, usually unstructured documents (generally text) that satisfies a need for information within large stored collections in computers, through the development and implementation of intelligent techniques such as incorporating a retrieval information model, use of ontologies, etc [2].

Information Retrieval used to be an activity that only one person was dedicated to do it as a librarian, secretaries, etc., but with the exponential growth of information (text), makes it hard for a person to retrieve some information in a manner quick. Now the world changed and the information retrieval is performed through information retrieval models, training corpus, techniques of queries expansion, etc [2].

Information Retrieval System (IRS) are designed for processing text in natural language, rarely structured, and usually of semantic ambiguous. They allow

pp. 69–76; rec. 2016-07-18; acc. 2016-09-19

information retrieval, previously stored, through performing a series of queries. They have an information retrieval model and an inverted index [3].

One model of information retrieval is the Boolean model based on set theory and Boolean algebra. In this initial model, the user specifies a its query a boolean expression that it consists of a series of terms commonly linked by boolean operators such as *and, or* and *not*. Given the logical expression of the query, the system will return those documents that meet and form the set of relevant documents. Thus, the system partitions the collection of documents in two sets, those that meet the specified condition (relevant), and those that do not comply (not relevant). A document is therefore simply relevant or not. The popularity of the Boolean model, especially in the beginning, is given by its simplicity both conceptually, for the clarity of its formalism, as at the implementation level. Moreover, since queries are formulated as a Boolean expression, with semantic highly accurate, the user knows the reason why a document has been returned by the system, this does not always happen in other more complex models. Furthermore, since the documents are bag-of-terms, the recovery process is extremely fast [4].

In this paper, we present an algorithm to expand queries in a Boolean retrieval system that is used to find evidence of the concepts and semantic relationships of four domain ontologies in the corpus. The aim is to find evidence in the corpus of the concepts that exist in the ontologies, which are then used for the evaluation of them [5].

This paper is organized as follows: The most related works to query expansion in BIRS are discussed in section 2. In section 3, the proposed algorithm is explained. The experimental results and the data set are shown in section 4. Finally in section 5, the conclusion and future work are provided.

2 Related Work

Information Retrieval Systems with query expansion are systems that incorporate techniques commonly used in information retrieval to improve the documents retrieved by reformulation or by expansion of the original query, either by adding new terms or by weighting the query terms original. The terms can be automatically extracted from documents or taken from a linguistic resource [6].

In contrast the above, information retrieval systems without query expansion are those that process the query made by the user, only the information retrieval system is able to interpret, and recover the relevant documents to the query. But given that not perform query expansion by any technique, the system is only able to retrieve information associated with such query as it was entered, and as result of not add an expansion technique, IRS does not have the ability to access the most relevant documents to the user [7].

Some authors have used several techniques of queries expansion, linguistic tools and information retrieval models. Below, we expose some research related with this work.

In Chunyan et al. [8] propose a query expansion method based on Markov technique, discovering a relationship of terms through the concepts of a tree Markov model. For the query expansion, a given query and a set of documents, the authors calculate the probability. The results obtained in their experiments were effective.

Olufade et al. [9] proposes an improved method of document clustering for automatic query expansion tagged concept based thesaurus network was extended for the authors in other to obtain from an optimal cluster. Each query or terms is represented as a matrix of documents where each column describes a document and each row a query. The Fuzzy Latent Semantic Query Expansion Model process achieved a better precision and recall rate values on experimentation and evaluations when they compared with others existing information retrieval approaches.

De Campos et al. [10] given a document collection, the authors built a thesaurus based on a bayesian network, their method learn a poly-tree of terms. The poly-tree nodes represent terms of the collection in the form of binary variables. Given a query submitted to their system, the query expansion process starts placing the evidences in the learned polytree. This action means looking for the terms that appear in the query in the poly-tree nodes and setting their states to term more relevant. As their network is a poly-tree, they can use an exact and efficient inference method to propagate the probabilities. The results obtained from their experiments realized, show a improvement of the retrieval effectiveness using their query expansion technique proposal.

Lozada et al. [11] propose user relevance feedback, introduced two algorithms of the query expansion based on the function $VP{-}IDF$, build the first algorithm of query expansion, this algorithm receives as input the original query and submit the original query with it is expansion with weight for all terms of the query. Designate that the perfil user is composed by the number of the documents than the users has been evaluated, the number the documents relevant evaluated and a list the terms of user, in the that registry each term than appear in the documents evaluated by the user. The second algorithm named CE-DF is the that complete the query of user, with the terms more relevant profile and gives with results a list of terms in a text string, similarity to the gives for the user.

Mata et al. [12] propose an algorithm for Polish ontologies for task of information retrieval and for the query expansion, the authors used two types of cross references SeeRelatedDescriptor and ConsiderAlso and of the hierarchical structure of the ontologie MeSH, the expansion is produced to level term, and the terms that is descriptors in MeSH expands with the content of element SeeRelatedDescriptor or ConsiderAlso. SeeRelatedDescriptor associated the descriptor with others descriptors related through cross references, their object is provide others descriptors. ConsiderAlso reference to others descriptors related through linguistic roots. Their third strategy of expansion, is based in the tree structure as the structure of MESH for its descriptors. Also UMLS is very wide than MeSH, and it does use of concepts instead of terms and the authors developed two strategies of expansion, the first, is expand the concepts through relations between concepts and the second the expansion it is realized through the concepts relations for relations between concepts. The experiments realized demonstrated that the use of the hierarchical structure as the MESH was effectiveness.

In contrast these authors, in this article we present the expansion of queries using synonyms. They are extracted from WordNet, that are processed and incorporated in the SRIB, it is able to recover not only the documents that contain the original query, but also documents that contain synonyms of that query. According to the experimental results, an increase in terms of the information retrieved, compared with the system that does not perform expansion was obtained.

3 The Proposed Algorithm

In this section, we propose a general algorithm to expand the queries. They are initially formed by the words that make up to the concepts extracted from each domain ontology. Subsequently, the queries are expanded with the corresponding synonyms of each word, these are obtained from WordNet. The expanded queries are used by the Boolean Information Retrieval System (BIRS). The algorithm is described below.

1. For each domain ontology to extract concepts and semantic relationships.
2. For every word that form part of the concepts to extract the synonyms corresponding from WordNet [13].
3. Pre-processing domain corpus, concepts, relationships and synonyms. This step involves the following:
   (a) The reference corpus is split into sentences.
   (b) To remove special characters, punctuation, numbers and empty words.
   (c) To apply the Porter algorithm to the information [14].
4. To build queries. There are three types of queries:
   (a) Queries formed with the words of the concepts.
   (b) Queries formed with the synonyms of each word that integrate to the concepts.
   (c) Queries formed with the words of the concept that form the semantic relationship.
5. To apply the Boolean Information Retrieval System to the concepts, without expansion.
6. To apply the Boolean Information Retrieval System to the synonyms of the concepts, with expansion.
7. To mix and to join posting list, i. e., the results gotten with the BIRS with synonyms and without them.
8. To apply of the AND operator to the query that includes the two concepts that form the semantic relationship. The AND operator performs the intersection of the sentences that make up the posting of both concepts that form the semantic relationship.

In case of the evaluation of the results gotten, we use the Equations (1) and (2) for measuring the precision at the level of concepts and relationships:

$$P_C = \frac{Recovered\ concepts}{Total\ concepts}, \tag{1}$$

$$P_R = \frac{Recovered\ relationships}{Total\ relationships}, \tag{2}$$

where: *Recovered concepts* is the total of concepts obtained by the BIRS, and *Total concepts* is the number total of concepts in the domain ontology. In the case of recovered relationships, we evaluate for separated the taxonomic relationship and non-taxonomic relationship (for more information see [5]). The *Total relationships* correspond to the number total of relationships of each type in the domain ontology evaluated independently.

4 Results

In this section, we present the datasets used (4.1) and the results obtained in the experiments carried out (4.2).

4.1 Datasets

In Table 1 we present the number of concepts (C), taxonomic relations (T) and non-taxonomic relations (NT) of the ontology evaluated. The characteristics of its reference corpus are also given in the same Table: number of documents (D), number of tokens (T), vocabulary dimensionality (V), and the number of sentences (O). The domains used in the experiments are Artificial Intelligence (AI), e-learning (SCORM) [15], Oil (OIL) and Tourism (Tourism).

Table 1. Datasets.

Domain	Ontology			Reference corpus			
	C	T	NT	D	T	V	O
AI	276	205	61	8	11,370	1,510	475
SCORM	1,461	1,038	759	36	1,621	34,497	1,325
OIL	48	37	-	577	546,118	10,290,107	168,554
Tourism	963	1,016	-	1,801	877,519	32,931	36,505

4.2 Experimental Results

Below, we present the experimental results obtained by the algorithm developed and its comparison, i. e., the results of BIRS without query expansion and BIRS with query expansion. The results obtained by both BIRS in the algorithm,

Table 2. Results of the proposed algorithm for concepts.

	CO	C	F	P	CE	FE	PE	O	OE	DI	%
AI	276	274	2	0.992	274	2	0.992	1,994	3,332	1,338	67.10
SCORM	1,461	1,434	27	0.981	1,436	25	0.982	23,406	41,461	18,055	77.13
OIL	48	48	0	1.00	48	0	1.00	232,603	298,752	66,149	28.43
Tourism	963	682	281	0.708	788	175	0.818	86,353	249,087	162,734	188.45

for the case of the concepts, are shown in the Table 2 for each revised ontology (domain). We also show the total number of concepts extracted from the ontology (CO), concepts recovered by the BIRS without expansion (C), concepts that did not obtain associated sentences (F) and precision (P); concepts recovered by the BIRS with expansion (CE), concepts that failed to recover the SRIB with expansion (FE) and precision obtained (PE). In addition, the table shows the number of sentences obtained by the BIRS without expansion (O), with expansion (OE), the difference in the number of sentences recovered with expansion and without expansion (DI) and the percentage increase (%). In base to the results obtained for the concepts, we note that in the case of domains SCORM and Tourism, the number of concepts recovered is higher than the result of the BIRS without expansion. Furthermore, the number of sentences that contain the synonyms of the concepts increases the amount of sentences associated for each concept in the ontology, this occurs in each domain. The percentage increase of the information retrieved by the BIRS with expansion is greater than 28%, indicating that the concept can be represented in the corpus by its synonym corresponding and that this information is additional to that presented by the BIRS without expansion.

In Table 3 shows the results obtained by both Information Retrieval Systems with expansion and without it for taxonomic relationships. The RT column corresponds to the number of taxonomic relationships included in the corresponding domain ontology. The RR column is the number of recovered taxonomic relationships with BIRS without expansion and with BIRS with expansion (RRE). The F column shows the difference of relationships recovered by the BIRS without expansion and with expansion (FE). The precision of the system without expansion (P) and with expansion (PE). The amount of sentences recovered by the BIRS without expansion (O) and with expansion (OE) for this type of relationship, the difference obtained (DI) and the percentage difference (%) are also included. In base to the results obtained we observe that the number of relations taxonomic for two ontologies are maintained by the two algorithms designed. But in the case of SCORM ontology, the number of concepts is incremented by one, while for the Tourism ontology the number of concepts is increased from 291 to 441 this indicates that there are more concepts in the corpus that can only be found by its corresponding synonym. Also, the amount of sentences retrieved with the BIRS with expansion is increased for the four ontologies and even more for the ontology of Tourism, it support the existence of the synonyms for the concepts found in the corpus newly.

In the case of the non-taxonomic relationships, that only AI and SCORM

Table 3. Results of the proposed algorithm for taxonomic relationship.

	RT	RR	F	P	RRE	FE	PE	O	OE	DI	%
AI	205	205	0	1.00	205	0	1.00	782	972	190	24.29
SCORM	1,038	1,002	36	0.965	1,003	35	0.966	10,640	15,897	5,257	49.40
OIL	37	32	5	0.864	32	5	0.864	12,696	13,410	714	5.62
Tourism	1,016	291	725	0.286	441	575	0.434	5,606	22,552	16,946	302.283

Table 4. Results of the proposed algorithm for non-taxonomic relationship.

	RNT	R	F	P	RE	FE	PE	O	OE	DI	%
AI	61	61	0	1.000	61	0	1.000	108	136	28	25.92%
SCORM	759	738	21	0.972	739	20	0.973	8,728	10,155	1,427	16.40%

have it, we observe that the amount of recovered relationships is the same by both systems.

The RNT column corresponds to the number of non-taxonomic relationships included in the corresponding domain ontology. The R column is the number of recovered non-taxonomic relationships with BIRS without expansion and with BIRS with expansion (RE). The F column shows the difference of relationships recovered by the BIRS without expansion and with expansion (FE). The precision of the system without expansion (P) and with expansion (PE). The amount of sentences recovered by the BIRS without expansion (O) and with expansion (OE) for this type of relationship, the difference obtained (DI) and the percentage difference (%) are also included. In this case, some sentences were increased in the results with this type of relationship (see Table 4).

5 Conclusions

After experiments, with the proposed algorithm, we observe that the BIRS with expansion increase the number of sentences recovered for the domain ontologies. Even more, the number of concepts or relationship found are incremented in some case. Therefore, to expand the query with synonyms is a good alternative to get better precision of the system. The domain of Tourism had a satisfactory behavior with this algorithm, because its ontology has synonyms in the corpus that it's not possible recover with a traditional retrieval system.

As future work, we propose the use of lexico-syntactic patterns for the extraction of alternative synonyms from the text, synonyms than WordNet probably has not stored.

Acknowledgments. This research work has been partially supported by Sectoral Research Fund for Education CONACyT project 257357, VIEP-BUAP project 00570 and PRODEP-SEP project 00570 (EXB-792) DSA/103.5/15/10854.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
2. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Query expansion in information retrieval systems using a bayesian network-based thesaurus. CoRR abs/1301.7364 (2013)
3. Kuna, H.D., Rey, M., Podkowa, L., Martini, E., Solonezen, L.: Expansión de consultas basada en ontologías para un sistema de recuperación de información. In: XVI Workshop de Investigadores en Ciencias de la Computación (2014)
4. Kwak, B.K., Kim, J.H., Lee, G., Seo, J.Y.: Corpus-based learning of compound noun indexing. In: Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11. pp. 57–66. Association for Computational Linguistics (2000)
5. Lozada, C.A.C., Mendoza, E.E., Becerra, M.E.M., Flórez, L.C.G., Guzmán, E.L.: Algoritmos de expansión de consulta basados en una nueva función discreta de relevancia. Revista UIS Ingenierías 10(1) (2012)
6. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
7. Mata, J., Crespo, M., Maña, M.J.: Estudio del uso de ontologías para la expansión de consultas en recuperación de imágenes en el dominio biomédico. Procesamiento del lenguaje natural 47, 39–46 (2011)
8. Miller, G.A.: WordNet: A Lexical Database for English, vol. 38. ACM, New York, NY, USA (1995)
9. Onifade, O.F., Ibitoye, A.O.: Fuzzy latent semantic query expansion model for enhancing information retrieval. International Journal of Modern Education and Computer Science 2, 49–53 (2016)
10. Porter, M.F.: Readings in information retrieval. chap. An Algorithm for Suffix Stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
11. Tovar Vidal, M., Pinto Avendaño, D., Montes Rendón, A., González Serna, J.G., Vilariño Ayala, D.: Evaluation of ontological relations in corpora of restricted domain. Computación y Sistemas 19(1), 135–149 (2015)
12. Vechtomova, O., Wang, Y.: A study of the effect of term proximity on query expansion. Journal of Information Science 32(4), 324–333 (2006)
13. Vilares Ferro, J.: Aplicación del procesamiento del lenguaje natural en la recuperación de información en español. Ph.D. thesis, Universidad da Coruña, Departamento de Computación (Mayo 2005)
14. Yuan, C.: Concept tree based information retrieval model. Journal of Multimedia p. 652 (2014)
15. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)

ISSN 1870-4069

A Comparative Analysis of Learning Techniques for Cancer Risk Prediction based on Medical Textual Records

Carolina Fócil-Arias, Grigori Sidorov, Alexander Gelbukh,
Miguel A. Sanchez-Perez

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación,
Mexico City, Mexico

focil.carolina@gmail.com, sidorov@cic.ipn.mx, www.gelbukh.com,
masp1988@hotmail.com

Abstract. In this paper, we compare the performance of a variety of machine learning algorithms, including supervised Naïve Bayes, J48, SVM, Random Tree, Random Forest, and non-supervised KNN for determining the type of cancer a patient is suffering using medical textual records. We train these classifiers on different sets of features such as unigrams and bigrams of words, character n-grams using tf-idf weighting scheme and binary feature representation. We evaluated performance of the classifiers in terms of accuracy, precision, recall, and F-measure. The obtained results show that Naïve Bayes and SVM achieve the best performance in this task.

Keywords: Cancer classification, medical records, supervised learning, SVM, Random Forest, KNN, Naïve Bayes, J48, natural language processing.

1 Introduction

Colon cancer is considered to be the major cause of death in the world [13]. According to [16], more than 1.2 million people are being diagnosed with this disease every year. Based on the information provided by American Cancer Society[1] in 2016, 95,270 cases of colon cancer are estimated with 49,190 death cases. Brain cancer has led to 16,050 deaths only in USA during the last year. Cancer is produced when an uncontrollable growth of cells occurs, and there is a spread of abnormal cells [6]. An early detection of these diseases allows significantly increasing survival rates.

One of the most commonly used approaches for classification of any type of data is based on machine learning (ML) techniques. In fact, the machine learning

[1] http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2016/ [last access: 17.07.2016].

pp. 77–88; rec. 2016-06-12; acc. 2016-09-19

strategy coupled with annotated corpora is considered the most efficient method known up to date to solve many natural language processing (NLP) tasks, such as authorship identification related tasks [9, 10, 14], plagiarism detection [21], tasks related to text similarity [23, 24], among many others.

In this work, we apply a machine learning approach using unigrams and bigrams of words and character n-grams as features to identify the type of cancer a patient is suffering from. We focus on two types of cancer: colon and brain cancer. We conducted experiments on a large dataset, which was collected and consolidated by Styler *et al.* [26]. We examined the performance of several machine learning algorithms and compared their results.

The rest of the paper is organized as follows. Section 2 presents several studies related to the usage of machine learning techniques for cancer detection. Section 3 describes the materials and methods used for determining whether a person has colon or brain cancer. Section 4 summarizes the results of using various machine learning algorithms. Finally, Section 5 draws the conclusions and points out possible directions of future work.

2 Related Work

In this section, we overview several works related to the prediction of cancer at the early stage. We focus on the studies that used machine learning approaches and consider several systems that were able to produce good results in clinical domain.

The system proposed in [12] is able to detect 20 common types of cancer, e.g., bronchus and lung, prostate, colon, breast, pancreas, etc., as well as whether cancer was the cause of death. The system is composed of two following stages: processing natural language pipeline for extracting features and using Support Vector Machines (SVM) algorithm. The authors obtained 94.2% in terms of F-measure.

Sparse Compact Incremental Learning Machine (SCILM) is a method that was proposed in [17] for cancer classification. This algorithm is based on a network structure on small dataset with a high number of dimensions, and it achieved an accuracy of 88.75%. Wang *et al.* [27] applied the same approach in order to identify breast cancer, the most common type of women cancer. The proposed method performed at the accuracy of 96.19%.

The approach proposed by Liu *et al.* [13] is able to detect cancerous colon tissues through the pattern recognition of spectra data. A total number of 60 colon tissues, which form two classes (normal and adenocarcinoma), were selected in order to perform the classification task. This method combines a large variety of features, and uses Principal Component Analysis (PCA) and Fisher's Discriminant Analysis (FDA). The method achieved the classification rate of 90.3%. The study of Rathore *et al.* [20] proposed a CBIC system, which performs a classification of colon cancer using SVM classifier.

As one can see, machine learning techniques are successfully used in medical domain. However, the question which classifier performs better in this task is still an open research question and the main motivation of the current work.

3 Materials and Methods

In this study, the idea is to classify whether a patient has symptoms associated with colon or brain cancer. The approach consists of seven following steps:

1. Accessing the clinical data obtained from Semantic Evaluation Exercises 2016 [26] (SEMEVAL is based on series of evaluations to explore meaning in language).
2. Data preprocessing for improving the quality of the dataset.
3. Extraction of features: unigrams and bigrams of words.
4. Construction of the vector space model.
5. Selection of machine learning algorithms based on the state of the art.
6. Evaluation of the results.
7. Comparison of the machine learning algorithms used to perform this task.

Figure 1 shows the methodology for conducting machine learning experiments.

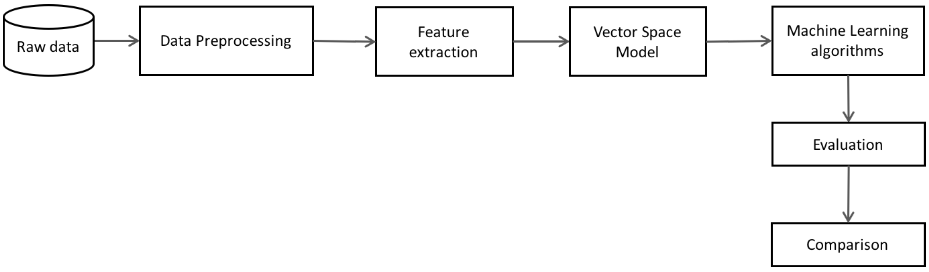


Fig. 1: Machine learning basic steps.

In our research, the corpus is divided into two subsets: training and test. A hold out approach was used to estimate the performance of machine learning classifiers. In this validation method, a certain amount for training and test data is reserved, and it is often used with independent test set [28].

3.1 Dataset

The dataset used in this research is the THYME (Temporal Histories of your Medical Events) [26] corpus, which was designed by clinic Mayo, University of Colorado, and the Harvard Medical School/Boston Children's Hospital. The corpus consists of a total number of 1200 documents describing 400 patients and is divided into two major diseases within oncology: colon and brain cancer.

Each patient is associated with three types of documents: clinical notes, radiology files, and pathology reports. However, in this research, we use only pathology reports, since they provide an important information concerning the patients [26]. For example, the confirmation of a benign or malignant tumor.

Thus, the number of pathology reports used in this research is 400, where 200 documents correspond to brain cancer and 200 correspond to colon cancer.

To the best of our knowledge THYME corpus is used for a classification task for the firs time.

3.2 Preprocessing Steps

An important step for improving data quality is data preprocessing, which has been used in several natural language processing tasks [10, 15, 18, 25] in order to increase the efficiency of the classifiers. Unlike conventional data preprocessing methods applied in the field of pattern recognition, the area of natural language processing (NLP) proposes alternative techniques, such as stop words extraction, stemming, detection of sections, etc. to enhance the quality of input data representation to be fed into a machine learning algorithm.

In our case, first, each pathological document in the corpus was segmented into sections. Then, we expanded the contractions (e.g., can't → cannot, I've → I have, etc.), which was required to normalize certain words. Afterwards, word span tokenizer [3] was applied using regular expressions, since some words were not correctly separate by a white space. Then, all the stop words were omitted, since they are used in all of the documents and do not provide useful information for classifiers. For instance, we removed the words "the", "my", "can", "ever", "by", "of", "now", and others that showed a high frequency in all the documents.

3.3 Classification

In this study, to determine the importance of a word in a document, the metric known as *tf-idf* is used. The *tf* (term frequency) value is the frequency of the word in document, when *idf* (inverse document frequency) is the inverse proportion of the frequency of the word in a set of documents [29]. We also try binary feature representation, that is, whether a feature exists or no exists in the corpus.

We examine five machine learning algorithms: Naïve Bayes, SVM, Random Tree, J48, Random Forest, and KNN with $K = 3$. According to [8, 27], the most commonly used algorithms in the field are SVM and KNN. However, we also examine Random Forest, Random Tree, Naïve Bayes, and J48, since these algorithms are considered among the best ones to tackle classification tasks.

Machine learning is a branch of artificial intelligence which aims at detecting meaningful patterns in data and is based on statistics and computer science. This field is divided into several subfields dealing with different types of learning tasks: supervised, unsupervised, active, passive and others [22].

The algorithms used in this study employ supervised learning. This means that the algorithm can estimate the success of prediction using the labeled training data. A brief description of each of the classifiers is presented below.

Naïve Bayes is commonly used in classification tasks. This algorithm belongs to the probabilistic classifiers, where features are conditionally independent. This algorithm ignores possible dependencies among the inputs and reduces a multivariate problem to a group of univariate problems [1]. It is based on Bayesian theorem, which shows a relation between marginal probabilities and conditional probabilities [11].

Support Vector Machine (SVM) is a powerful classification algorithm, which belongs to linear model. SVM allows separating the data when it is tend to be linearly or non-linearly separable through a linear decision surface (hyperplane) based on the maximum distance found between the surface and the nearest points of the two classes [20]. The distance between the hyperplane and the closest examples should be maximized to separate its input space into two classes. Also, this algorithm can be used in a non-linear classification using a non-linear kernel, which is a mathematical function that transforms input data to a high dimensional feature space [7].

J48 belongs to decision tree family. It is considered being a powerful classifier and hierarchical structure for supervised learning. This classifier can be used for both classification and regression, even though they it is more frequently for classification [1]. The learning process is based on splitting the labels of training data into subsets according to statistical tests under divide-and-conquer heuristic [1, 22].

Random Forest [4] consists of many decision trees for making a decision based on a response of each decision tree. This algorithm is able to handle the missing values, as well as to compute generalization errors, and to identify relevant variables. Also, it is considered as a potential algorithm for building classifiers due to the selection of a random subset of input features [6, 19]. It has shown to work well on a large corpus with a large number of features. This classifier has a collection of decision trees [22] and combines multiple Random Tree, which is built with a number of random features and stochastic process [28].

K-Nearest Neighbor is a non-supervised classifier based on pattern recognition. The basic idea is to classify a new pattern with the most probable class according to its k nearest neighbors [2].

A simplified representation of several machine learning algorithms is given in Figure 2. Table shows the advantages and limitations of the algorithms used in this study [5, 7, 22].

A pictorial depiction of the classification task performed in this study is shown in Figure 3.

4 Experimental Results

The aim of this study is to determine with a high accuracy the type of cancer a patient is suffering from. This means whether he/she has colon or brain

Table 1: A comparison among the algorithms used in this study.

Supervised algorithm	**Advantages**	**Limitations**
SVM	– Lower risk of over-fitting. – Can achieve a nonlinear separating hyperplane. – Computational complexity reduced to quadratic optimization problem.	– Training can be slow. – Difficult when training data is not linearly separable. – The structure of algorithm is difficult. – Lack the transparency of results. – Speed and size for both training and test. – Selection of kernel function parameters. – High complexity and extensive memory requirements.
KNN $K = 3$	– Fast classification of instances. – The cost of learning process is zero. – The local approximation is used to learn complex concepts. – Tolerant with missing values and noise. – Assumes similar classification when the instances have similar features. – Can be used with categorical features.	– Computationally expensive complex when number of attributes increases. – The performance depends on the number of dimensions. – Assumes that attributes will be equally relevant. – Slower to update.
Naïve Bayes	– Efficient training algorithm. – Consider the relationships between attributes. – Handles discrete, real data and streaming. – Fast to classify instances. – Irrelevant attributes do not affect the performance.	– Assumes independence of features. – Classes must be mutually exclusive. – Frequency of attributes and classes can affect the performance.
Decision Tree	– Very intuitive predictors. – Very simple to understand and to interpret – Discover nonlinear relations and interactions. – Can generate rules for helping the knowledge. – Are not affected by outliers.	– Computationally hard to learn. – Not guarantee to return the globally optimal decision tree. – Can be complex and time consuming with large decision tree. – Large trees are not intelligible. – The cost of analysis can be an expensive option.

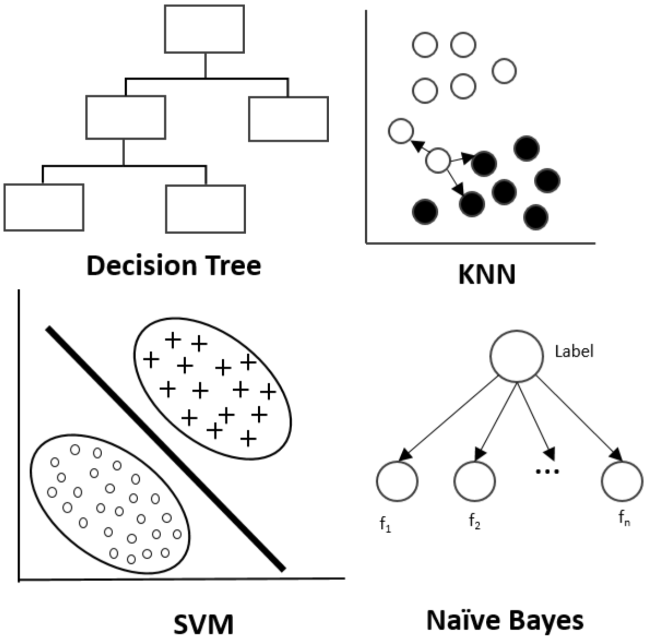


Fig. 2: An example of the algorithms used in this study.

cancer according to pathological reports [26]. The evaluation of the classifiers was carried out in terms of the following metrics: accuracy (A) provides the number of instances that are correctly predicted; precision (P) is the number of retrieved documents that are relevant; recall (R) gives the number of relevant documents that are retrieved; F-measure (F1) denotes a combination of precision and recall [28].

The experiments were carried out on the test set using the well-known data mining tool, WEKA (Waikato Environment for Knowledge Analysis) [28], which has a large collection of implemented machine learning algorithms to perform classification tasks.

We conducted four series of experiments. The first set of experiments consisted in using bag-of-words (BoW) approach with a total 5,860 features. Based on this approach, Naïve Bayes and J48 achieved a classification accuracy of 100.0% as can be seen in Table 2. SVM, Random Forest, Random Tree and KNN also showed high accuracy of 99.01%, 99.01%, 97.03% and 94.06%, respectively. For classification task, a proximity baseline was used. Each document was predicted to be "Brain cancer", which is the majority class.

In the next stage of experiments, we used bigrams of words as features to perform the task. The total number of such features for our dataset is 22,501. The usage of bigrams showed the same accuracy scores for predicting two types of cancer as when using unigrams of words as features. Based on the classifier

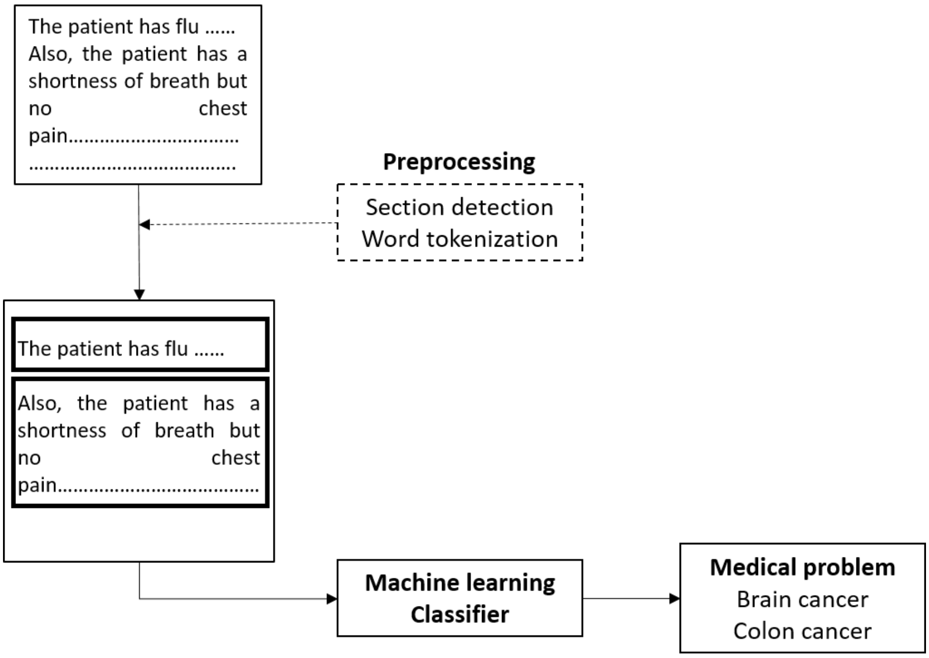

The patient has flu ......
Also, the patient has a shortness of breath but no chest pain.................................
....
Preprocessing
Section detection
Word tokenization
The patient has flu ......
Also, the patient has a shortness of breath but no chest pain.......................................
Machine learning Classifier
Medical problem
Brain cancer
Colon cancer

Fig. 3: An example of document review process.

Table 2: Classification results using bag-of-words (BoW) model and tf-idf weighting scheme.

Classifier	Acc., %	Precision		Recall		F-measure	
		Brain	Cancer	Brain	Cancer	Brain	Cancer
Naïve Bayes	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
J48	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Forest	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Tree	97.03	98.00	96.20	96.00	98.00	97.00	97.00
KNN $K = 3$	94.06	89.30	100.00	100.00	88.20	94.30	93.06
Baseline	0.5016	1	0	0.5016	0	0.6680	0

performance on the test data, Naïve Bayes and J48 achieved 100.0% of accuracy for detecting two types of cancer (colon and brain), followed by SVM, Random Forest, Random Tree and KNN (see Table 3) with 99.01%, 98.02%, 95.05% and 81.19%, respectively.

Next, another set of experiments are conducted using character n-grams ($n = 3$) as features. The results for this experiment are shown in Table 4. Here, Naïve Bayes and J48 outperformed the other classifiers, achieving 100% of accuracy. SVM, KNN, and Random Forest showed 99.01% of accuracy followed by Random Tree with 97.03% of accuracy.

Table 3: Classification results using bigrams of words as features and tf-idf weighting scheme.

Classifier	Acc., %	Precision		Recall		F-measure	
		Brain	Cancer	Brain	Cancer	Brain	Cancer
Naïve Bayes	**99.01**	**98.00**	**100.00**	**100.00**	**98.00**	**99.00**	**99.00**
SVM	**99.01**	**98.00**	**100.00**	**100.00**	**98.00**	**99.00**	**99.00**
Random Forest	**99.01**	**98.00**	**100.00**	**100.00**	**98.00**	**99.00**	**99.00**
J48	98.02	96.20	100.00	100.00	96.10	98.00	98.00
Random Tree	95.05	94.10	96.00	96.00	94.10	95.00	95.00
KNN, $K = 3$	81.19	72.50	100.00	100.00	62.70	84.00	77.10

Table 4: Classification results using character-level trigrams and tf-idf weighting scheme.

Classifier	Acc., %	Precision		Recall		F-measure	
		Brain	Cancer	Brain	Cancer	Brain	Cancer
Naïve Bayes	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
J48	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
SVM	99.01	98.00	100.00	100.00	98.00	99.00	99.00
KNN $K = 3$	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Forest	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Tree	97.03	94.30	100	100.00	94.10	97.10	97.00

Finally, another set of experiments was conducted using bag-of-words and binary feature representation (see Table 5). J48 yielded the best accuracy of 100%, followed by Naïve Bayes, SVM, Random Forest, and Random Tree with an accuracy of 99.01%, 99.01%, 99.01%, 95.04% and 78.21%, respectively.

Table 5: Classification results using bag-of-words model and binary feature representation.

Classifier	Acc., %	Precision		Recall		F-measure	
		Brain	Colon	Brain	Colon	Brain	Colon
J48	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Naïve Bayes	99.01	98.00	100.00	100.00	98.00	99.00	99.00
SVM	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Forest	99.01	98.00	100.00	100.00	98.00	99.00	99.00
Random Tree	95.04	90.90	100.00	100.00	90.20	95.20	94.80
KNN, $K = 3$	78.21	69.40	100.00	100.00	56.90	82.00	72.50

The results presented in Tables 2 – 5 indicate that it is possible to identify the type of cancer from the pathology reports using either unigrams, bigrams of words, or character n-grams as features for machine learning algorithms.

As can be seen in Figure 4, the results are high for the majority of the examined classifiers. It can be explained by the fact that we considered just two classes representing colon and brain cancer. The very interesting result is that character trigrams with unsupervised KNN algorithm obtained practically the same results as the supervised algorithms (99%). It shows the importance of character n-grams for these type of tasks.

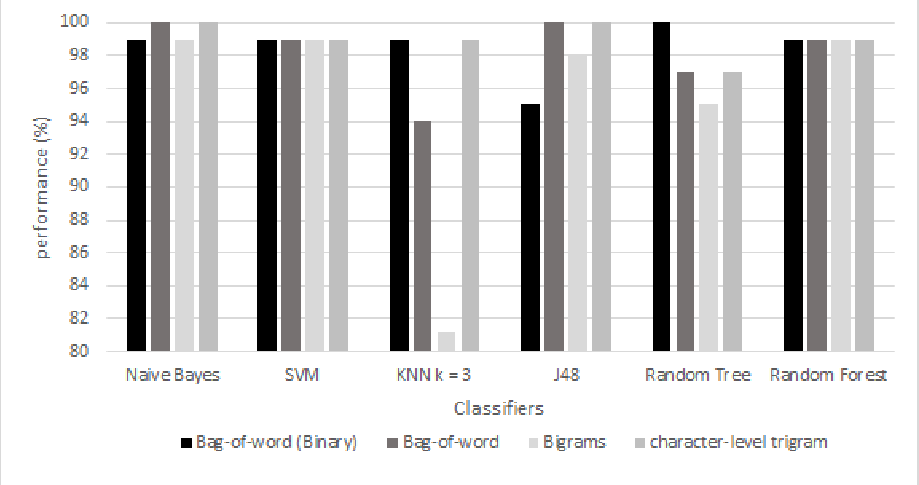


Fig. 4: Results in terms of accuracy.

5 Conclusions and Future Work

The main goal of this study consisted in predicting two deadliest types of cancer in the world: colon and brain cancer. The effectiveness of six machine learning classifiers (J48, Naïve Bayes, SVM, Random Forest, Random Tree, and KNN) was examine using different features and feature representations.

We empirically showed that these algorithms are effective for classification task of the leading types of cancer in the world. Naïve Bayes and J48 produced the highest results in the experiments, achieving an accuracy of 100% for detecting colon or brain cancer, when bag-of-words (BoW) and character n-grams with tf-idf weighting scheme were used. However, J48 generated the best result with 100 % when bag-of-words model and binary representation is used followed by Naïve Bayes, SVM, Random Forest with 99.01% of accuracy.

Overall, we believe that KNN uses all features for determining the class of a patterns assuming that attributes are equally relevant, unlike decision tree algorithm and Naïve Bayes, which use features that distinguish a disease. This means that Naïve Bayes reduces the number of parameters that must be estimated to learn. It is very interesting that the performance of KNN increases drastically using character trigrams as features (from 84% ot 99%).

The focus of this research aimed at predicting cancer risk using a corpus with pathological information without regard to the limitations in the corpus, which has two classes. This study can provide a great help to physicians to detect colon and brain cancer at the early stage, which can contribute to assign a curative treatment on time and save lives.

In future work, we will conduct experiments with more classes to make the task more challenging, that is, we will include other types of cancer to be classified. Furthermore, we will apply Latent Semantic Analysis (LSA) in order to reduce the number of dimensions in the vector space model.

References

1. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, 2nd edn. (2010)
2. Bhuvaneswari, P., Therese, A.: Detection of cancer in lung with k-nn classification using genetic algorithm. Procedia Materials Science 10, 433–440 (2015)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc., 1st edn. (2009)
4. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (Oct 2001), http://dx.doi.org/10.1023/A:1010933404324
5. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2(2), 121–167 (Jun 1998), http://dx.doi.org/10.1023/A:1009715923555
6. Chen, H., Lin, Z., Wu, H., Wang, L., Wu, T., Tan, C.: Diagnosis of colorectal cancer by near-infrared optical fiber spectroscopy and random forest. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy 135, 185–191 (2015)
7. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Informatics 2, 59–77 (2006)
8. Durgalakshmi, B., Vijayakumar, V.: Progonosis and modelling of breast cancer and its growth novel naïve bayes. Procedia Computer Science 50, 551–553 (2015)
9. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Fócil-Arias, C.: Compilación de un lexicón de redes sociales para la identificación de perfiles de autor [Compiling a lexicon of social media for the author profiling task] (in Spanish, abstract in English). Research in Computing Science 115 (2016)
10. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience 2016 (2016)
11. Karabatak, M.: A new classifier for breast cancer detection based on naïve bayesian. In: Measurement: Journal of the International Measurement Confederation. vol. 72, pp. 32–36 (2015)
12. Koopman, B., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Automatic icd-10 classification of cancers from free-text death certificates. International Journal of Medical Informatics 84(11), 956–965 (2015)
13. Liu, L., Nie, Y., Lin, L., Li, W., Huang, Z., Xie, S., Li, B.: Pattern recognition of multiple excitation autofluorescence spectra for colon tissue classification. Photodiagnosis and Photodynamic Therapy 10(2), 111–119 (2013)

14. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: Adapting cross-genre author profiling to language and corpus. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, vol. 1609, pp. 947–955. CLEF and CEUR-WS.org (2016)
15. Meystre, S., Haug, P.: Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. Journal of Biomedical Informatics 39(6), 589–599 (2006)
16. Mohammed, A., El-tanni, H., El-khatib, H., Mirza, A., El-kashif, A.: Molecular classification of colorectal cancer: Current perspectives and controversies. Journal of the Egyptian National Cancer Institute (2016)
17. Nayyeri, M., Sharifi Noghabi, H.: Cancer classification by correntropy-based sparse compact incremental learning machine. Cold Spring Harbor Labs Journals (2015)
18. Olson, D.L., Delen, D.: Advanced Data Mining Techniques. Springer Publishing Company, Incorporated, 1st edn. (2008)
19. Rastghalam, R., Pourghassem, H.: Breast cancer detection using mrf-based probable texture feature and decision-level fusion-based classification using hmm on thermography images. Pattern Recognition 51, 176–186 (2014)
20. Rathore, S., Hussain, M., Aksam Iftikhar, M., Jalil, A.: Ensemble classification of colon biopsy images based on information rich hybrid features. Computers in Biology and Medicine 47(1), 76–92 (2014)
21. Sánchez-Pérez, M., Sidorov, G., Gelbukh, A.: The winning approach to text alignment for text reuse detection at pan 2014, in working notes of clef 2014. In: Conference and Labs of the Evaluation forum. CEUR Workshop Proceedings, vol. 1180, pp. 1004–1011. CEUR (2015)
22. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York, NY, USA (2014)
23. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Detección automática de similitud entre programas del lenguaje de programación Karel basada en técnicas de procesamiento de lenguaje natural [Automatic detection of similarity of programs in Karel programming language based on natural language processing techniques (in Spanish, abstract in English)]. Computación y Sistemas 20(2), 279–288 (2016)
24. Sidorov, G., Ibarra Romero, M., Markov, I., Guzman-Cabrera, R., Chanona-Hernández, L., Velásquez, F.: Measuring similarity between Karel programs using character and word n-grams. Programming and Computer Software 43, (in press) (2017)
25. Singh, G., Samavedham, L.: Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of parkinson disease. Journal of Neuroscience Methods 256, 30–40 (2015)
26. Styler, W., Savova, G., Palmer, M., Pustejovsky, J., O'Gorman, T., Groen, P.: Thyme annotation guidelines (2014)
27. Wang, P., Hu, X., Li, Y., Liu, Q., Zhu, X.: Automatic cell nuclei segmentation and classification of breast cancer histopathology images. Signal Processing 122, 1–13 (2016)
28. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
29. Yun-tao, Z., Ling, G., Yong-cheng, W.: An improved tf-idf approach for text classification. Journal of Zhejiang University-SCIENCE A 6(1), 49–55 (2005)

ISSN 1870-4069

Entropy of Eye Fixations: a Tool for Evaluation of Learning Objects

Carlos Lara-Alvarez[1], Hugo Mitre-Hernandez[2], Maria Alvarado-Hernandez[2]

[1] CONACYT Research Fellow – Centro de Investigación en Matemáticas (CIMAT),
Zacatecas, Mexico

[2] Centro de Investigación en Matemáticas (CIMAT),
Laboratorio de Interacción Humano-Computadora,
Zacatecas, Mexico

carlos.lara@cimat.mx, {hmitre,maria.alvarado}@cimat.mx

Abstract. Learning objects (LOs) are important information resources that support traditional learning methods. To evaluate the impact, effectiveness, and usefulness of learning objects it is necessary a theoretically, reliable, and valid evaluation tool. This paper presents a metric to compare the design of LOs, it uses the information provided by visual fixations measured from a small focus group. We conducted an experiment with children of elementary school (n=23). Results showed that images with higher values of the proposed metric were faster to read (Mean = 0.629 min/image) than those images of LO selected randomly (Mean = 0.782 min/image). The metric is useful to optimize the fluency, this is an important step through obtain a fully automated tool to evaluate LOs.

Keywords: Eye tracking, evaluation of learning objects, entropy.

1 Introduction

There are many definitions of Learning Objects (LOs) in the literature [1,10]. For the purposes of this paper the following definition is adopted:

> *"Learning objects are information resources or interactive software used in online learning [10]".*

A single image, a page of text, an interactive simulation, or an entire course could all be examples of learning objects. Thousands of LOs are currently available through the web [10]; hence, it is necessary an automatic tool for evaluating the impact, effectiveness, and usefulness of learning them.

In this paper we are interested on evaluating LOs composed of a sequence of images. The order and content of these images are designed to ensure a given learning objective. When designing such a learning object, every person on the design team could have different ideas of what the student needs or wants.

Let us suppose that the content and sequence of a LO is already defined; but, for every position in the sequence there are several options with different

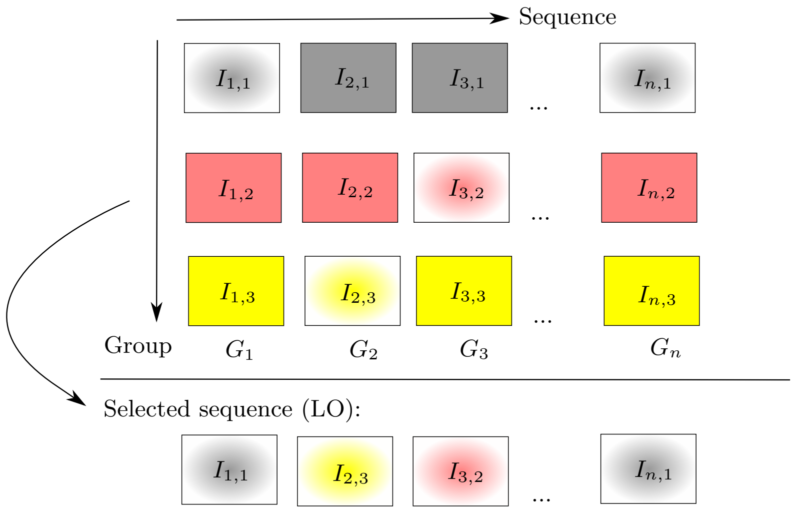


Fig. 1. For each group G_i the designer must select the best image to fill the i–th position of the LO. In this example, image $I_{1,1}$ was selected from $\{I_{1,1}, I_{1,2}, I_{1,3}\}$ to fill position 1 of the LO, image $I_{2,1}$ to fill position 2, etc. This paper proposes an entropy-based metric calculated from eye fixations to select the best image.

graphical design. This problem is illustrated in Fig. 1; for instance, the group G_1 is composed of three images; i.e. $G_1 = \{I_{1,1}, I_{1,2}, I_{1,3}\}$. One of these images must be selected to fill the first position in the LO. Analogously, one image of the group $G_2 = \{I_{2,1}, I_{2,2}, I_{2,3}\}$ must be selected to fill the second position in the LO, and so on. The problem can be stated as follows:

Given a set of candidate images:

$$G_i = \{I_{i,j} \mid j = 1 \cdots, N\},$$

where all images $I_{i,j} \in G_i$ has the same information but different design, select the image $\hat{I}_i \in G_i$ that optimize certain predefined evaluation criteria.

The main contribution of this paper is an entropy-based metric able to compare the images of a group. The proposed metric uses the information provided by visual fixations measured from a small focus group; hence, it does not require the experts' intervention, and the information is more accurate as it is originated directly from students. The empirical evaluation shows that images selected by this metric are faster to read.

The rest of this paper is organized as follows: section 2 discusses previous approaches to evaluate LOs, section 3 introduces the entropy-based metric, section 4 presents an experiment with students from primary school, section 5 presents the results and discusses the pros and cons of the proposed technique; finally, section 6 concludes this article.

2 Related Work

In general, LO evaluation approaches can be categorized as:

Indirect methods. These approaches consider that experts, aided with guidelines or other evaluation instruments, can improve the design and content of LOs.

Direct methods These approaches obtain information of the quality of LO from the focus group. A simple strategy is using questionnaires or surveys. Technological advances had open the possibility of automating the evaluation of LOs. A widely used sensor for this purpose is the eye tracker device commonly known as eye-tracker that is capable of obtaining data of when and what the user views on a screen.

There are several studies that suggest guidelines that can be used for indirect evaluation; for instance, Ozcelik et al. [5] study how color coding affects multimedia learning, they suggest that color coding can reduce unnecessary search processes in comparison to black and white material because it guides the student attention by salient information. The Learning Object Review Instrument (LORI) [10] is aimed to evaluate different aspects of a LO; reviewers can rate and comment with respect to nine items: content quality, learning goal alignment, feedback and adaptation, motivation, presentation design, interaction usability, accessibility, reusability, and standards compliance. The metric proposed in this paper is focused on the presentation design; this aspect evaluates the design of visual and auditory information for enhanced learning and efficient mental processing.

An example of a direct instrument is the Learning Object Evaluation Scale for Students (LOES-S) which includes three main categories: learning, quality, and engagement [2]. In this way, after using a LO, students complete a survey to determine their perception of (a) how much they learned, (b) the quality of the learning object, and (c) how much they were engaged with the learning object. We consider that it is possible to generate an automatic method that evaluates these three aspects. The first aspect is easily evaluated through the student's achievements, the second aspect requires to know the emotional state of the player [6] or it can be inferred by the analysis of the data usage [7]. The proposed approach is an effort towards a direct evaluation of the material quality.

Several studies on eye tracking supports the metric proposed in this paper. Tsai et al. [9] state that successful problem-solving students tend to spend more fixation time on inspecting relevant than irrelevant factors. Vatrapu et al. [11] suggest that observation times are longer for harder to understand representations. Considering these facts, we assert that a good design must have an equal distribution of observation times along the image; thus, avoiding hard-to-understand or irrelevant zones. In this sense, an entropy-based metric, such as the proposed in this paper, is a good alternative.

3 Proposed Approach

Entropy can be described qualitatively as a measure of energy dispersal. The concept itself is linked to disorder: entropy is a measure of disorder, and nature tends toward maximum entropy for any isolated system. The information entropy is defined as:

$$H = -\sum_i p_i \log p_i, \tag{1}$$

where p_i is the probability of occurrence of the i-th symbol of an alphabet.

To obtain an entropy-based metric, a grid that covers the region of interest is superimposed on the image. This representation adjusts to a variety of elements and structures that may contain the image (e.g. text or pictures). Because of the size of objects and their spatial relations among them are unknown, two strategies were implemented:

Scale-space. The concept of octave was implemented in order to obtain a robust metric that is invariant to changes of objects' size. Given the initial grid size of $n \times m$ cells, each octave is obtained by multiplying the original grid size by $k = 2^{s-1}$. That is, the grid size G_s for the s-th octave is

$$\begin{aligned} G_s &= m_s \times n_s \\ &= 2^{s-1}m \times 2^{s-1}n. \end{aligned} \tag{2}$$

For instance, given the grid of original size $G_1 = 2 \times 3$ shown in Fig. 2a, the grid for the octave 2 is $G_2 = 4 \times 6$ (Fig. 2b), and the grid for the octave 3 is 8×12 (Fig. 2c).

The cell size for an image of $x \times y$ pixels and a grid of m_s rows and n_s columns is

$$S_x = x/n_s, \tag{3}$$

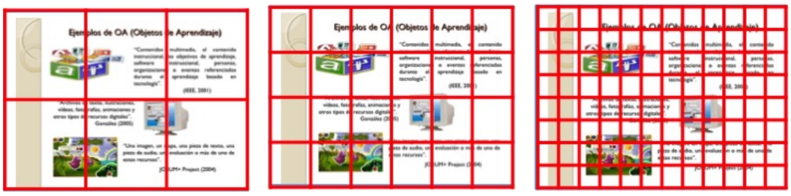
$$S_y = y/m_s. \tag{4}$$

Sliding window. This concept is used because the position of objects is unknown. The grid is successively moved in both directions using increments of:

$$\Delta_x = S_x/p, \tag{5}$$

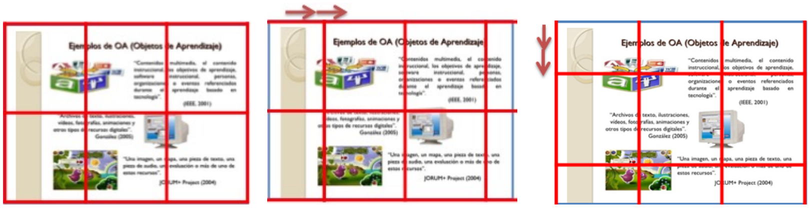
$$\Delta_y = S_y/p, \tag{6}$$

where p is the predefined number of steps. As shown in Fig. 3 every pixel in the region of interest is always covered by a single cell. Henceforth, G_s^{hv} denotes a grid of size $G_s = m_s \times n_s$ moved h and v steps in horizontal and vertical directions, respectively.



(a) (b) (c)

Fig. 2. Illustration of the scale-space strategy: (a) Grid of original size $G_1 = 2 \times 3$ (b) second octave, $G_2 = 4 \times 6$, and (c) third octave $G_3 = 8 \times 12$.



(a) (b) (c)

Fig. 3. Illustration of the sliding window strategy: (a) grid at the initial position, (b) grid displaced horizontally, (c) grid displaced vertically.

Entropy-based metric using eye tracking data

Given the observation time over an image (using an eye tracking sensor) and a grid G_s^{hv}, the entropy is calculated as

$$H(G_s^{hv}) = -\sum_{i=1}^{m_s}\sum_{j=1}^{n_s} p(A_{ij}) \log p(A_{ij}), \tag{7}$$

where A_{ij} is the grid cell at row i and column j, and $p(A_{ij})$ is:

$$p(A_{ij}) = \frac{t_{ij}}{\sum_i \sum_j t_{ij}}, \tag{8}$$

here t_{ij} is the total the time cell A_{ij} was observed by the student.

Eye entropy for scale s is

$$H_{\text{eye}}(s) = \frac{1}{p^2}\sum_{h=0}^{p-1}\sum_{v=0}^{p-1} \frac{H(G_s^{hv})}{H_s'}, \tag{9}$$

where $H_s' = -\log \frac{1}{n_s m_s}$ is the maximum entropy for a grid of size $n_s \times m_s$.

The process to calculate the entropy of an Eye tracking data is described in Algorithm 1; finally, Algorithm 2 describes the process of consolidating the LO.

```
Algorithm 1: EyeEntropy(G,p,$\tau$, E)
Input: $G = m \times n$: initial grid size; $p$: number of steps, $\tau$: minimum cell size, $E$:
       eye tracking data.
Output: $H_{eye}$: entropy
s ← 1
$H_{eye} \leftarrow 0$
Calculate $G_s$ (eq. 2)
while $\tau < \min(m_s, n_s)$ do
    $H_{eye}(s) \leftarrow$ Calculate entropy for scale $s$ (eq. 9 )
    $H_{eye} \leftarrow H_{eye} + H_{eye}(s)$
    $s \leftarrow s + 1$
    Update $G_s$ (eq. 2)
end
return $H_{eye}$
```

```
Algorithm 2: Consolidating LO by using $H_{eye}$
Input: $[G_1, G_2, \dots G_n]$: a proposed sequence, where all images $I_{i,j} \in G_i$ has the
       same information but different design.
Output: $\hat{L}$: consolidated LO
Initialize grid size $G$, the minimum cell size $\tau$, and the number of steps $p$
foreach $i \in \{1, \dots n\}$ do
    foreach $j \in \{1, \dots, |G_i|\}$ do
        $H_j \leftarrow 0$
        foreach participant $p$ do
            $E \leftarrow$ Eye tracking data of participant $p$ observing image $I_{ij}$
            $H_j \leftarrow H_j +$ EyeEntropy($G$,$p$, $\tau$,$E$)
        end
    end
    $\hat{L}[i] \leftarrow$Select the image $I_{ij} \in G_i$ that maximize $H_j$
end
return $\hat{L}$
```

4 Materials and Methods

The aim of this research is to assess whether H_{eye} is useful as a mechanism to select the best design of a LO.

An eye-tracking device type "Eye Tribe" model ET1000 with 60Hz sampling frequency was used in a screen with 1440 × 960 pixels resolution. Eye tracker device was located at a distance of 50cm from the student's face. The device calibration was made with OGAMA [12] (using 12 calibration points). In addition, this tool was used for generating the image sequences for evaluation.

For the calculation of the entropy-based metric, the full image was taken as region of interest and a grid of 2 × 3 grid was used as initial configuration.

4.1 Participants

Thirty-two fifth grade children from the "Pedro Coronel" Elementary School in Zacatecas, Mexico participated in this study. Nine of them participated in the design phase of LOs, and the other 23 participated in the testing phase.

4.2 Learning objects

Two sets of images were used in this study: "legends" and "theater". In the design phase, two LO were generated from each image set:

H_{eye} **LOs.** Two learning objects (one for "legends" and another for "theater") were generated by following the strategy illustrated in Fig. 1 and using the proposed metric described in (9). The image at each position was selected from a set of three equivalent images. For this aim, each sequence was composed from eye tracking data generated by nine students (each one observing one of the three possible images). This process and some images are shown in Fig. 4.

Random LOs. Two learning objects (one for "legends" and another for "theater") were selected randomly.

Every LO used in the experiment is a sequence of six images; in every case, the predefined order of the instructional content was preserved. At the end of each LO, a five item questionnaire was included to know the student's achievements.

4.3 Procedure

For the testing phase, participants were allocated in two groups as follows:

Group 1 Eleven children studied the H_{eye} LO of the "legends" theme, followed by the random LO of the "theater" theme.

Group 2 Twelve children studied the random LO of the "legends" theme, followed by the H_{eye} LO of the "theater" theme.

Each session took approximately 30 minutes per participant, until each participant completed the entire LOs presentation and their respective questions.

4.4 Metrics

The following metrics were used:

Final score. A questionnaire that has five multiple-option questions was used to evaluate student's achievements.

Median observing speed. The observing speed is the total time that a student observed a given image. The median observing speed is calculated from the data generated by a number of participants who observed an image.

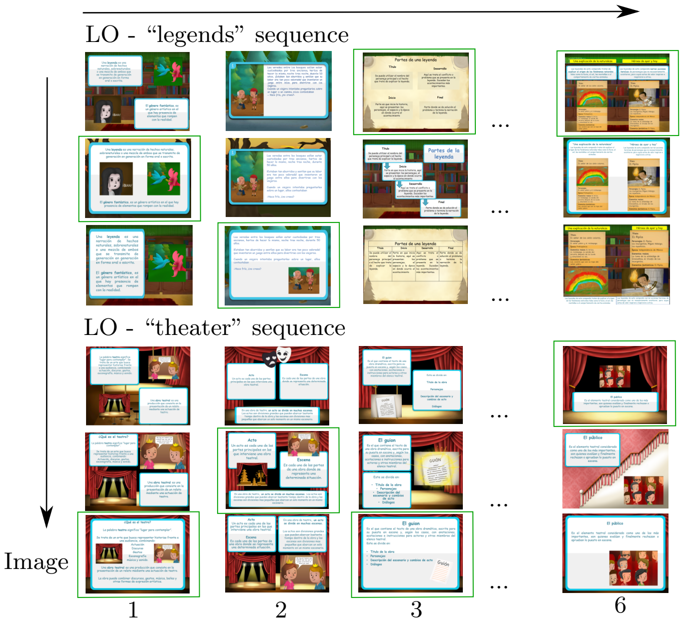


Fig. 4. Illustration of the composition of H_{eye} LOs (design phase). Nine children evaluate groups of three images to obtain the one that maximizes H_{eye}. A green rectangle shows the selected image.

4.5 Statistical analysis

Data are represented as mean ± SD, and the significance was assessed by Student’s t-test for paired data.

5 Results and Discussion

Results of the median reading speed are described in Table 1. There was a extremely significant difference in the median observing speed for H_{eye} LOs (0.629 ± 0.235 min/image) and random LOs (0.782 ± 0.329 min/image); t(11)= 4.66, p=0.0007. In other words, the observing speed is faster for those LOs composed of images that maximize H_{eye}. As it is expected, the observing time is correlated to the number of words. One special case is the image 6 of the theater sequence which obtained a lower observing speed compared to the randomly selected image; but, this image also has the lowest number of words. We speculate that the observation time for graphics in this image becomes relevant; i.e. the observation time for graphics is of the same order that the reading time of text.

There was not a significant difference in the final score for H_{eye} LOs (0.513 ± 0.207 points) and random LOs (0.496 ± 0.199 points); $t(22) = 0.7843$. It means that the final score was not affected by the aesthetics. We believe that the students’ achievement outcomes are mainly driven by the LO content itself. In

Table 1. Results for the observing speed in minutes per image. Best results are marked in bold.

	“legends” sequence					
	1	2	3	4	5	6
# words	45	62	88	78	62	141
H_{eye} LO	**0.498**	**0.723**	**0.747**	**0.698**	**0.607**	**1.163**
random LO	0.772	0.827	0.976	0.988	0.791	1.495

	“theatre” sequence					
	1	2	3	4	5	6
# words	65	75	52	43	49	25
H_{eye} LO	**0.782**	**0.719**	**0.497**	**0.386**	**0.475**	0.247
random LO	0.988	0.822	0.552	0.438	0.533	**0.212**

this direction, it is important to choose meaningful content that focuses on the learning objective [8].

Images of H_{eye} LOs have the following qualitative characteristics: (i) their elements (text or graphics) are visually balanced, (ii) element's placement on the screen establish and strengthen visual relationships between items, and (iii) their color scheme is harmonious. These characteristics agree to guidelines for authors of Learning Objects [8].

Based on the previous facts, we claim that by maximizing the proposed metric, we also maximize the fluency of LO. Fluency as a subjective experience of ease or difficulty associated with a mental process is part of phenomenon obtained from the aesthetics aspects in design [4]. Aesthetics changes as font condition or figure-ground contrast could infer on perceptual fluency.

The advantage of the proposed metric is twofold: (i) it does not require experts intervention, nor surveys applied to students, and (ii) the metric cope different graphical elements. The main drawback is that it is not possible to evaluate dynamic content; e.g. animations.

6 Conclusion and Future Work

This paper presents a metric to compare learning objects with eye tracking. The proposed metric uses the information provided by visual fixations measured from a small focus group. Images selected by maximizing the proposed metric characteristics agree to guidelines for authors of Learning Objects; and they also maximize the fluency. Results show that the aesthetic content can be evaluated by the proposed metric.

This paper opens the opportunity of automating other aspects of the evaluation of LO to obtain an holistic and fully automatic tool. We are also planning a more detailed study to compare the reading speed and the time students spend to analyze graphics. For this aim, it is necessary to detect the reading activity [13,3].

Acknowledgments. We thank all of the children who participated in the study. This research was funded by the National Council of Science and Technology of Mexico (CONACyT) through the project CATEDRAS-3163.

References

1. IEEE, Learning Technology Standards Committee: IEEE standard for learning object metadata. IEEE Standard 1484(1), 2007–04 (2002)
2. Kay, R.H., Knaack, L.: Assessing learning, quality and engagement in learning objects: the learning object evaluation scale for students (loes-s). Educational Technology Research and Development 57(2), 147–168 (2009)
3. Kunze, K., Masai, K., Inami, M., Sacakli, Ö., Liwicki, M., Dengel, A., Ishimaru, S., Kise, K.: Quantifying reading habits: counting how many words you read. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. pp. 87–96. ACM (2015)
4. Oppenheimer, D.M.: The secret life of fluency. Trends in cognitive sciences 12(6), 237–241 (2008)
5. Ozcelik, E., Karakus, T., Kursun, E., Cagiltay, K.: An eye-tracking study of how color coding affects multimedia learning. Computers & Education 53(2), 445–453 (2009)
6. Peter, C., Urban, B.: Emotion in human-computer interaction. In: Expanding the Frontiers of Visual Analytics and Visualization, pp. 239–262. Springer (2012)
7. Sanz-Rodriguez, J., Dodero, J.M.M., Sánchez-Alonso, S.: Ranking learning objects through integration of different quality indicators. IEEE transactions on learning technologies 3(4), 358–363 (2010)
8. Smith, R.S.: Guidelines for authors of learning objects. New Media Consortium (2004)
9. Tsai, M.J., Hou, H.T., Lai, M.L., Liu, W.Y., Yang, F.Y.: Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education 58(1), 375–385 (2012)
10. Vargo, J., Nesbit, J.C., Belfer, K., Archambault, A.: Learning object evaluation: computer-mediated collaboration and inter-rater reliability. International Journal of Computers and Applications 25(3), 198–205 (2003)
11. Vatrapu, R., Reimann, P., Bull, S., Johnson, M.: An eye-tracking study of notational, informational, and emotional aspects of learning analytics representations. In: Proceedings of the Third International Conference on Learning Analytics and Knowledge. pp. 125–134. ACM (2013)
12. Voßkühler, A.: Ogama description (for version 2.5). Berlin, Germany: Freie Universität Berlin, Fachbereich Physik (2009)
13. Yamaya, A., Topić, G., Martínez-Gómez, P., Aizawa, A.: Dynamic-programming–based method for fixation-to-word mapping. In: Intelligent Decision Technologies, pp. 649–659. Springer (2015)

ISSN 1870-4069

A User-Centered Approach for Cloud Service Selection and Recommendation

Aqsa Ashraf[1], Imran M. Rabbani[1], Ana Maria Martinez-Enriquez[2],
Aslam Muhammad[1]

[1] Department of CS & E, University of Engineering & Technology,
Lahore, Pakistan

[2] Department of CS, CINVESTAV-IPN, Ciudad de Mexico,
Mexico

ammartin@cinvestav.mx, aqsa.ashraf4@gmail.com,
imranmrabbani@gmail.com, maslam@uet.edu.pk

Abstract. Cloud computing is one of the most emerging industries and offers scalable computing resources and on-demand packages. A large number of cloud services are being offered by many service providers, a vendor supplies many services or same service is provided by many vendors, which creates the selection issue of appropriate cloud service from a big pool. Quality of service plays important role for better service choice. For this reason, cloud service recommendation systems are proposed in the literature that makes it easy for a user to decide which service to choose. However, user needs and requirements like economy, feedback, etc. are least considered, In our approach, an efficient cloud service recommendation system has been developed enabling users to hire a cloud service according to their own requirements. The aim is to test services according to user needs and some stated parameters. Users select the criteria from given parameter repository, the quality attributes like security, privacy, data storage, etc. and thus services are evaluated. Finally, results are shown providing user with the flexibility and opportunity to select best cloud service.

Keywords: Cloud computing, recommendation system, quality attributes, user feedback.

1 Introduction

Information technology has become very popular among organizations by providing them to communicate and create more efficiently and reliably [1]. Cloud computing is making its place in this area. It is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [2][3]. Instead of keeping data in our own hard drive or updating applications for our needs, we use a service over the Internet, at some other location, to store our information or use its applications.

pp. 99–110; rec. 2016-07-02; acc. 2016-09-19

Main focus of a cloud is to provide reliable, secure, sustainable, fault tolerant data sharing through web based applications [4]. Several well-known and huge business enterprises like Google, Amazon and IBM are offering different cloud services to users now [5][6].

The infrastructure of cloud computing allows more efficient use of hardware and software. Cloud service selection is different from web service selection in terms of quality of service and Service Level Agreement [7]. There exist challenges from service adoption perspective. The basic issue of cloud computing is to provide reliable services and allow consumers to access applications and data on demand. Different vendors follow different criteria and fulfil certain demands. Any cloud service offered by a service vendor is different in various perspectives and a description of these characteristics is as follows [8]:

On-demand service: Allows the user to unilaterally provision the computing resources required for that particular service without human interaction. Cloud providers automatically handle the user request for extra storage or computing time. The advantage of provisioning is that it allows the cloud to manage its resources on demand basis and allows it to keep the security, availability and functionality of a service at high level.

Network access: Services are available to user and are accessed through the internet or devices that are connected or networked. There is another scenario in which a software component is installed on the consumer device in order to use the service but somehow this approach is a least favorite to users due to the reason of a third party component installation as it puts an organization’s security and trust on risk.

Rapid Elasticity: Certain capabilities to use a service are provided to the user which may sometime seem as unlimited to user so these capabilities at any time can be appropriated in any quantity.

Measured Service: In case of occurrence of an issue during the provision of services, it is supervised to get resolved and the services are monitored and measured for billing according to their utilization. The service monitoring helps in recovering from a fetal error in the production environment and for system crashes as well. Logging history is also maintained with service monitoring and this history is used as an archive in order to debug a service if any problem arises in service functionality.

Resource Pooling: The computing resources of a service provider are pooled so that it can serve multiple service users by using a multi-tenant model. These virtual and physical resources are assigned and reassigned dynamically to the user according to their demand.

A thorough introduction of service models given by cloud is as follows[9][10][11]:

Software as a Service (SaaS) provides users to connect with the service providers to use the application, but they do not control the infrastructure, operating system or network infrastructure. There is no need for user to have heavy resources with installed application software. Users only need to have browsers on their devices to connect to internet in order to use the application in cloud. A daily, weekly, monthly and annually subscription period is provided and the end user is charged according to that duration working on “pay-as-you-go” mechanism [12]. An example of SaaS would be Google docs. A Cloud Service Broker helps finding suitable SaaS services [13].

Platform as a Service (PaaS) provides users to utilize all facilities on the cloud to develop and deliver their Web application and services to the end users. This service may include development, testing or the storage resources to complete the life cycle of services. Some examples of PaaS are Microsoft's Azure, Google App Engine, Engine Yard etc. A framework for user management, resource management and access management in PaaS is maintained [14].

Infrastructure as a Service (IaaS) provides computing infrastructure and resources like firewalls, image library file based storage and load balancers. Users can use the visualization resources as a fundamental infrastructure for their applications. User has access to computational resources and these resources may be a CPU, network or storage [4]. Some of the known IaaS vendors are Amazon AWS, Go Grid, etc.

DataCenter as a Service (DCaaS) offers all the features of IaaS plus some additional features such as providing the infrastructure ownership to user are included in this model. Major cloud vendors of DCaaS model are Rackspace, GoGrid, Amazon AWS and business entities usually use services by these vendors if there is need to have wide control over equipment.

A large number of cloud services are being available on internet with time. Some of the examples are network and storage services provided by different vendors. The important point is to make a decision in selecting the best suited service when there are multiple service providers. Problem arises when a user cannot decide which service to select when that particular service is being provided by multiple vendors. Internet community lacks information about various services, their performance and other quality attributes. A service provided by different vendors would have different quality attributes so the user might end up selecting a service that is not quite up to the mark and does not nearly fulfil user's requirements. So, a system is developed that is easy to use and provides users with best results to their requirements.

The related work is given in Section 2, followed by the proposed system in Section 3. A case study is discussed in Section 4 and conclusions are presented in Section 5.

2 Related Work

In the cloud recommender in [15][16], a repository is maintained in which infrastructure services are stored from different providers. Domain knowledge of multiple configurations of infrastructure services is identified and formalized. This knowledge is then implemented in a recommender service on top of a relational data model. It implements transactional execution procedures and applies well defined SQL semantics for query, insert and delete infrastructure services configuration.

The algorithm proposed in paper [13] gives ranking of SaaS providers according to their offerings matching the consumer requirements. A CS Recommender in [17] consists of five main components as crawler, cloud service identifier, indexer, search engine and recommender system. Recommender uses both collaborative and content based approaches by calculating an average. System starts with the content based recommendation results and then shifts to collaborative results when number of user increases. Approach in [18] uses the data mining technique K-means clustering for service recommendation which provides computational simplicity.

A Service Measurement Index (SMI) framework given in [19, 20] was designed for the comparison and ranking of cloud services. Cloud customers can also use these indexes to compare different cloud services. Pseudo ratings are constructed based on implicit feedback in [21]. It implements matrix factorization. Ratings are extracted from user watch list. Model is retrained for new user. Ratings are gathered into a dataset and a strategy is applied to recommend a service to user. Online pattern is used for new user ratings and the model is retrained according to the new feedback for future recommendations.

Service selection and recommendation are supported by these solutions but not all the systems take into account the feedback provided by previous users of a particular service and most importantly the requirements of a user regarding service quality attributes. Moreover, user needs are approximated and not thoroughly looked into. The proposed system makes service selection easy for cloud users by evaluating services and leaving the choice of quality parameters totally on service user.

3 Cloud Service Recommender System

The proposed solution, illustrated in Figure 1 has been developed and divided into separate components and these are done using software approach. The service recommendation task has been accomplished in a simple way by the system gathering data, implementing existing technologies, collecting users, taking comments in feedback form and maintaining quality attributes. Along with the collection of data, the conditions or constraints for criteria were also collected from service seeker or user which is a very important step that should be carried out during service search. This makes the commencement of a very first step as data collection which is the most important information as the service evaluation is carried out on this information basis.

The core element of CSRS is the repository which is a collection of various quality attributes. Due to the rapid increase in the number of web services it has become very important to build a service which is trustworthy and for this reason, QoS attributes have become an important factor [22]. A large number of service attributes are available and different sets of attributes have been developed [23]. Some of the quality attributes supported by this system are Stability, Cost, Security, Availability, Performance and services are thoroughly checked and evaluated by the system. Services are analysed and the analysis algorithm is performed on services according to the selection of quality of service attributes. All the related parameters and features of services and characteristics of cloud service providers are gathered and the attributes are selected and saved for the specific cloud services selected for evaluation. After the gathering and saving of data, pre-processing steps are taken such as filling some missing information, incorrect or duplicate information removal from the data collected.

Vendors are maintained along with their services on a page and links are given for each service of a service provider. User is then provided with the option of selecting any vendor and its service from vendor service list or can seek the help of cloud recommender system by selecting the desired parameters and let the system evaluate all the services of all the vendors available and this evaluation is done on the basis of selected parameters by the user.

In order to collect user reviews or feedback according to their service usage, users are presented with a survey to input their response for a service based on the attributes stated in the survey form. Major module is the quality of service attribute rating collection in the form of feedback for each attribute of a service and weightage associated with each attribute is also collected. All of this collected information is then passed onto the database created for storing user feedback. As services are evaluated based on the review or feedback of user on its features, properties or attributes so ranking of each attribute is collected from user review. Another major module is the dashboard, where all the available services are shown to the user along with each service description, its service type whether it is storage, computing or security service and its vendor information. User can navigate to individual and separate pages for every service and can view its details and also provide feedback for service. The first module is the graphical user interface where users are given the option whether to choose any service by themselves or direct to system repository and select the desired parameters and let the system do the calculation.

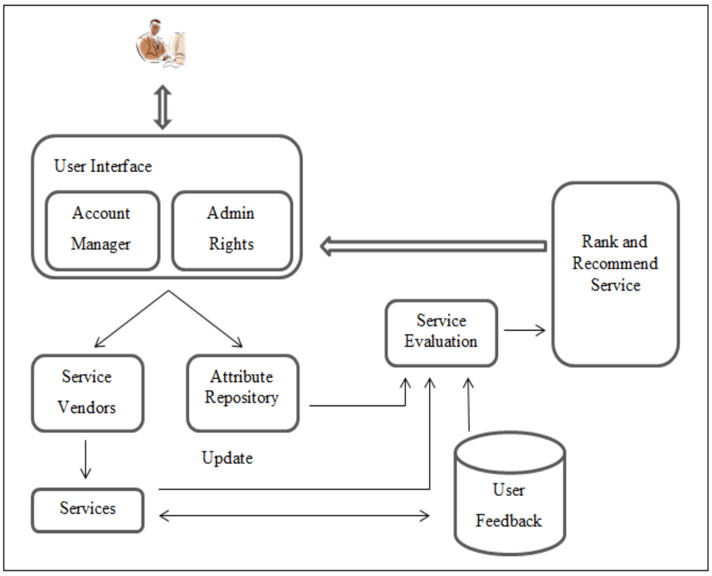


Fig. 1. Architecture of Recommender System.

The QoS parameters selected by user are passed on further to give query results for specified criteria. This modular approach of the system offers user to work in a number of phases and makes selection easy for user for best service based on user criteria. Performance of a cloud service provider depends on the performance of all its services provided. Performance of a service provider is calculated using Eq. 1.

$$P_{sp} = \sum_{m=1}^{n} (p_m(s)), \tag{1}$$

where n is the total number of services of cloud service provider supported by the system. This $p_m(s)$ is the score of a service based on the quality attributes selected from repository given in Eq. 2.

$$p_m(s) = \sum_{i=1}^{R_U} (f_i - t_i), \quad R_U \subseteq R \tag{2}$$

Where R represents the attribute repository which contains different quality of service parameters and R_U is the number of user stated or user selected quality parameters which is clearly a subset of the original parameter repository as shown above. f_i is the user feedback with respect to attribute i and t_i is the company value for the attribute or quality parameter i as a threshold value. Total feedback against a single service is calculated by using Eq. 3:

$$\begin{gathered} f_{total} = \frac{f_1 + f_2 + \cdots + f_n}{n}, \\ 0 < f_{total} < 10, 0 < t_i < 10, \end{gathered} \tag{3}$$

where n is the total number of service users.

This module functions according to the results collected from users in the form of feedback and the threshold value assigned to the services. The system is implemented using these algorithms on asp.net software in the form of a web solution. When a service is rated by a user, the ratings for all its attributes along with user information are stored in database. End user is first registered with the system and after registration assigned an account. On signing in, user is given the option of selecting services by themselves or move to the system repository.

In system repository, quality parameters are maintained and users are asked to select the desired parameters on which they want services to be evaluated and recommended. On selecting parameters by users, the evaluation processing of services starts which takes user feedback and company value as inputs and after processing according to the designed algorithm, produces some results which evaluates all services based on the parameter selection accordingly. According to the results presented, user can choose first service as it is ranked as the best among all others based on user criteria. Working of the recommender system is illustrated in Figure 2.

Finding Service Rank

```
Login profile
ServiceList[]
Select Service Type p
Select n parameter/parameters from R
Services← Apply algorithm with f and t
Get MAX
//Load Service
If(Service k=MAX)
        Set Service k rank→ 1
        Else if(Service k=MAX-1)
        Set Service k rank→ 2
        Output SortedServiceList[]
```

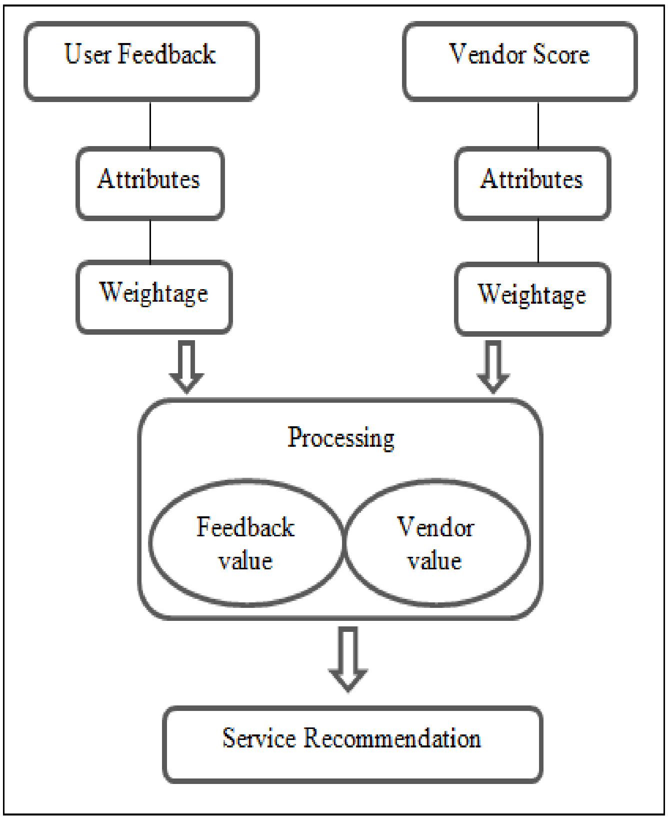


Fig. 2. Recommender System Working.

This pseudocode shows the initial step of user login and then selecting the type of service needed to be evaluated and recommended. The service type which is used as a case study is storage services. One or more than one parameters are selected from repository and algorithm is applied on all the services of given type using feedback and threshold values. Service with maximum value(MAX) obtained is ranked highest among the others and is recommended as the best suited service according to user's quality requirements.

4 Storage Service Recommendation

The services evaluated by the recommender system are given in Table 1. The table shows some of the cloud storage services provided by different vendors over the internet. The need for storage service arises for several reasons such as if the user device space is not enough to keep large amount of data or organizations have the need to keep data on cloud. The user data is stored on a space provided and managed at the server side and only users with registered accounts have access to their own data.

Table 1. Cloud Services with Type.

Service Name	Service Type
GoogleDrive	Storage
OneDrive	Storage
IDrive	Storage
DropBox	Storage
SugarSync	Storage
JustCloud	Storage
BackBlaze	Storage
CrashPlan	Storage

An introduction to the services shown in the Table I is given as follows.

- GoogleDrive provides a free storage of 15Gb initially along with presentation builder, spreadsheet and word processor. This service also has applications for managing files and data from cellular phone. Third party applications are also supported to sign documents or send faxes etc.
- OneDrive provides free 15Gb of initial storage which is further extended by completing a few offers. At present, it offers storage of 5Gb to the users whereas, the Microsoft office subscribers have a storage limit in terabytes. It implements modern design of Microsoft user interface. There is an option of creating files and folders on the web using the Microsoft office online platform. Social networks could be linked to the storage account which allows file sharing and permissions are set for the other users to edit the data.
- IDrive provides 5Gb of free storage on account registration and 1Tb for personal pro level means who are the payers for this service. After account creation, a key type is selected by user whether it is a private encrypted key or IDrive managed default key. IDrive also supports social media backup and this feature lets user to back up all the data such as audio files, video files and pictures.
- DropBox provides storage service to linux and blackberry users as well and for windows it offers an official phone application as well. Upto 2Gb storage is free whereas a specific amount needs to be paid in order to get more storage. Certain offers are available with which the providers provide user with free upgrades or some storage as a reward.
- SugarSync provides free storage of 5Gb after creating an account which is more than the storage provided by Dropbox. But free account is provided only for 30 days after that files access or account usage is charged as per usage. Free trial period is increased if this service is recommended and referred to other users as well and the amount is 500mb per person and another way to increase amount is

by completing different tasks. Data uploading speed can be managed at user end based on the internet speed of service user.

- JustCloud gives free storage of 5Gb on new account registration and on recommending this service to other friends gives additional storage at the rate of 1Gb per person. JustCloud servers implement AES 256 bit encryption but its security is somehow limited to just basic features. File sharing with the access limitation of view or read only and edit is available and it also supports file versioning.

Table 2. Weighted Score of Services.

Service	Availability	Cost	Security	Stability	Performance
IDrive	6.87	6.62	6.75	5.5	6.37
JustCloud	7	6.57	6.42	7	6.28
OneDrive	7	6.57	6.42	7	6.28
SugarSync	7	7.12	6.62	7.12	6.87
BackBlaze	7	7.12	6.62	7.12	6.87
CrashPlan	7.12	7.75	6.12	7.37	7.37
DropBox	6.28	6.57	7	5.85	5.71
GoogleDrive	6.62	6.87	6.37	6.37	5.87

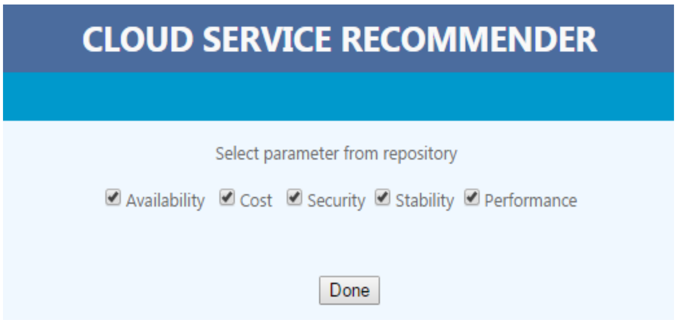
- BackBlaze offers an unlimited storage for about 50 to 58 dollars a year which is a very small amount compared to other backup services. The problem with this service is that it is very slow and offers a minimum amount of features which is not an attractive property for a service and the decision for backing up data is taken by the service instead of service user and there is no availability of sharing files stored on the cloud. From user level SSL/TLS encryption is done for transferring files to cloud and AES encryption is done at the service server level. A key is stored on the servers in a secure format.

- CrashPlan provides a 30 day free trial and after that pricing according to selected plan is started. There are no limits in file sizes and their storage but it does not support features like collaboration, file sharing or synchronization. The initial backup time is slow and it may take almost a day to backup all files on cloud but is a very powerful online backup solution. A peer-to-peer backup is supported which allows a user to back-up their data on someone else’s computer who have also an account for CrashPlan and allow some space for their friends data and all of this is done with encryption. Level of security differs for different plans. Files are transferred using 128 bit SSL encryption.

The domain of cloud storage services is very vast and a huge number of storage and backup related services are available which cover major areas of this field. Table 2 shows the weighted score of each parameter supported in the system repository for given storage services.

All Parameter Selection

A scenario is the selection of all quality parameters and it will be checked in this scenario that which storage service is recommended by the system. The selection is shown in Figure 3 which shows that all the parameters are selected from parameter repository page for services to be evaluated based on them and these are availability, cost, security, stability and performance.



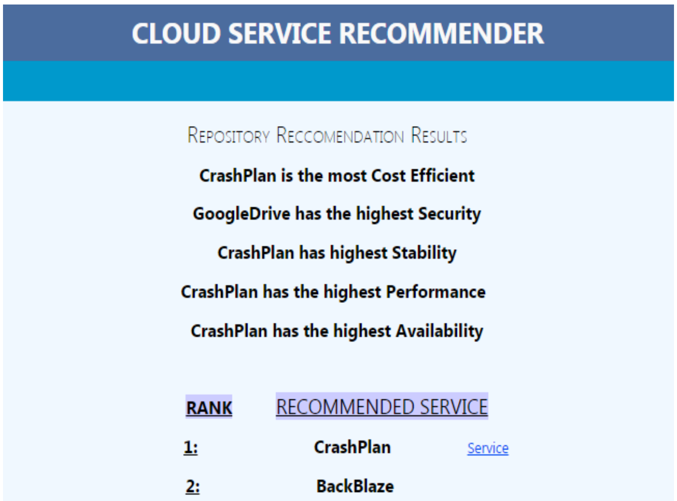
CLOUD SERVICE RECOMMENDER

Select parameter from repository

☑ Availability ☑ Cost ☑ Security ☑ Stability ☑ Performance

Done

Fig. 3 All Parameters Selection.



CLOUD SERVICE RECOMMENDER

Repository Reccomendation Results

CrashPlan is the most Cost Efficient

GoogleDrive has the highest Security

CrashPlan has highest Stability

CrashPlan has the highest Performance

CrashPlan has the highest Availability

RANK	RECOMMENDED SERVICE	
1:	CrashPlan	Service
2:	BackBlaze	

Fig. 4 Service Recommendation for All Parameters.

System now evaluates all the services according to all of the selected parameters and its result is shown in Figure 4.

Figure 4 shows that according to the selection of all the parameters, the best suited cloud storage service is CrashPlan and it is the service being recommended by the system and the second best service according to the result is Backblaze being shown at the bottom. The top of the above Figure shows results for all the parameters individually meaning that according to the system evaluations, which service falls at top for any individual quality parameter. This shows the fine and efficient working of the recommender system which provides detailed description of services.

5 Conclusion

In widely spreading and complex domain of cloud computing, it is necessary to assist cloud users in selection of services offered by this paradigm. Problem arises with multiple options which makes decision making difficult for user. Selection may be affected by multiple parameters like cost, efficiency, resource nature, satisfaction, among other. So, we develop a powerful technique to recommend cloud service to its seekers which not only considers the user requirements but also quality of service and feedback of several other clients.The service which fulfills the maximum requirements of the user is recommended consequently the surfing time is decreased and service seeker can avail the best of all available cloud services. In future, the feedbacks and opinions of users could be organized according to their dates so that it defines the concept and working of a service in a particular time.

References

1. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud Computing and Grid Computing 360-Degree Compared. In: Workshop on Grid Computing Environments (2008)
2. Technology, N.I.o.S.a.: The NIST Definition of Cloud Computing. Special Publication, pp. 800–145 (2011)
3. Buyya, R., Broberg, J., Goscinski, A.: Cloud Computing Principles and Paradigms (2011)
4. Neela, K.L., Kavitha, V.: A Survey on Security Issues and Vulnerabilities on Cloud Computing. Int. Journal of Computer Science and Technology (2013)
5. Kumar, S.,Versteeg, S., Buyyaa, R.: A Framework for Ranking of Cloud Computing Services. Elsevier, pp. 1012–1023 (2013)
6. Iosup, A., Ostermann, S., Yigitbasi, N., Prodan, R., Fahringer, T., Epema, D.: Performance analysis of cloud computing services for many-tasks scientific computing. IEEE Transactions on Parallel and Distributed Systems, pp. 931–945 (2011)
7. Preeti Gulia, S.S.: Dynamic Ranking and Selection of Cloud Providers Using Service Level Agreements. Int. J. of Adv. Research in Computer Science and Software Engineering (2013)
8. Talia, D.: Cloud computing and Software Agents: Towards Cloud Intelligent Services. Workshop on Objects and Agents (2011)
9. Dimpi Rani, R.K.R.: A Comparative Study of SaaS, PaaS and IaaS in Cloud Computing. Int. J. of Adv. Research in Computer Science and Software Engineering, pp. 458–461 (2014)
10. Serra, J.: IaaS, PaaS, and SaaS explained. James Serra's Blog (2014)
11. IaaS, PaaS and SaaS Terms Clearly Explained and Defined, http://www.silverlighthack.com/post/2011/02/27/IaaS-PaaS-and-SaaS-Terms-Explained-and-Defined.aspx
12. Tsidulko, J.: Pole Position, Ranking The Top 5 IaaS, PaaS and Private Cloud Providers. http://www.crn.com/slide-shows/cloud/300076702/pole-position-ranking-the-top-5-iaas-paas-and-private-cloud-providers.htm (2015)

13. Badidi, E.: A Framework for Software-as-a-Service Selection and Provisioning. International Journal of Computer Networks & Communications (IJCNC), pp. 189–200 (2013)
14. Banerjee, S., Gupta. N., Gupta, V.: Implementation and Management of framework for PaaS in Cloud Computing. International Journal of Innovations & Advancement in Computer Science, pp. 38–49 (2014)
15. Zhang, M., Ranjan, R., Nepal, S., Menzel, M., Haller, A.: A Declarative Recommender System for Cloud Infrastructure Services Selection. In: 9th International Conference on Economics of Grids, Clouds, Systems and Services (2012)
16. Papaioannou, I.V., Tsesmetzis, D.T., Roussaki, I.G., Anagnostou, M.E.: A QoS ontology language for Web-services. In: 20th Int. Conference on Advanced Information Networking and Applications (2006)
17. S.K.G.e.: SMICloud: A framework for comparing and ranking cloud services. In: Fourth IEEE International Conference on Utility and Cloud Computing (2011)
18. Tahir, Z., Rabbani, M., Muhammad, A., Martinez-Enriquez, A. M.: Cloud Service Recommender System Using Clustering. In: 11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), IEEE (2014)
19. Gang, J.W., Keqing, H., Weidong, Z., Panpan, G.: Cold-Start Web Service Recommendation Using Implicit Feedback. In: SEKE, pp. 371–376 (2014)
20. Koren, Y.: Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In: International Conference on Knowledge Discovery and Data Mining, pp. 426–434 (2008)
21. Al-Moayed, A., Bernhard, H.: Quality of Service Attributes in Web Services. In: 5th International Conference on Software Engineering Advances, pp. 367–372 (2010)
22. O'Brien, L., Merson, P., Bass, L.: Quality Attributes for Service-Oriented Architectures. In: SDSOA Proceedings of the International Workshop on Systems Development in SOA Environments (2007)

ISSN 1870-4069

Spoken English Learner Corpora

Olga Kolesnikova, Oscar-Arturo González-González

Instituto Politécnico Nacional, Escuela Superior de Cómputo,
Ciudad de México, Mexico

kolesolga@gmail.com, oscar.ar-56@hotmail.com

Abstract. In this paper we present a survey of some most significant spoken English learner corpora created up to date. Spoken learner corpora which include speech generated by learners are important in many areas of research and practice, in particular, for identifying typical pronunciation errors of learners of English as a second language (ESL), English as a foreign language (EFL), or English as a lingua franca (ELF). The data on common errors is helpful in designing more effective methods of pronunciation teaching as an aspect of language training. Also, error patterns can be implemented in intelligent tutor systems for English learning in order to design explanations and exercises in the error-preventive way and to generate a relevant feedback to the learner. The corpora we survey in this article include various types of English speech generated by learners with Arabic, Chinese, French, German, Greek, Japanese, Korean, Norwegian, Polish, Spanish, among others, as their first language (L1). Some English learner corpora described here are created for a single L1, other corpora are compiled for various first languages. Also, learner corpora vary depending on what type of English they exhibit: ESL, EFL, ELF or their combinations.

Keywords: Spoken English learner corpus, accented English speech, English as a second/foreign language, English as lingua franca, pronunciation errors.

1 Introduction

An English learner corpus is a collection of written and/or spoken texts produced by learners of English as a second language (ESL or English as L2), or English as a foreign language (EFL), or English as a lingua franca (ELF). Learner corpora are used by researchers, teachers of English, and learners for various purposes, one of them is error recognition and analysis. The results of error analysis can be applied in English language teaching in a conventional classroom environment as well as in computer assisted training or intelligent tutor systems.

Pronunciation is one of the major aspects of English training. For an efficient acquisition of English pronunciation, the knowledge of most common errors can aid in the development of adequate methods which will help the learner to understand how the English sounds are generated and recognized in speech. Another element of English teaching is the use of pronunciation exercises for speaking practice; such exercises can be designed in a more effective manner taking into account common error patterns in order to prevent mispronunciations on the one hand, and on the other hand, correct them with drills targeted at the specific errors known in advance.

In order to successfully implement the data on learner errors, the latter must be identified, studied, classified, and formalized. Formalization of error patterns or rules is necessary for their application in intelligent tutor systems and various forms of online learning. Learner corpora include the systematized and usually annotated material necessary for researchers and education professionals to work on errors, therefore, the importance of such corpora can hardly be overestimated.

In this paper we present a survey of some most significant spoken English learner corpora collected for various first languages (L1), i.e., the speech recorded in such corpora is accented by different mother tongues of English learners. For instance, a Spanish L1 – English L2 spoken corpus contains texts spoken by English learners whose first language is Spanish. In our survey, we give description of spoken English corpora with various first languages: Arabic, Chinese, French, German, Greek, Japanese, Korean, Norwegian, Polish, Spanish, among others. Some English learner corpora described here are created for a single L1, other corpora are compiled for several first languages. Also, learner corpora vary depending on what type of English they exhibit: ESL, EFL, ELF, or their combinations.

Our descriptions of corpora are structured as follows. First, we give the title of the corpus, then we indicate the name of the coordinator/s and/or head/s responsible for the corpus compilation, in what institution or organization the corpus was created, the aim of the corpus, the subjects, themes, topics covered in texts, some details on speech recording, then how the recorded material was processed, transcribed, and annotated. The extension and completeness of the descriptions depend on information found in public domain to the best of our knowledge.

In some cases, a collection of spoken texts is called a database instead of corpus. The difference between a corpus and a database is that the latter is aimed to include a wide range of data types: from spontaneously generated texts (for example, speech in informal interviews or chats) to texts read by learners in a formal environment of a classroom or during a test or exam. Sometimes, such databases include not only speech but also written texts: essays, summaries, description, reports, etc.

There are English learner corpora which form a part of larger corpora: for example, a multi-layered learners' corpus called AixOx (Herment, Tortel, Bigi, Hirst, Loukina 2012) includes French native speakers reading French passages, English native speakers reading English passages, as well as French speakers reading English and English speakers reading French.

The rest of the paper is organized as follows. Sections from 2 to 12, arranged in the alphabetic order of the first languages of English learners, describe English learner corpora created for a particular single F1. Section 13 presents corpora compiled for multiple first languages, each corpus is described in a separate subsection arranged in the alphabetic order of the corpora titles.

2 Arabic

2.1 Qatar Learner Corpus

The Qatar learner corpus is an Arabic L1 – English L2 spoken corpus. It was created by Yun (Helen) Zhao, at Modern Language Department of the Carnegie Mellon University, USA. The recorded participants are ESL learners with Arabic, mostly Qatar, as their first language. The corpus includes spoken interviews of 19 participants. For each participant, the following metadata is available: first name, grade, nationality, gender, the rates of the participant's reading skills and language usage, as well as the average English rate, see Table 1. The corpus is freely available online[1]. The link to the Qatar learner corpus is located at the website of Université catholique de Louvain[2]. At this page, an extensive list of learner corpora for various first and second languages around the world can be found.

Table 1. Data on some participants recorded in the Qatar learner corpus.

Name	Grade	Nationality	Gender	Reading Skills	Language Usage	Average English
Sam	12	Qatari	Male	39.61	65.76	52.685
Abe	12	Qatari	Male	62.57	73.67	68.12
Charles	11	Qatari	Male	61.12	80.31	70.715
Tom	12	Qatari	Male	44.35	39.31	41.83
Larry	12	Qatari	Male	93.66	89.91	91.785
Bill	12	Qatari	Male	75.32	99	87.16
Jenny	12	Qatari	Female	77.63	97	87.315
Nancy	12	Qatari	Female	96.81	99	97.905
Lucy	12	Qatari	Female	99	97.2	98.1
Anne	12	Qatari	Female	94.54	87.75	91.145
Alice	11	Qatari	Female	78.46	98.8	88.63
Paula	11	Qatari	Female	53.06	57.4	55.23
Pat	12	Qatari	Female	53.37	84.62	68.995
Tina	12	Qatari	Female	79.65	89.32	84.485
Linda	11	Qatari	Female	66.95	90.51	78.73
Donna	11	Kuwaiti	Female	71.32	91.1	81.21

3 Chinese

The interest of Chinese researches and educationalists for the English language acquisition is quite big as evidenced by many spoken English learner corpora created for this first language. In this section we give a brief description of most significant of them.

[1] http://talkbank.org/ data/SLABank/English/

[2] https://www. uclouvain.be/en-cecl-lcworld.html/

3.1 BICCEL

BICCEL stands for the Bilingual Corpus of Chinese English Learners. It is a spoken corpus which contains the recordings of speech produced at the National Oral English Test by forth year students majoring in English. The collected material spans the years from 2001 to 2005 and includes 1,100 test participants. This corpus also contains written texts which are in-class assignments. The work on the BICCEL was supervised by Prof. Wen Qiufang and done by Dr. Wang Jinquan at the National Research Center for Foreign Language Education, Beijing Foreign Studies University, China.

3.2 CUCASE

CUCASE is the City University Corpus of Academic Spoken English compiled by David Yong Wey Lee at the City University of Hong Kong. It is a multimedia collection of learner speech, and it also includes data produced by native English speakers.

3.3 COLSEC

COLSEC is the acronym for the College Learners' Spoken English Corpus. This corpus consists of the transcribed speech of non-English university majors produced at the National Spoken English Test. This corpus was created by Yang and Wei (2005). The COLSEC was used in the work of Luo, Yang, and Wang (2011) to define mispronunciation rules. These rules and the statistics of mispronunciations observed in the COLSEC allowed the authors to construct pronunciation lexicons in which prior probabilities was indicated. Prior probabilities reflect how likely each type of error might occur as language models for automatic speech recognition (ASR) systems and applications.

3.4 ESCCL

ESCCL is the English Speech Corpus of Chinese Learners. It contains dialogues read aloud, and it was created by Chen Hua from Nantong University, Wen Qiufang from the Beijing Foreign Studies University, and Li Aijun from the Chinese Academy of Social Sciences (Hua, Qiufang, Aijun 2008). The recorded speech was produced by participants at four different educational backgrounds. They were asked to complete two tasks: to read a dialogue aloud and to produce a spontaneous dialogue on a given topic. The recordings were collected in different parts of China and in its ten major dialectal areas. The authors claim that the quality of their recording is higher than the recordings stored in other corpora: the recording for the ESCCL was done in language laboratories by MP3- H06 at the sample rate of 16,000 (16 kHz, 16 bit mono PCM).

The recorded participants cover almost all learners under formal classroom instruction, with an interval of three years between adjacent groups. In each part of China (also in each dialectal district), at least 30 junior middle school students, 30 senior middle school students, 30 college English majors, and 30 English majors for Master degree were willingly recorded. In each group, the number of male and female students was well balanced.

The corpus also includes prosody annotations using both British and American annotation systems. The annotations were made on the computer with Praat software[3] (Boersma 2002) by 15 college English researchers. All the data were cross-checked by three phoneticians in China.

3.5 SWECCL

SWECCL stands for the Spoken and Written English Corpus of Chinese Learners. As its title shows, the corpus includes a spoken part, entitled as the Spoken English Corpus of Chinese learners (SECCL), and a written part, called the Written English Corpus of Chinese learners (WECCL). The SWECCL was compiled by Wei Qiufang, Wang Lifei, and Liang Maocheng (Wen, Wang, Liang 2005). Two versions of this corpus have been created up to date: the first version was completed in 2005, and the work on the second version was finished in 2007. The spoken section (SECCL) in its second version includes speech produced at the National Spoken English Test: in the years 2003-2006 by second-year English majors, and in the years 2000-2006 by forth-year English majors. The corpus also comprises longitudinal data of 40 hours of speech within the years 2000-2004.

3.6 TSLC

TSLC is the TELEC Secondary Learner Corpus, which includes written and spoken English Chinese (Allan 2002). TELEC stands for the Teachers of English Language Education Center. This institution maintains a computer network called TeleNex designed to provide support for language teachers in Hong Kong. The work on the TSLC corpus was done under TELEC auspices. The corpus was developed within a number of years beginning from 1994 and comprises now ten million words of running text, mostly written compositions; however, it also includes a small spoken part. The work on the corpus was led by Quentin Allan, the University of Hong Kong.

The TSLC has been used primarily for pedagogical purposes (Allan 1999). The TELEC staff members use the TSLC for developing teaching materials, designing lessons that address common problem areas, and answering questions asked by teachers on TeleNex webpage[4]. Through TeleNex, teachers who do not have time or expertise to carry out their own corpus investigations can still enjoy the benefits of learner corpora research.

4 French

4.1 ANGLISH

The ANGLISH database (Tortel 2008) was created for British English as L2. It includes L1 as well as L2 French speakers. The recording was done in an anechoic room, and 63 participants were recorded while reading and repeating texts. The reading part

[3] http://www.praat.org/

[4] http://www.telenex.hku.hk/telec/

includes 1,260 utterances. Continuous unprepared speech was also recorded. The participants included native speakers of British English (23 speakers: 13 female and 10 male), non-specialist working adult French speakers of English (20 learners: 10 female, 10 male), and second-year university French students of English (20 learners: 10 female, 10 male).

The recordings of the reading part of the corpus were manually segmented into phonemes and labeled with CVC codes using the Praat software[5] (Boersma 2002). The corpus is freely available on SLDR (Speech and Language Data Repository[6],) and its description can also be found at the website of Université catholique de Louvain[7].

4.2 AixOx

AixOx, a multi-layered learners' corpus, (Herment, Tortel, Bigi, Hirst, Loukina 2012) includes French native speakers reading French passages, English native speakers reading English passages, as well as French speakers reading English, and English speakers reading French.

4.3 Learners' Corpus of Reading Texts

This learner corpus includes unprepared reading of English texts. The texts are short abstracts of fiction or made-up dialogues. The corpus was compiled by Sophie Herment, Valérie Kerfelec, Laetitia Leonarduzzi, and Gabor Turcsan from Laboratoire parole et langage, Aix Marseille Université in Aix-en-Provence, France. The 54 recorded participants were first-year French students of the English Department at the above mentioned university. The corpus is accessible online[8] and can be freely downloaded.

4.4 CoNNECT

CoNNECT is the Corpus of Native and Non-native EFL Classroom Teacher Talk. It contains transcripts of native and non-native English lesson audio recordings performed in a secondary classroom of students ranging from A1 to B2 levels. The data was collected within the period from January 2009 till March 2011. The recordings were made in French-speaking Belgium and in Britain. The CoNNECT includes two sub-corpora: the native English sub-corpus with 108,988 words and non-native English sub-corpus with 56,526 words.

The native English recordings include 24 lessons, and the non-native English recordings include 14 lessons. The corpus has been used to analyze the linguistic features of native-speaker teachers' classroom language that could be useful to non-native foreign language teachers. Its native English part can also serve as a baseline for comparison with the non-native sub-corpus.

[5] http://www.praat.org/

[6] http://sldr.org/, http://sldr.org/sldr000731/

[7] https://www.uclouvain.be/en-cecl-lcworld.html/

[8] http://sldr.ortolang.fr/voir_depot.php?lang=en&id=15&allpreview=1/

The CoNNECT has been transcribed according to the guidelines used for the Louvain International Database of Spoken English[9] (LINDSEI).

The main objective of the CoNNECT (Meunier 2016) was to study native versus non-native teachers' speech, in particular, lexical choices and prosody. It was found in the analysis of the corpus that native teachers use prosody as a strategic pedagogical tool: they employ rising intonation to draw the learners' attention and longer pauses to prompt learners' feedback and reactions. Another result of the research is findings with respect to rephrasing strategies of teachers. Native teachers rephrase guidelines, feedback, task descriptions more often that non-native teachers do, and the former also use several types of rephrasing for the same turn, sometimes up to three variants. Non-native teachers tend to use full forms of terms in spoken interactions, for instance, *is not*, *could not*, *is going to*, while native teachers use contracted forms instead.

5 German

5.1 LeaP

The LeaP corpus (LeaP stands for Learning Prosody in a Foreign Language) by Milde and Gut (2002), see also (Gut 2004, 2012), consists of two sub-corpora: the first sub-corpus includes recordings of ESL learners, and the second one contains the speech of learners of German as a second language. The corpus is available to the scientific community[10] and can be downloaded upon the request to the authors.

The LeaP corpus was collected within the frame of the LeaP project, which was led by Ulrike Gut at the University of Bielefeld, Germany, within the time period from 2001 to 2003. The aim of the LeaP project was related to the acquisition of prosody by non-native speakers of German and English; therefore, the researchers were concerned with phonetic and phonological description of non-native prosody and exploration of learner variables that influence the language acquisition process.

The corpus data covers a wide range of speakers in terms of age, sex, native languages, level of competence, length of exposure to the target language, age at first exposure to the target language, and non-linguistic factors such as motivation to learn the language, musicality, and so forth. The age of the non-native speakers at the time of the recording ranged from 21 to 60. The data was collected from different groups of speakers: learners before and after a period abroad, before and after a four-month prosody training course, learners with different levels of competence; special attention was given to advanced learners who are hardly distinguishable from native speakers.

Four types of speech styles were recorded: nonsense word lists, readings of a short story (about 2 minutes), retellings of the story (between 2 and 10 minutes), free speech in an interview situation (between 10 and 30 minutes) The recordings were annotated manually and automatically on eight different tiers including pitch, tones, segments, syllables, words, phrasing, parts-of speech, and lemmas. The entire corpus consists of

[9] http://www.uclouvain.be/en-307849.html/

[10] http://www.philhist.uni-augsburg.de/de/lehrstuehle/anglistik/applied-inguistics/Forschung/leap/

359 annotated files, includes a total of 131 speakers, and the total amount of recording time is more than 12 hours.

Many research works have been performed on this corpus. For example, one the works (Carson-Berndsen, Gut, Kelly 2006) discovered regularities in non-native speech which can be used in a variety of pedagogical activities as well as in computer assisted training and automatic speech recognition.

5.2 GLBCC

GLBCC stands for the Giessen-Long Beach Chaplin Corpus, it includes transcribed interactions between native English speakers, ESL and EFL speakers. The corpus was compiled by Andreas Jucker and Sara Smith at the University of Giessen, Germany. The corpus can be accessed online[11] and downloaded upon a request to the authors.

In the process of corpus creation, pairs of students, in California (for English as native and second language) and in Giessen (for English as foreign language), participated in the experiment. They were asked to watch the first part of a silent Charlie Chaplin movie. One participant, called speaker A, was then asked to retell in a monologue what he/she had seen so far, while the other participant, called speaker B, watched the rest of the movie and told his/her partner the second part of the movie. In the end of the conversation, the two participants discussed several aspects of the movie on the basis of a few written prompts.

In the process of corpus compilation, 108 sessions were recorded involving 191 speakers. There were 83 A-speakers, 90 B-speakers, and altogether, the corpus comprises 35 American, 4 British, and 2 Australian native speakers. 77 non-native speakers are Germans, the others have a variety of linguistic backgrounds, including Hispanic, Japanese, and Korean.

6 Greek

6.1 YoLeCorE

YoLeCorE stands for the Young Learner Corpus of English, it is an English Greek spoken pedagogic corpus of video-recorded EFL classes. The corpus was created by Marina Mattheoudakis and Thomas Zapounidis at the Aristotle University of Thessaloniki, Greece (Mattheoudakis 2014). This audiovisual written and spoken corpus was compiled at the Third Model Experimental primary school in Evosmos, and it is an innovative pedagogic corpus which includes all language instances produced in a class of 8-9-year-old learners during one school year.

[11] http://ota.oucs.ox.ac.uk/headers/2506.xml/

7 Japanese

7.1 NICT JLE

NICT JLE corpus is the Japanese Learner English (JLE) corpus compiled at the National Institute of Information and Communications Technology (NICT) in 2004 by Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, Kyoto, Japan. The corpus data includes the transcripts of audio-recorded speech samples of English oral proficiency interview test which is called ACTFL-ALC SST (Standard Speaking Test). The corpus contains 1,281 samples, 1.2 million words, 300 hours in total. This corpus is available online[12] and can be freely downloaded.

The metadata includes the proficiency level of the participants (9 levels) based on the SST scoring method; this makes it possible to easily analyze and compare the characteristics of interlanguage of each developmental stage. This is one of the advantages of the NICT JLE corpus.

The corpus is annotated with more than 30 basic tags for all files and with error tags for 167 files. The basic tags include tags for representing the structure of the interview, tags for the interviewee’s profile, tags for speaker turns, and tags for representing utterance phenomena such as fillers, repetitions, self-corrections, overlapping, etc. The error tag set includes 47 tags of lexical and grammatical errors of learners.

In order to compare native and non-native English speech, the NICT JLE corpus includes a native English speakers’ sub-corpus.

8 Korean

8.1 ETC

ETC stands for English (as a foreign language) Teacher Corpus, it contains teacher talks in language classrooms and includes a total of 247,398 words compiled through 62 hours of recording of EFL classes. The corpus was created by Ye-Eun Kwon and Eun-Joo Lee (Kwon, Lee 2014).

The EFL corpus consists of two sub-corpora: first, the teacher talk was collected from four Korean EFL teachers, and second, the speech recordings of five native university EFL instructors were made. All the non-native English teachers were teaching general English classes at three different universities in Seoul at the time of data collection. For comparative purposes, five native teachers who were teaching at two different universities in Seoul participated in the study. The ETC two sub-corpora are called non-native EFL teacher (NNET) corpus and native English teacher (NET) corpus. The NNET corpus includes 123,122 words, and the NET corpus contains 124,275 words.

The EFL teachers’ age ranged from the late twenties to mid-forties. The Korean EFL teachers held graduate level degrees in English education or general education while the native EFL teachers had graduate level degrees in fields other than English or education. The participants’ teaching experience ranged from three to nine years, and

[12] http://alaginrc.nict.go.jp/nict_jle/index_E.html/

the Korean EFL teachers had on average slightly less experience teaching English than native EFL teachers.

8.2 NICKLE

NICKLE stands for the Neungyule Interlanguage Corpus of Korean Learners of English. It includes the written part and the spoken part. The spoken part consists of student interviews and transcriptions of oral speech tests. The corpus was created by Ji-Myung Choi at the Yonsei University, Seoul, Korea.

9 Norwegian

9.1 EVA

The EVA Corpus of Norwegian school English was compiled as a part of the government-sponsored EVA Project (Evaluation of English in Norwegian schools), with Angela Hasselgren as the project leader and Anna-Brita Stenström as its advisor, both from the English Department at the University of Bergen, Norway. The recorded speakers were Norwegian pupils of 14–15 years.

The corpus consists of the transcripts of 62 pupils taking the EVA 8th grade oral test. This test includes three picture-based tasks: the first task involves describing, narrating, and discussing, the second task involves giving instructions and checking for understanding, and the third task consists in a role play, with one role fixed, i.e. read by the pupil.

The main part of the corpus includes about 35,000 words. Together with it, a smaller control corpus was compiled with 16 native British teenager speakers carrying out the same tasks.

10 Polish

10.1 PLEC

PLEC is the PELCRA Learner English Corpus created by Piotr Pęzik, Barbara Lewandowska-Tomaszczyk, University of Łódź, Poland[13]. PELCRA stands for Polish and English Language Corpora for Research and Applications, and it is also the name of a research group at the Department of English Language at the University of Łódź.

PLEC includes the written part and the spoken part. The corpus makes it possible to analyze many aspects of the phonetic, lexical, grammatical, and phraseological competence of Polish learners of English using quantitative and qualitative methods.

The corpus contains time-aligned interviews and other spoken interactions of Polish learners of English. The transcriptions of the corpus include manual annotations of mispronounced words, this permits researchers to study the relative frequency of word mispronunciations as well as possible patterns among them. The results of research on

[13] http://pelcra.pl/plec/ research/

pronunciation errors can be used to prioritize certain lexical items in pronunciation courses and thus they will help to develop syllabuses and materials in teaching English to Polish-speaking students.

11 Spanish and Catalan

11.1 BELC

BELC stands for the Barcelona English Language Corpus (Muñoz 2006); it is a written and spoken corpus created with the objective to do research on how age affects the acquisition of English as a foreign language. BELC was compiled by the research group Grup de Recerca en Adquisició de Llengües (GRAL) in the Department of English at the University of Barcelona. The coordinator of the research group is Dr. Carmen Muñoz Lahoz, other details and information can be found at the website of the University of Barcelona[14].

The recorded participants were 2,063 students from state schools in Catalonia, Spain. As they are residents of Catalonia, they are bilingual: their native languages are Spanish and Catalan.

According to the objective of the corpus, students of various age groups were recorded. Alongside with the age factor, another parameter was taken into account, namely, the number of hours a student passed learning English as a second language. So the students were recorded after having 200, 416, 726, and 826 hours of English language instruction in schools.

The written part of the corpus includes compositions dealt with a familiar topic: *Me: my past, present and future*, it was the first task completed by the subjects.

The spoken part of the corpus includes three other tasks performed by the subjects: first, oral narrative prompted by six pictures at which the subjects were to freely look before and during their spontaneous talk; second, oral semi-guided interview which began with a warm-up in the form of questions about the subject's family, daily life, and hobbies and then continued as a spontaneous talk on any other topics initiated by the interviewer as well as by the subject; third, role-play performed in randomly chosen pairs: one of the students played the role of the mother/father, and the second student, the role of the son/daughter. The latter had to ask permission to have a party at home and both role play partners had to negotiate the setting, time, music, eating, drinking, and any other activities and details.

The BELC data consists of the recordings of those students who could be followed longitudinally and for whom there are two, three, or four collection times over a period of seven years. However, not all subjects performed all the four tasks.

BELC was updated in 2014: its spoken part was expanded by adding more recordings of oral narratives which were also transcribed, and to its written part, more compositions were aggregated. Other details on BELC can be found online[15].

[14] http://www.ub.edu/web/ub/en/recerca_innovacio/recerca_a_la_UB/grups/fitxa/G/ADQULENG/equipInvestigador/index.html/

[15] https://www.uclouvain.be/en-cecl-lcworld.html/

11.2 SULEC

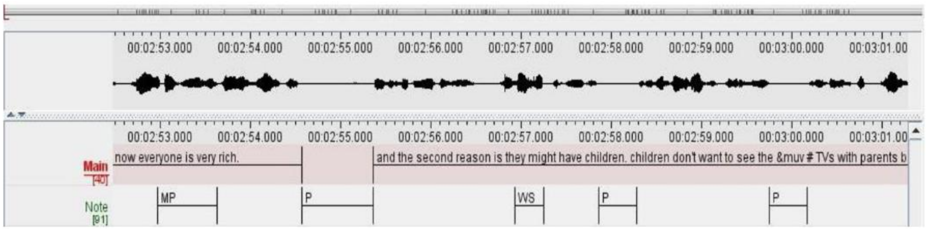
SULEC is the Santiago University Learner of English Corpus which includes written and spoken data. The written part contains compositions and argumentative essays, while the spoken section includes semi-structured interviews, short oral presentations, and brief story descriptions. The corpus was created starting in 2002 by a research group at the University of Santiago de Compostela led by Ignacio M. Palacios Martínez, see (Palacios Martínez 2005).

12 Taiwanese

12.1 LTTC English Learner Corpus

The LTTC English Learner Corpus consists of language samples produced by Taiwanese learners of English. LTTC stands for the Language Teaching and Testing Center. This center in cooperation with another institution, Graduate Institute of Linguistics (GIL), started the work on the corpus in 2007. The project director is Prof. Hintat Cheung from GIL and the co-directors are Dr. Zhao-Ming Gao from the Department of Foreign Languages at the National Taiwan University and Dr. Siaw-Fong Chung from the Department of English at National Chengchi University.

The participants were English learners who took the General English Proficiency Test (GEPT), a language proficiency examination developed and applied by the LTTC. The corpus includes 2,000 written samples and also includes a spoken section with 400 speech samples; both written and oral samples were collected at the Intermediate GEPT examination. For each participant the following metadata is given: the region of Taiwan (North, East, etc.) where the test was taken, the age, gender, education level of the test-taker, his/her major if the test-taker was a college graduate, whether the test-taker was a student or not, and whether he or she had lived in an English-speaking country, and if so, for how long.



now everyone is very rich.

and the second reason is they might have children. children don't want to see the &muv # TVs with parents b

Figure 1. Soundwave and transcription of spoken data segment from the LTTC English Learner corpus.

The spoken data was first recorded on cassette tapes, then digitized and transcribed using the software ELAN (EUDICO Linguistic Annotator) (Hellwig, van Uytvanck, Hulsbosch 2008) as well as tagged using the CHAT (CHILDES) format (MacWhinney 2008). Tags were added inside the body of the transcriptions for repetitions, self-corrections, incomprehensible sounds, lengthened vowels, and other characteristics. Tags for filled and unfilled pauses, mispronunciations, and word stress errors were added in the tier below that of the main transcription. Figure 1 presents an example of a soundwave and the corresponding transcription of the data after its processing with the ELAN software.

13 Various First Languages

13.1 ITAcorp

ITAcorp is the International Teaching Assistants corpus. International teaching assistants (ITA) are international graduate students employed usually by North American universities especially in the areas of engineering, mathematics, and sciences. ITAs participate in such activities as grading tests for large lectures, teaching break-out discussion sessions, and conducting office hours. The latter activity involve ITAs as tutors of undergraduates on homework problems, preparing them for tests, and answering questions on behalf of a supervising professor.

In practice, it turned out that many ITAs are not prepared for their roles in these sessions. Therefore, they are suggested to take advanced ESL for academic purposes and ITA preparation courses to improve their posterior performance as teaching assistants. The goal of these preparation courses is two-fold: to give instruction in English language usage in the ITAs' activities context and to teach pedagogy.

The ITAcorp consists of transcribed recordings of such preparation courses at a large Northeastern American university, which included different classroom activities and computer-mediated activities (chats): classroom discussions, lecture preparation, question answering, concept presentations, and office hours role plays. The corpus was started in 2005, a detailed description of its creation is given in (Reinhardt 2007). The work on corpus creation was done by Steven L. Thorne, Paula Golombek, and Jonathon Reinhardt at the Pennsylvania State University, USA (Thorne, Reinhardt, Golombek 2008; Reinhardt 2010).

The purpose of the corpus was to inform instruction in advanced spoken English for academic purposes and to do research on intercultural pragmatics and sociolinguistic issues.

Chinese, Thai, Korean, and other L1 English learners were recorded. The sub-corpus of office hours role plays includes approximately 103,000 tokens. In these role plays, the students played the ITA and student roles, and also, the ITA and an evaluator roles in a post-semester evaluation. Each role play is about 4 minutes long, the context of each role play is a student approaching an ITA with a typical problem that would need to be negotiated.

13.2 LONGDALE

LONGDALE is the Longitudinal Database of Learner English at the University of Louvain, Belgium. The project director is Fanny Meunier, and the team also includes Sylviane Granger, Damien Littré, and Magali Paquot.

The objective of this project is to construct a Learner English longitudinal database from learners of various L1s. These learners are followed over a period of at least three years of their studies. Up to date the data was collected within 2008-2009-2010 and 2010-2011-2012 periods including only written data from German, French, Italian, Dutch, Turkish, and Brazilian learners[16]. The project team also plans to record interviews but this work is still in progress.

The LONGDALE also includes recordings of English learners of French (EN_FR), but only of young children. In this respect, this corpus is similar to FLLOC (French Learner Language Oral Corpus, see Myles and Mitchell 2007), in which the children are aged 7 to 11, or to CYLIL (The Corpus of Young Learner Interlanguage, see Housen 2002), which contains English L2 recordings of school pupils of different European nationalities, French being one of them.

13.3 LINDSEI

LINDSEI is the Louvain International Database of Spoken English Interlanguage (Gilquin, De Cook, Granger 2010). It is a collaborative project between several universities internationally, coordinated at the University of Louvain, Belgium. Started in 1995, this database now includes 21 sub-corpora, of which 14 are complete (Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, 2 Spanish sub-corpora, Swedish, Taiwanese, Turkish), and seven are in progress (Arabic of Saudi Arabia, Basque, Brazilian Portuguese, Czech, Finnish, Lithuanian, Norwegian)[17].

The LINDSEI corpus is offered online[18] on CD-ROM containing over 1 million words, of which almost 800,000 were produced by learners, representing 11 different mother tongue backgrounds: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish. The corpora include 525 interviews and each interview consists of three tasks: first, a warm-up, in which learners were given a few minutes to talk about one of three set topics, second, a free informal discussion as the main part of the interview, and third, a picture description. The interviews are transcribed according to the transcription guidelines which can be found at the website of the University of Louvain[19].

13.4 MAELC

MAELC is the Multimedia Adult ESL Learner Corpus started in 2001 by Stephen Reder, Kathryn Harris, and Kristen Setzler of the Department of Applied Linguistics in Portland State University, USA, together with Portland Community College which provides adult ESL courses (Reder, Harris, Setzler 2003).

The corpus is a database of videos of classroom activities from four years of adult ESL classes from beginning to upper-intermediate proficiency. More than 3,600 hours of classroom interaction were recorded by six cameras and multiple wireless microphones.

[16] According to http://www.uclouvain.be/en-314347.html/ as of July 7, 2016.

[17] This data on sub-corpora was retrieved on July 7, 2016 from http://www.uclouvain.be/en-307845.html/.

[18] http://www.i6doc.com/en/collections/ cdlindsei/

[19] http://www.uclouvain.be/en-307849.html/

By now, about 150 hours of student language have been transcribed, this includes language of 250 low-level students with known background characteristics. The corpus also includes scanned copies of classroom written materials, student work, teacher logs, and teacher reflections. Many students were recorded several times per term and often in consecutive terms in different levels, which allows for longitudinal study of ESL acquisition.

The corpus was created with the purpose to do research on diverse ESL acquisition issues including longitudinal studies as mentioned above, in-depth case studies of adult learners of English, close examinations of dyadic and small-group interactions, which can focus on interactions between students from different L1 backgrounds, developmental studies of individual students who were recorded throughout several terms of study, among other themes.

13.5 T2K-SWAL Corpus

The T2K-SWAL Corpus is the TOEFL 2000 Spoken and Written Academic Language Corpus (Biber, Conrad, Reppen, Byrd, Helt, 2002; Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay, Urzua 2004). It was created with the purpose to do an empiric study of texts used on listening and reading exams and determine if they accurately represent the linguistic characteristics of spoken and written academic registers, i.e., to diagnose the representativeness of English as a Second Language/English as a Foreign Language (ESL/EFL) materials and assessment instruments and see if it corresponds to real-life language usage.

Table 2. Data on T2K-SWAL corpus.

Register	Number of texts	Number of words
Spoken:		
Class sessions	176	1,248,800
Classroom management	40	39,300
Labs/in-class groups	17	88,200
Office hours	11	50,400
Study groups	25	141,100
Service encounters	22	97,700
Total speech:	**291**	**1,665,500**
Written:		
Textbooks	87	760,600
Course packs	27	107,200
Course management	21	52,400
Institutional writing	37	151,500
Total writing:	**172**	**1,071,700**
Total corpus:	**423**	**2,737,200**

The corpus includes 2.7 million words and is representative of the range of spoken and written registers that students encounter in U.S. universities. The data in the corpus includes written and spoken texts associated with academic life, including classroom teaching, office hours, study groups, on-campus service encounters, textbooks, course packs, and institutional written materials (e.g., university catalogs, brochures), see Table 2.

Spoken texts were transcribed using a consistent transcription convention (see Edwards, Lampert 1993), then speakers were distinguished to a possible extent and some demographic information was added for each speaker (their status as instructor or student, etc.). The texts were annotated using various grammar tags, for details of annotation see (Biber et al. 2004).

13.6 ICNALE

ICNALE stands for the International Corpus Network of Asian Learners of English. It is a corpus of controlled speech and essays produced by learners of English in ten Asian countries and areas. For performing comparative studies, the corpus also includes speech of English native speakers. The project director is Shin'ichiro Ishikawa, Kobe University, Japan (Ishikawa 2011, 2012, 2013, 2014). ICNALE, as many other learner corpora, includes both written and spoken sections. The corpus can be accessed freely at its web page[20].

The spoken part of ICNALE contain the recordings of participants performing the following tasks: first, they respond to the personal attribute questions; second, they respond to the learning history questions; third, they answer the vocabulary size test, and forth, they take a telephone interview and respond to the questions concerning student's name, country, college, self-introduction (60 sec. speech), then each participant was asked to speak on some topics defined in advance, and the speech was recorded according in the following modes: Topic 1, Trial 1 (60 sec. speech after 20 sec. preparation), Topic 1, Trial 2 (60 sec. speech after 10 sec. preparation), Topic 2, Trial 1 (60 sec. speech after 20 sec. preparation), Topic 2, Trial 2 (60 sec. speech after 10 sec. preparation), then self-evaluation (0 to 5 points).

The questions and tests for the first three tasks were taken from the data collection sheet which can be obtained upon the request to the author of the corpus, as well as the topics used for the forth task.

The recorded participants were learners of English from the following countries: Hong Kong, Pakistan, Philippines, Singapore, Indonesia, and Thailand. For each participant the following data is included in the corpus: country or area of origin, sex, age, school grade (1, 2, 3, 4 ...), major or occupation; in case of students, their major at colleges; in case of employed people, their job.

The following academic genres were employed only for students: humanities, social sciences, science and technology, and life science. Information on the participants' proficiency test also included test name (TOEIC, TOEFL, etc.), the score in the above test; information on participants' motivation (using the scale from 1 to 6 points): integrative or instrumental motivation, strength of motivation, and the integrative motivation orientation score.

Also, information on the participants' English learning experiences (using the scale from 1 to 6 points): how much a participant studied English in their primary school days, in their secondary school days, in their college days, how much a participant

[20] http://language.sakura.ne.jp/icnale/index.html/

studied English in class, outside class, namely, at home, in the community, etc., how much a participant studied listening, reading, speaking, writing, how much a participant has been taught by English native participant, how much a participant has been taught pronunciation, presentation, essay writing.

13.7 CYLIL

The Corpus of Young Learner Interlanguage (CYLIL) contains English L2 data elicited from European School pupils and recorded at different levels of the participants' development including a longitudinal dataset of 6 learners followed during the period of three years as well as speech produced by other 40 learners. The corpus also includes the speech of eight native English children produced on the same tasks performed by other English learners recorded for the corpus; this serves as a baseline for comparative studies. The participants' L1 background was one of the following: Dutch, French, Greek, or Italian. In total, the corpus currently amounts to 500,000 words.

The creation of the corpus started in 1990 at Vrije Universiteit Brussel, Belgium, by Alex Housen (Housen 2002). The oral interviews consisted of informal free conversation and semi-guided speech tasks. Informants were asked to talk about events in their past, to describe pictures, to share opinions about movies they had seen, and to retell three picture stories with a variety of characters and actions.

The recorded speech was transcribed, segmented, coded, and annotated in CHAT format. The latter permits researchers to perform computer-aided analysis using the CLAN software. CHAT stands for Codes for the Human Analysis of Transcripts and CLAN is the acronym of Computerized Language Analysis. These too toolkits were designed for studying language learner speech by the CHILDES organization.

The purpose of the CHILDES system (Child Language Data Exchange System) is the study of child language and first language acquisition. However, it has been used also to do research on second language acquisition (SLA), speech pathologies, and discourse.

The CHILDES system[21] also contains electronically available corpora on child language, interlanguage, bilingual speech, and speech disorders (MacWhinney 2000). Access to the CHILDES database, the CHAT conventions, and the CLAN software is free. In its turn, CHILDES is a component of the TalkBank database[22] where many freely accessible corpora can be found.

13.8 ISLE

The ISLE speech corpus (Menzel, Atwell, Bonaventura, Herron, Howarth, Morton, Souter 2000) was created with the objective to implement a speech recognition method based on Hidden Markov Model in a computer assisted environment for teaching English at the intermediate level. The acronym ISLE stands for Interactive Spoken Language Education.

[21] http://childes.psy.cmu.edu

[22] http://talkbank.org/

Table 3. Data on the ISLE corpus.

Corpus section	Number of sentences	Linguistic issue	Exercise type	Examples
A B C	27 33 22	Wide vocabulary coverage (410)	Adaptation/ reading	"In 1952 a Swiss expedition was sent and two of the men reached a point only three hundred metres from the top before they had to turn back."
D	81	Problem phones, weak forms	Minimal pair item selection/ combination	"I said bad not bed." "She's wearing a brown wooly hat and the red scarf."
E	63	Stress, weak forms, problem phones, consonant clusters	Reading	"The convict expressed anger at the sentence." "The jury took two days to convict him."
F	10	Weak forms, problem phones	Description/Item selection/combination	"I would like chicken with fried potatoes, broccoli, peas and a glass of water."
G	11	Weak forms, problem phones	Description/Item selection/combination	"This year I'd like to visit Rome for a few days."

The corpus includes recorded speech of different types: reading simple sentences, pronouncing minimal pairs, giving answers to multiple choice questions by selecting an item from a list of options or/and combining items from different selections. English learners recorded for the corpus had German (23 learners) or Italian (23 learners) as their first language. English was learnt in its British variant.

The produced speech was recorded directly into WAV format, using a sampling rate of 16 kHz at a resolution of 16 bits. Some examples of the data included in the ISLE corpus are presented in Table 3.

Since the aim of this corpus is more specific in comparison with other more general purpose learner spoken corpora, the ISLE corpus was compiled to be applied as a tool in order to train the parameters and rules used in the recognition and diagnosis systems, to test the performance of the system on a known dataset, and to evaluate the contribution of speaker adaptation for improving the reliability of the native British English recognizer.

Taking the above mentioned tasks into account, the ISLE corpus had to be annotated at multiple levels: the word level, the phone level, and the stress level. This was necessary for determining the pronunciation errors (for instance, phone realization problems and misplaced word stress assignments, see examples in Table 4 and Table 5). Also, the corpus had to include various types of speech because the actual system for teaching EFL/ESL usually includes exercises of various grades of complexity (elementary, intermediate, advanced exercises).

Table 4. Examples of phone level errors from the ISLE corpus.

German			Italian		
from	**to**	**Example**	**from**	**to**	**Example**
oh	ow	pr**o**duce	eh	ey	s**ai**d
ax	ao	c**u**pboard	eh	ae	b**e**d
uw	ao	pn**eu**matic	ae	ey	pl**a**nning
aw	ow	**ou**tside	ih	iy	t**i**cket
aa	ae	st**a**ff	ay	iy	b**io**log**i**c**a**l
ih	iy	d**e**ssert	oh	ow	
-	p	**p**neumatic	ih	iy	
s	z	**s**aid	ax	ae	
v	w	**v**isa	-	ax	sheep
w	v	**w**eekend	-	hh	**h**onest
dh	d	**th**e	th	t	**th**in
-	w	bisc**u**it	s	z	**s**leep
-	b	thum**b**	jh	g	**g**in**g**er
g	-	fin**g**er	t	-	bai**t**
t	-	desser**t**			

Table 5. Examples of stress level errors from the ISLE corpus.

German	Italian
´report	´photographic
´television	´convict/con´vict
´contrast/contr´ast	´components

The ISLE corpus includes almost 18 hours of annotated speech and is based on 250 utterances selected from typical second language learning exercises. The ultimate purpose of the corpus usage is production of a relevant detailed feedback to English learners based on detection of their errors as well as selection targeted pronunciation exercises for error correction and further practice. The researchers who created and used the ISLE corpus on the tasks described above claim that the results of their work can be applied to any L1 (Menzel et al. 2000).

13.9 TCEEE

The Tübingen Corpus of Eastern European English (TCEEE) includes spontaneous spoken production data obtained by means of a semi-structured interview (Salakhyan 2012). The TCEEE was compiled by Elena Salakhyan at the Eberhard Karls University of Tübingen, Germany. The native languages of the participants were Russian, Ukrainian, Polish, and Slovak. The corpus includes a total of 60,000 words.

The corpus was created with the objective to study the Eastern European variety of English as one of the World Englishes (Berns 2005; House 2002; Jenkins 2007). However, since English speech was produced by non-native speakers of English, it can be viewed as an English learner corpus. The proficiency of the participants was in the range from the B1 to the C1 level according to Common European Framework of

Reference for Languages classification. The questions used in interviews elicited information about each speaker's English learning history, his/her profession and career, experience of participation in international projects as well as speech produced in a spontaneous conversation.

The TCEEE was applied in studies of tense and aspect usage and it was found (Salakhyan 2012) that Slavic speakers of English do not use the tense forms to a full extent which results in the phenomenon of simplifying and reducing the English system of tenses and aspects. These observations yield a conclusion of a possibility that Eastern European Englishes are emerging. However, in order that such varieties of English to be recognized, further studies concerning morphology, syntax, vocabulary, and semantics are required.

Acknowledgements. The author appreciates the support of Mexican Government which made it possible to complete this work: SNI-CONACYT, BEIFI-IPN, SIP-IPN grants 20162064 and 20161958, and the EDI Program.

References

1. Allan, Q. G.: Enhancing the language awareness of Hong Kong teachers through corpus data: The Telenex experience. Journal of Technology and Teacher Education, 7, pp. 57–74 (1999)
2. Allan, Q. G.: The TELEC secondary learner corpus: a resource for teacher development. Computer learner corpora, second language acquisition and foreign language teaching, pp. 195–212 (2002)
3. Berns, M.: Expanding on the Expanding Circle: Where do WE go from here? World Englishes, 24(1), pp. 85–93 (2005)
4. Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M.: Speaking and Writing in the University: A Multidimensional Comparison. TESOL Quarterly, 36(1), pp. 9–48 (2002)
5. Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., Urzua, A.: Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus. Test of English as a Foreign Language (2004)
6. Boersma, P.: Praat, a system for doing phonetics by computer. Glot international, 5(9/10), pp. 341–345 (2002)
7. Carson-Berndsen, J., Gut, U., Kelly, R.: Discovering regularities in non-native speech. Language and Computers, 56(1), pp. 77–89 (2006)
8. Edwards, J. A., Lampert, M. D.: Talking data: Transcription and coding in discourse research. Hillsdale: Erlbaum (1993)
9. Gilquin, G., De Cook, S., Granger, S.: LINDSEI Louvain International Database of Spoken English Interchange. Handbook and CD-ROM. Louvain-laNeuve: Presses universitaires de Louvain (2010)
10. Gut, U.: The LeaP corpus. http://www.phonetik.unifrieburg.de/leap/ LeapCorpus.pdf (2004)
11. Gut, U.: The LeaP corpus A multilingual corpus of spoken. Multilingual corpora and multilingual corpus analysis, 14, pp. 3–23 (2012)
12. Hellwig, B., Van Uytvanck, D., Hulsbosch, M.: EUDICO Linguistic Annotator (ELAN) version 3.6 manual, http://www.lat-mpi.eu/tools/elan (2008)

13. Herment, S., Tortel, A., Bigi, B., Hirst, D., Loukina, A.: AixOx, a multi-layered learners' corpus: automatic annotation. In Proceedings of the 4th International Conference on Corpus Linguistics, Jaèn, Spain (2012)
14. House, J.: Developing pragmatic competence in English as a lingua franca. In: Knapp, K., & Meierkord, C, Lingua Franca Communication, Frankfurt: Peter Lang, pp. 245–269 (2002)
15. Housen, A.: A corpus-based study of the L2-acquisition of the English verb system. Computer learner corpora, second language acquisition and foreign language teaching, 6, pp. 2002–2077 (2002)
16. Hua, C., Qiufang, W., Aijun, L.: A Learner Corpus-ESCCL. In Proceedings of the Speech Prosody Conference, pp. 155–158 (2008)
17. Ishikawa, S.: A new horizon in learner corpus studies: The aim of the ICNALE project. In Weir, G., Ishikawa, S., & K. Poonpon, Corpora and language technologies in teaching, learning and research, Glasgow, UK: University of Strathclyde Publishing, pp. 3–11 (2011)
18. Ishikawa, S.: Basic Corpus Linguistics (in Japanese), the original title is Beshikku Kopasu Gengogaku. Tokyo: Hitsuji Shobo (2012)
19. Ishikawa, S.: The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In: Ishikawa, S. (Ed.), Learner corpus studies in Asia and the world, Kobe, Japan: Kobe University, 1, pp. 91–118 (2013)
20. Ishikawa, S.: Design of the ICNALE-Spoken: A new data base for multi-modal contrastive interlanguage analysis. In Ishikawa, S. (Ed.), Learner corpus studies in Asia and the world, Kobe, Japan: Kobe University, 2, pp. 63–76 (2014)
21. Jenkins, J.: English as a Lingua Franca. Attitude and Identity, Oxford, Oxford University Press (2007)
22. Kwon, Y. E., Lee, E. J.: Lexical bundles in the Korean EFL teacher talk corpus: A comparison between non-native and native English teachers. The Journal of Asia TEFL, 11(3), pp. 73–103 (2014)
23. Luo, D., Yang, X., Wang, L.: Improvement of Segmental Mispronunciation Detection with Prior Knowledge Extracted from Large L2 Speech Corpus. In Interspeech, pp. 1593–1596 (2011)
24. Mattheoudakis, M.: Learner Corpora of English: Glimpses into learners' L2 development. In Proceedings of 4th Postgraduate Student Conference: Assessing and Analyzing Discourses, Faculty of English Language and Literature National and Kapodistrian University of Athens (2014)
25. MacWhinney, B.: The CHILDES project: The database. Psychology Press, Vol. 2 (2000)
26. MacWhinney, B.: The CHILDES project. Tools for analyzing talk – Electronic version, Part 1: The CHAT transcription format, http://childes.psy.cmu.edu/manuals/chat.pdf (2008)
27. Menzel, W., Atwell, E., Bonaventura, P., Herron, D., Howarth, P., Morton, R., Souter, C.: The ISLE corpus of non-native spoken English. In: Proceedings of LREC 2000: Language Resources and Evaluation Conference, European Language Resources Association, 2, pp. 957–964 (2000)
28. Meunier, F.: Learner corpora and pedagogical applications. The Routledge Handbook of Language Learning and Technology, pp. 376 (2016)
29. Milde, J. T., Gut, U.: A prosodic corpus of non-native speech. In Bel, B., & Marlien, I. (Eds.), Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence: Laboratoire Parole et Langage, pp. 503–506 (2002)
30. Myles, F., Mitchell, R.: French learner language oral corpora (FLLOC) (2007)
31. Palacios, I.: The Santiago University Corpus of Learner English. Santiago, University of Santiago de Compostela, http://www.sulec.es (2005)

32. Muñoz, C.: Age and the Rate of Foreign Language Learning. Clevedon: Multilingual Matters (2006)

33. Reder, S., Harris, K., Setzler, K.: The multimedia adult ESL learner corpus. TESOL Quarterly, 37(3), pp. 546–557 (2003)

34. Reinhardt, J. S.: Directives usage by ITAs: An applied learner corpus analysis. Doctoral dissertation, Pennsylvania State University (2007)

35. Reinhardt, J.: Directives in office hour consultations: A corpus-informed investigation of learner and expert usage. English for Specific Purposes, 29(2), pp. 94–107 (2010)

36. Salakhyan, E.: The Tübingen Corpus of Eastern European English (TCEEE): From a small-scale corpus study to a newly emerging non-native English variety. A Journal of English Linguistics, Jan Kochanowski University Press, 1, pp. 143–157 (2012)

37. Thorne, S., Reinhardt, J., Golombek, P.: Mediation as objectification in the development of professional discourse: A corpus informed curricular innovation. In Lantolf, J., & M. Poehner (Eds.) sociocultural theory and the teaching of second languages, London: Equinox, pp. 256–284 (2008)

38. Tortel, A.: ANGLISH. Une base de données comparatives de l'anglais lu, répété et parlé en L1 & L2. TIPA, Travaux interdisciplinaires sur la parole et le langage, 27, pp. 111–122 (2008)

39. Wen, Q. F., Wang, L. F., Liang, M. C.: Spoken and written English corpus of Chinese learners. Foreign Language Teaching and Research Press (2005)

40. Yang, H., Wei, N.: Construction and data analysis of a Chinese learner spoken English corpus. Shanhai Foreign Languse Eduacation Press (2005)

ISSN 1870-4069

Efecto del pre-procesamiento en la detección automática de plagio para PAN 2014 y PAN 2015

Jovani Armeaga García, Yulia Ledeneva, René Arnulfo García-Hernández

Universidad Autónoma del Estado de México,
México

UAP Tianguistenco Instituto Literario, Toluca, Estado de México,
México

jovani_2807@hotmail.com, yledeneva@yahoo.com, renearnulfo@hotmail.com

Resumen. Dentro de la detección automática de plagio, el alineamiento de texto en [1] lo define como el descubrimiento de fragmentos similares de texto entre dos documentos. La cual puede utilizarse en: detección de plagio, identificación de autoría, detección de reúso de texto, recuperación de información, entre muchas otras. El pre-procesamiento consta de diversas técnicas que se aplica en la mayoría de las tareas del Procesamiento del Lenguaje Natural (PLN), en este caso, las heurísticas presentadas son tomadas de los trabajos [1] y [2] de las mejores participaciones en la competencia internacional de detección automática de plagio PAN 2014 y PAN 2015 en la sub-tarea alineamiento de texto monolingüe, con la finalidad de conocer el efecto que tiene la eliminación de *stopwords* y el uso o no de *stemming* en las heurísticas antes mencionadas, que son técnicas dentro del pre-procesamiento.

Palabras clave: Procesamiento de lenguaje natural, alineamiento de texto, detección automática de plagio, competiciones PAN 2014 y PAN 2015.

1. Introducción

El PLN es una sub-disciplina de la inteligencia artificial y rama de la ingeniería lingüística computacional, la cual busca construir sistemas y mecanismos que permitan la comunicación entre personas y máquinas por medio de lenguajes naturales. El lenguaje natural en [3] se entiende como el lenguaje hablado y escrito que tiene como propósito que exista una comunicación entre una o varias personas. Algunas de las aplicaciones del PLN son:

- Recuperación de información,
- Traducción automática,
- Extracción de información.

La recuperación de información según [4] es, teniendo una necesidad de información y un conjunto de documentos, se ordenan los documentos por relevancia para esa necesidad y se presenta un sub-conjunto de los más relevantes. Según [5] dice que

pp. 133–152; rec. 2016-07-05; acc. 2016-09-21

"cualquier sistema de recuperación de información puede ser descrito como un conjunto de ítems de información (DOCS), un conjunto de peticiones (REQS) y algún mecanismo (SIMILAR) que determine qué ítem satisface las necesidades de información expresadas por el usuario en la petición".

De la recuperación de información se desprenden diversas tareas como:

- Generación automática de resúmenes,
- Detección automática de plagio,
- Extracción automática de palabras clave.

En la detección de plagio, en [6] el plagio se define como, copiar en lo sustancial obras ajenas, dándolas como propias, es una de las definiciones más aceptadas, lo que se refleja en los trabajos [7, 8]. Una obra es toda creación original que puede reproducirse por cualquier medio o forma [7]. El plagio puede aparecer en diversas obras como se menciona en [9]:

- Literarias,
- Fotográficas,
- Programas de cómputo,
- Musicales,
- Arquitectónicas,
- Danzas,
- Esculturas,
- Programas de radio y TV,
- Cinematográficas y audiovisuales,
- Obras de arte.

En los últimos años el uso incremental de los medios digitales, ha provocado el incremento de plagio textual o en documentos, según [8] esto es tomar el texto de un autor y hacerlo pasar como propio. Esto abarca desde la copia sin modificar nada, hasta el parafraseado del documento que modifica las palabras, pero manteniendo la idea central del texto. Esto se debe a la enorme cantidad de información que se encuentra disponible en dichos medios.

Actualmente las instituciones académicas, es en donde más se presenta el plagio en tareas de los alumnos [7], siendo un acto muy poco castigado, surgiendo así los sistemas de detección automática de plagio. Los primeros sistemas en [10, 11] se menciona que fueron WcopyFind desarrollado por la universidad de Virginia, Ferret Plagiarism Detector por la universidad de Hertfordshire y SCAM (*Stanford Copy Analysis Mechanism*) por la universidad de Stanford, estos primeros sistemas no mostraron mucha eficiencia, para la detección de documentos plagiados, al final el humano decide que es plagio y que no.

Un ejemplo en donde se puede aplicar la detección automática de plagio textual para evitar duplicados, es en los sistemas de creación de documentos. En los trabajos [8, 12] mencionan dos formas en las que se puede detectar plagio. La primera es realizando un análisis intrínseco, el cual solo busca cambios de estilos de escritura en un documento; y el basado en documentos de referencia, en donde se comparan documentos sospechosos contra documentos fuente.

Dentro de la detección automática de plagio basado en documentos de referencia se encuentra el alineamiento de texto, es una tarea en donde normalmente, los documentos contienen diferentes tipos de ofuscación, con la finalidad que el nuevo documento sea similar al original [13]. Generalmente los corpus que existen para la detección de plagio, implementan diferentes técnicas de ofuscación elaboradas por herramientas comerciales, en donde, hay variaciones semánticas de palabras, operaciones con el texto de forma aleatoria. Básicamente el plagio de los corpus es creado de manera artificial.

A partir del 2009 hasta la fecha según [14], PAN es la competición más grande de detección de plagio, identificación de autoría y mal uso de software social. Entre el 2012 y 2015 en la competición PAN, la tarea de detección de plagio se dividió en dos sub-tareas: recuperación de fuentes y alineamiento de texto. Para el alineamiento de texto [13, 14, 15], los sistemas deben identificar todos los pasajes de máxima longitud de texto reusado entre un par de documentos.

Como primera etapa de los enfoques mostrados des PAN 2012 a PAN 2015, se aplican distintas técnicas de pre-procesamiento, ésta es una etapa que se aplica en diversas tareas de PLN. En el caso de alineamiento de texto algunas técnicas que se han aplicado son:

- Eliminación de caracteres especiales: se refiere en [7, 16, 17, 18, 19] a remover los signos de puntuación y algunos caracteres que puedan generar ruido en el documento.
- Eliminación de números: en [20] la importancia de los dígitos no es prioridad y remueve los números que aparezcan en el texto.
- Eliminación de espacios en blanco: En [16, 21] se cambian los espacios en blanco de todo el texto por algún carácter, separando cada *token*, o simplemente se elimina el espacio en blanco.
- Conversión a mayúsculas: en [7] todo el texto se deja en un solo formato, con la finalidad de dejarlo normalizado.
- Conversión a minúsculas: en [1, 2, 18, 19, 20, 21, 22, 24, 25] todo el texto se deja en un solo formato, con la finalidad de dejarlo normalizado.
- Eliminación de *stopwords*: en [18, 19, 21, 22, 24] se menciona que son palabras que por sí solas no dicen nada del documento y pueden ser pronombres, artículos, preposiciones, etc.
- *Stemming*: en [1, 2, 19, 20, 21, 22, 26] obtiene la raíz de una palabra truncando una palabra en relación a otras. Por ejemplo sonrisa, sonríe, sonrían y sonreíste se obtiene la raíz *"sonri"*.
- División en *tokens*: en [1, 2, 18] el texto del documento es fragmentado, es decir, se divide en palabras o *tokens*.

En este artículo, se enfoca a la utilización de diferentes listas de *stopwords* en la etapa de pre-procesamiento, para saber cómo afecta la eliminación de esta información sobre las heurísticas de [1] y [2], las cuales están disponibles en código abierto[1].

[1] http://www.gelbukh.com/plagiarism-detection/PAN-2015

2. Estado del arte

En diversas aplicaciones del PLN se han hecho trabajos sobre pre-procesamiento uno de ellos es el de Ledeneva [27], en donde se analiza la importancia del pre-procesamiento, en la generación automática de resúmenes utilizando secuencias frecuentes maximales. Las técnicas de pre-procesamiento que utilizaron fueron análisis léxico como eliminación de signos de puntuación, normalización de números y algunas variantes de *stopwords* y *stemming*. Se detectó que al probar estas técnicas en el pre-procesamiento no afectan a la calidad de los resúmenes generados que comprueba que el método propuesto es bueno y no depende de la etapa de pro-procesamiento.

Al contrario en el trabajo de [28], se puede observar que al utilizar varias técnicas de pre-procesamiento, se mejoran los resultados considerablemente.

En la detección automática de plagio también se aplican técnicas de pre-procesamiento. En el trabajo de [7] se aborda la comparación de medidas de similitud en cadenas textuales, para identificar plagio en tareas escolares, en donde las técnicas de pre-procesamiento que utiliza son:

– Eliminación de números,
– Eliminación de espacios en blanco,
– Eliminación de signos de puntuación,
– Conversión a mayúsculas.

Aunque este trabajo se enfoca más a obtener una medida de similitud, es notable que inicialmente se aplique una etapa de pre-procesamiento, para reducir principalmente el ruido, que pueda ocasionar algunos caracteres.

Dentro de la competición internacional de plagio, identificación de autoría y mal uso de software social PAN, se desprende una sub-tarea que es alineamiento de texto en donde para la edición de PAN 2014 en [13] reportan que solo once participantes presentaron software para la evaluación y comparación, de los cuales solo diez reportaron la descripción de su enfoque.

Tabla 1. Técnicas de pre-procesamiento utilizadas por los participantes en la tarea de alineamiento de texto PAN 2014.

Pre-procesamiento	[1-2]	[23]	[29]	[16]	[17]	[18]
Eliminación de caracteres especiales	si	no	-	si	si	si
Eliminación de números	no	-	-	-	-	-
Eliminación de espacios en blanco	no	no	-	si	no	no
Conversión a mayúsculas	no	-	-	-	-	no
Conversión a minúsculas	si	no	-	no	no	si
Eliminación de *stopwords*	no	si	-	no	no	no
Stemming	si	no	-	no	no	no
División en *tokens*	si	no	-	no	no	si

Un análisis general de los métodos reportados, revela que para construir un algoritmo de alineamiento de texto se llevan a cabo las siguientes etapas: pre-procesamiento, pre-selección, extensión y filtrado. En esta edición la mayoría de los participantes se enfocó a predecir qué tipo de plagio se les presentaba. En la Tabla 1, se muestran las diferentes técnicas de pre-procesamiento utilizadas, se marcó con "-", cuando no se menciona que técnica se utilizó.

En la edición de PAN 2013 se reportó la participación de nueve equipos, los cuales presentaron su software para la evaluación y comparación, pero solo seis de ellos presentaron la descripción de su enfoque. De acuerdo a [30] algunos de los equipos que participaron en alineamiento de texto en PAN 2013 usaron las técnicas de pre-procesamiento que se muestran en la Tabla 2, así como las etapas mencionadas anteriormente que son pre-selección, extensión y filtrado. El corpus de entrenamiento tanto en PAN 2013 y PAN 2014 es el mismo.

Tabla 2. Técnicas de pre-procesamiento utilizadas por algunos participantes en la tarea de alineamiento de texto PAN 2013.

Pre-procesamiento	[19]	[22]	[24]	[25]	[20]
Eliminación de caracteres especiales	si	no	no	no	no
Eliminación de números	no	no	no	no	si
Eliminación de espacios en blanco	-	-	-	-	-
Conversión a mayúsculas	no	no	no	no	no
Conversión a minúsculas	si	si	si	si	si
Eliminación de *stopwords*	si	si	no	no	no
Stemming	si	si	no	no	si
División en *tokens*	-	-	-	-	-

Como se puede observar en la Tabla 1 y la Tabla 2, las técnicas de pre-procesamiento utilizadas, en la mayoría son las mismas. Los documentos del corpus de entrenamiento, son en texto plano. Cada equipo utiliza ciertas técnicas conforme al enfoque planteado. En la Tabla 3, se muestran los participantes que reportaron haber realizado alguna técnica de pre-procesamiento en PAN 2012, en general no se dice mucho del pre-procesamiento que realizaron.

Tabla 3. Técnicas de pre-procesamiento utilizadas por algunos participantes en la tarea de alineamiento de texto PAN 2012.

Pre-procesamiento	[21]	[26]
Eliminación de caracteres especiales	si	no
Eliminación de números	-	-
Eliminación de espacios en blanco	si	no
Conversión a mayúsculas	no	-
Conversión a minúsculas	si	no
Eliminación de *stopwords*	si	no
Stemming	si	si
División en *tokens*	-	-

En la primer edición de alineamiento de texto en el 2012 en donde en PAN reporta que once presentaron software para su evaluación y comparación, y solo diez reportaron la descripción de su enfoque, como se muestra en la Tabla 3 no hay muchos detalles acerca de técnicas empleadas de pre-procesamiento. El análisis de los métodos mostró

[15] algo en común, que fue las etapas de como construyen los algoritmos de alineamiento de texto las cuales son: pres-elección, fusión de partes y extracción filtrada. En general estos enfoques están basados en reglas, los cuales en cierta forma limitan la detección de plagio.

En general las técnicas que se utilizan en la preparación de los datos de entrada son muy similares en PAN 2012, PAN 2013 y PAN 2014. Inicialmente en PAN 2012 el corpus tenía diferentes tópicos los cuales se muestran a continuación y se explica con más detalle en [31]:

– No hay plagio (*no-plagiarism*),
– Ninguno (*no-obfuscation*),
– Artificial bajo (*artificial-low*),
– Artificial alto (*artificial-high*),
– Traducción (*translation*),
– Paráfrasis simulada (*simulated-paraphrase*).

Para PAN 2013 y PAN 2014 el corpus se conformó por los siguientes tópicos, siguiendo la misma forma de creación por parte de [31]:

– No hay plagio (*no-plagiarism*),
– Ninguno (*no-obfuscation*),
– Aleatorio (*summary-ofuscation*),
– Traducción (*translation-obfuscation*),
– Resúmenes (*summary-obfuscation*).

Como menciona en [31], los diferentes tipos de plagio que se crearon en su mayoría son elaborados de manera artificial, creados por herramientas comerciales, herramientas en línea, por operaciones de texto, etc.

En la edición de PAN 2015 la tarea de alineamiento de texto cambio un poco en cuanto a la temática, los equipos ahora debían elegir entre generar una colección con casos reales de reutilización de texto o plagio, o teniendo en cuenta los pares de documentos, generar pasajes de texto reutilizado o plagiados. Aplicando un tipo de ofuscación.

Debido a que el trabajo de [1] no se pudo comparar con otro de la edición PAN 2015, y había tenido una mejora del aplicado en PAN 2014 [2], no se declaró como ganador, sin embargo, analizando el enfoque de [1], utiliza las mismas técnicas de pre-procesamiento que [2], siendo hasta el momento el mejor en cuanto alineamiento de texto.

Las medidas de evaluación empleadas en PAN para alineamiento de texto son: *granularity, recall, presicion* y *plagdet.* En [12, 31] se dice que, d_{plg} denota un documento que contiene plagio. Un caso de plagio en d_{plg} es una 4-tupla $s = \{s_{\text{plg}}, d_{\text{plg}}, s_{\text{src}}, d_{\text{src}}\}$, donde, s_{plg} es un pasaje plagiado en d_{plg}, y s_{src} es el pasaje original correspondiente en el documento de referencia d_{src}. De forma similar, un caso de plagio detectado se expresa como $r = \langle r_{\text{plg}}, d_{\text{plg}}, r_{\text{src}}, d'_{\text{src}} \rangle$; donde r asocia un pasaje supuestamente plagiado r_{plg} en d_{plg} con un pasaje r_{src} en d'_{src}. Decimos que r detecta s si y solo si $r_{\text{plg}} \cap s_{\text{plg}} \neq \emptyset$ y $d'_{\text{src}} = d_{\text{src}}$.

Denotamos un documento d como un conjunto de referencias a sus caracteres d = $\{(1, d),\ldots, (|d|, \mathrm{d})\}$, donde (i, d) refiere el i-ésimo carácter en d. De esta forma un caso de plagio *s* puede ser representado como $\mathrm{s} = \mathrm{s}_{\mathrm{plg}} \cup \mathrm{s}_{\mathrm{src}}$, donde $\mathrm{s}_{\mathrm{plg}} \subseteq \mathrm{d}_{\mathrm{plg}}$ y $\mathrm{s}_{\mathrm{src}} \subseteq \mathrm{d}_{\mathrm{src}}$. Los caracteres referenciados en $\mathrm{s}_{\mathrm{plg}}$ y $\mathrm{s}_{\mathrm{src}}$ forman pasajes s_{plg} y s_{src} en la visión anterior. De forma similar una detección *r* puede ser representada como $r = r_{plg} \cup r_{src.}$ A partir menciona en [12] de esto podemos decir que *r* detecta *s* si y solo si $r_{plg} \cap s_{plg} \neq \emptyset$ y $r_{src} \cap S_{src} \neq \emptyset$. Por último, *S* y *R* denotan conjuntos de casos de plagios y detecciones respectivamente. Basado en estas representaciones *precision* y *recall* de *R* según *S* se define como:

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup s \in S (s \Pi r)}{|r|},$$

$$rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup r \in R (s \Pi r)}{|s|},$$

donde

$$s \Pi r = \{ s \cap r \; si \; r \; detecta \; s, \emptyset \; otro \; caso.$$

Además de *precision* y *recall* otro concepto importante que caracteriza la eficiencia de un algoritmo de detección de plagio, esto se refiere, si un caso de plagio es detectado como uno solo o en varias partes. Para esto se define *granularity* de *R* en *S*, en donde *S* y *R* denotan conjuntos de casos de plagio y de detecciones. Lo definen con la siguiente fórmula:

$$gran(S, R) = \frac{1}{|S_R|} \sum_{s \in S_R} |R_S|,$$

donde $S_R \subseteq S$ son los casos detectados por las detecciones de *R*, y $R_S \subseteq R$ son las detecciones de un caso *s* dado:

$$S_R = \{\mathrm{s} \mid \mathrm{s} \in S \wedge \exists\, r \in R : r \; detecta \; s\},$$

$$R_S = \{\mathrm{r} \mid \mathrm{r} \in R \wedge r \; \mathrm{detecta} \; s\}.$$

El dominio de *gran(S, R)* es [1,|*R*|], el 1 indica la correspondencia deseada uno a uno y |*R*| indica el peor de los casos, donde un solo caso $s \in S$, es detectado una y otra vez.

Teniendo en cuenta *precision*, *recall* y *granularity* que permiten un orden parcial entre los algoritmos de detección. Para obtener un orden general, estas medidas se combinan de la siguiente forma:

$$plagdet(S, R) = \frac{F_a}{\log_2 (1 + gran(S, R))}.$$

En donde F_a denota la Medida-F. Para las ediciones de PAN la medida armónica ponderada de *precision* y *recall* es $a = 1$, ya que no hay indicación de que una sea más importante que la otra.

3. Trabajo previo

De acuerdo a los equipos que participaron en alineamiento de texto en PAN 2014 y PAN 2015, en la Tabla 4, se reportan los mejores enfoques con el corpus de evaluación proporcionado por la competencia PAN.

Tabla 4. Resultados de [2], en alineamiento de texto PAN 2014.

Equipo	Plagdet	Recall	Precision	Granul.
Sanchez-Perez15 [1]	0.9010	0.8957	0.9125	1.0046
Sanchez-Perez14 [2]	0.8781	0.8790	0.8816	1.0034
Oberreuter [13]	0.8693	0.8577	0.8859	1.0036
Palkovskii [17]	0.8680	0.8263	0.9222	1.0058
Glinos [18]	0.8593	0.7933	0.9625	1.0169

Tomando como referencia los enfoques [1, 2, 16, 17, 18, 19, 20, 22, 23, 24, 25] en general se deducen las diferentes heurísticas en cuatro etapas: pre-procesamiento, pre-selección, extensión y filtrado.

Pre-procesamiento. Se refiere a preparar el texto antes de ser procesado en las siguientes etapas se puede hacer, por ejemplo: eliminar caracteres no alfanuméricos, eliminar palabras vacías, por mencionar algunas.

Pre-selección. Teniendo un documento sospechoso y un documento origen, el texto se divide en fragmentos, con el fin de encontrar coincidencias en ambos textos.

Extensión. En general, en esta etapa trata de formar pasajes de máxima longitud, con la unión de fragmentos, tanto en el documento origen como en el sospechoso.

Filtrado. Teniendo los pasajes alineados, se eliminan los que no cumplen ciertos criterios, esto se hace con el fin de maximizar el rendimiento de cada método.

Se cuenta con los trabajos de [1-2], los cuales están enfocados a la tarea de alineamiento de texto, en la Figura 1, se muestran las etapas del enfoque de [2].

En la Tabla 5, se muestran varios parámetros los cuales se enlistan a continuación:

- minSentLength: es la longitud mínima de un fragmento de texto, que son 3 palabras,
- th1: corresponde al parámetro de la similitud coseno,
- th2: corresponde al parámetro de la similitud de Dice,
- th3: corresponde al parámetro de similitud de estructura,
- minSize: longitud de pasajes pequeños,
- minPlagLength: longitud de un pasaje para ser considerado plagio,
- maxGap: define cuál es la separación máxima, en número de oraciones, que puede existir en dos oraciones seleccionadas para ser consideradas adyacentes.

Tabla 5. Lista de parámetros de la heurística [2].

Parámetro	Valor	Parámetro	Valor
minsentlen	3	maxgap	4
th_1	0.33	maxgap_least	2
th_2	0.33	minsize	1
Th_3	0.40	minplaglen	150

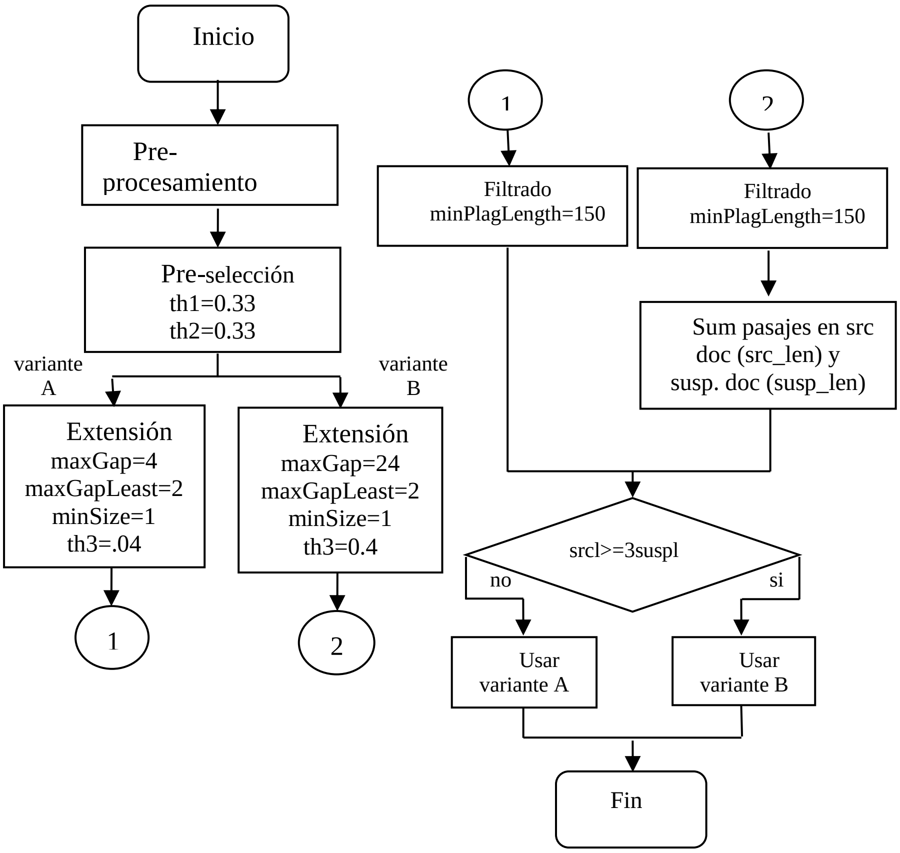

Inicio
Pre-procesamiento
Pre-selección th1=0.33 th2=0.33
variante A
variante B
Extensión maxGap=4 maxGapLeast=2 minSize=1 th3=.04
Extensión maxGap=24 maxGapLeast=2 minSize=1 th3=0.4
1
2
1
2
Filtrado minPlagLength=150
Filtrado minPlagLength=150
Sum pasajes en src doc (src_len) y susp. doc (susp_len)
srcl>=3suspl
no
si
Usar variante A
Usar variante B
Fin

Fig. 1. Enfoque de [2], para la detección automática de plagio mediante alineamiento de texto PAN 2014.

El enfoque de [2] fue el que obtuvo el mejor desempeño en la tarea de alineamiento de texto, en la competencia de detección plagio PAN 2014. Para el 2015, el trabajo de [1] mejoró su enfoque, superando sus resultados obtenidos un año atrás, este enfoque se muestra en la Figura 2.

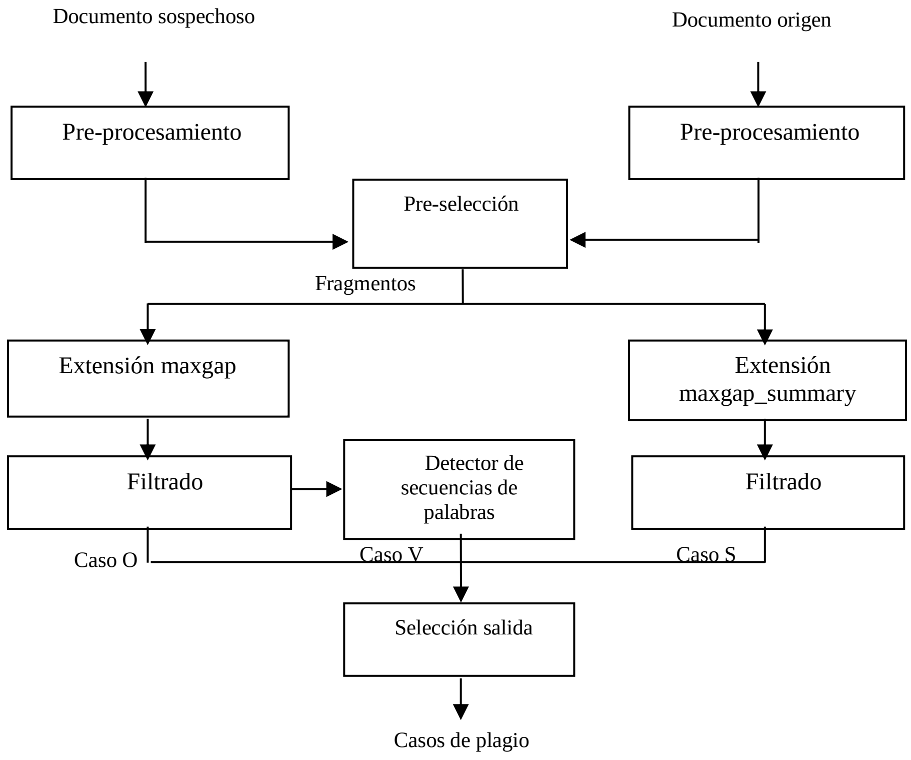


Fig. 2. Enfoque de [1], para la detección automática de plagio mediante alineamiento de texto PAN 2014.

Tabla 6. Lista de parámetros de la heurística [1].

Parámetro	Valor	Parámetro	Valor
minsentlen	3	maxgap_summary	24
th_cos	0.30	maxgap_least	0
th_dice	0.33	minsize	1
th_validation	0.34	minplaglen	150
maxgap	4	th_verbatim	256

Para [1], su enfoque va dirigido a la sub-área del corpus, *summary obfuscation*, siendo una de las más difíciles para detectar plagio. En la Tabla 4, se muestran los mejores resultados en el corpus de evaluación, dentro de la competencia PAN 2014 para alineamiento de texto, el enfoque de [1] es una mejora del enfoque [2], utilizando el corpus de PAN 2014. Los parámetros del enfoque de [1] son los que se muestran en la Tabla 6.

La descripción de cada parámetro se muestra a continuación:

- th_cos: corresponde al parámetro de la similitud coseno,
- th_dice: corresponde al parámetro de la similitud Dice,
- th_validation: corresponde al umbral de validación,
- maxgap: separación máxima de oraciones que son consideradas adyacentes,
- maxgap_summary: separación máxima de oraciones que son consideradas adyacentes, enfocado a la parte de ofuscación en
- minsize: longitud de pasajes pequeños,
- minplaglen: longitud de un pasaje para ser considerado plagio,
- th_verbatim: longitud de secuencia de palabras en la etapa de filtrado.

4. Metodología propuesta

Método general. En cada experimento, seguimos los siguientes pasos:

- Pre-procesamiento: se mantuvieron las técnicas de los métodos de [1, 2] los cuales son los siguientes:

    - Conversión a minúsculas,
    - Eliminación de caracteres especiales,
    - *Stemming*,
    - División en tokens.

  En cuanto a la implementación de *stopwords* en los enfoques de [1, 2] no se eliminan *stopwords*. En nuestro enfoque es en donde probamos diferentes listas de *stopwords ShortList* y *BigList* (las listas contiene diferente cantidad de *stopwords*). Implementamos estas dos listas en los enfoques de [1, 2], así como las dos listas que ellos mencionan en [12].

  El uso de la de técnica *stemming* que implementa [1, 2] decidimos no utilizarla en algunos experimentos, para conocer su importancia dentro de los enfoques trabajados.

- Pre-selección: Se utilizó la fragmentación de texto, tanto en los documentos sospechosos como el origen.

- Extensión: Teniendo los fragmentos de texto se procede a agruparlos, formando pasajes de máxima longitud. Para [1] se agrega una etapa de validación de los grupos creados.

- Filtrado: teniendo los pasajes alineados, se eliminan los que no superan la longitud de 150 caracteres, también se eliminan pasajes solapados.

Corpus de entrenamiento. Utilizamos el corpus de entrenamiento proporcionado por la competencia PAN 2014, en la sub-tarea alineamiento de texto, el cual está disponible[2]. El corpus está conformado por 5185 pares de documentos sospechosos de plagio en idioma inglés, están divididos en 5 formas de plagio: no hay plagio (*no plagiarism),* ninguno *(no obfuscation),* aleatorio *(random obfuscation),* traducción *(translation obfscation)* y resúmenes (*summary obfuscation).* En [31] menciona que algunas formas de plagio se realizaron con herramientas comerciales y en algunos casos se llegó a perder la coherencia de los textos. A continuación se dará una breve descripción de cada sub-área del corpus:

– No hay plagio (*no-plagiarism*). Esta parte del corpus no contiene ningún tipo de plagio.

– Ninguno (*no-obfuscation*). La ofuscación que se presenta solo es de *copy-paste.*

– Aleatorio (*summary-ofuscation*). Es una secuencia de operaciones de texto al azar, añadir, eliminar y reemplazar palabras o frases cortas en todo el texto.

– Traducción (*translation-obfuscation*). El texto se tradujo en por lo menos tres idiomas con diferentes herramientas comerciales, siendo inglés el idioma inicial y final.

– Resúmenes (*summary-obfuscation*). Incluye un resumen no atribuido en otro documento, las ideas principales del documento se mantienen. Puede ser visto como una forma de plagio de ideas.

Tabla 7. Resultados reportados por [2], en alineamiento de texto PAN 2014 sin eliminación de stopwords.

Resultados PAN 2014 sin *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.8938	0.9782	0.8228	1.0000
Aleatoria	0.8886	0.8581	0.9213	1.0000
Traducción	0.8839	0.8902	0.8777	1.0000
Resúmenes	0.5772	0.4247	0.9941	1.0434
Total	**0.8773**	0.8799	0.8774	1.0021

Evaluación. El marco de evaluación fue propuesto por [31], en donde se propone una medida (*plagdet*) que está en función de precisión, recuerdo y granularidad, lo cual ya se explicó anteriormente.

La Tabla 7 y la Tabla 8, muestran los resultados obtenidos en los diferentes enfoques para PAN 2014 [2] y PAN 2015 [1] utilizando el corpus de entrenamiento, estos resultados se encuentran reportados en [1], el resultado final se encuentra en negritas.

[2] http://pan.webis.de/clef14/pan14-web/plagiarism-detection.html

Los resultados obtenidos utilizando la lista de las 50 stopwords más frecuentes en inglés, reportada por Stamatatos en [32], para el trabajo de [2] los resultados se muestra en la Tabla 9 y el para el trabajo de [1] los resultados obtenidos se muestran en la Tabla 10.

Tabla 8. Resultados reportados por [1], en alineamiento de texto PAN 2015 sin eliminación de stopwords.

Resultados PAN 2015 sin *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9812	0.9761	0.9933	1.0048
Aleatoria	0.8847	0.8699	0.8999	1.0000
Traducción	0.8792	0.9128	0.8481	1.0000
Resúmenes	0.6304	0.4862	0.9739	1.0404
Total	**0.9025**	0.8937	0.9164	1.0036

5. Resultados experimentales

En el trabajo de [2], se encuentran dos listas de *stopwords* que en sus experimentaciones no reporta, nos dimos a la tarea de realizar la experimentación con estas dos listas de *stopwords*, una está descrita como las 50 *stopwords* más frecuentes en inglés por Stamatatos en [32] y la otra está contenida en el paquete NLTK de Python, la cual se encuentra en el trabajo de [12], se utilizó el corpus de entrenamiento de PAN 2014 en todos los experimentos.

En esta sección presentaremos en primer lugar los resultados reportados utilizando el trabajo de [2] en PAN 2014. Los resultados obtenidos utilizando la lista de las 50 *stopwords* más frecuentes en inglés, reportada por Stamatatos en [32], se muestran en la Tabla 9. Los parámetros utilizados en los resultados de la Tabla 9 a la Tabla 13 para PAN 2014 se encuentran descritos en la Tabla 5.

Tabla 9. Resultados obtenidos de [2], en alineamiento de texto PAN 2014 eliminando las 50 *stopwords* reportadas por Stamatatos [32].

Resultados PAN 2014 50 *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9002	0.9722	0.8380	1.0000
Aleatoria	0.8653	0.8099	0.9288	1.0000
Traducción	0.8768	0.8687	0.8850	1.0000
Resúmenes	0.4890	0.3530	0.9939	1.0924
Total	**0.8662**	0.8518	0.8866	1.0043

Con la lista de *stopwords* del paquete de Python NLTK, para el trabajo de [2] los resultados se muestran reportados en la Tabla 10.

Tabla 10. Resultados obtenidos de [2], en alineamiento de texto PAN 2014 eliminando las *stopwords* del corpus NLTK.

Resultados PAN 2014 NLTK *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.8968	0.9707	0.8334	1.0048
Aleatoria	0.8482	0.7846	0.9231	1.0000
Traducción	0.8677	0.8600	0.8755	1.0000
Resúmenes	0.4799	0.3517	0.9842	1.1136
Total	**0.8563**	0.8405	0.8797	1.0055

Las listas de *stopwords* que implementamos, en primer lugar en la Tabla 11, mostramos los resultados utilizando la lista de *stopwords ShortList.* La Tabla 12, muestra los resultados obtenidos con la lista de *stopwords BigList*.

Tabla 11. Resultados obtenidos de [2], utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, con la lista de stopwords *ShortList.*

Resultados PAN 2014 con *stopwords ShortList*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.8906	0.9696	0.8235	1.0000
Aleatoria	0.8442	0.7826	0.9163	1.0000
Traducción	0.8677	0.8586	0.8770	1.0000
Resúmenes	0.4976	0.3628	0.9789	1.0903
Total	**0.8541**	0.8398	0.8745	1.0043

Tabla 12. Resultados obtenidos del [2] del utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, con la lista de stopwords *BigList*.

Resultados PAN 2014 con *stopwords BigList*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.8818	0.9683	0.8096	1.0000
Aleatoria	0.8294	0.7574	0.9165	1.0000
Traducción	0.8621	0.8598	0.8644	1.0000
Resúmenes	0.5127	0.3829	0.9681	1.1000
Total	**0.8455**	0.8330	0.8647	1.0049

Se realizó un experimento sin implementar *stemming* ni eliminar *stopwords*, los resultados se muestran en la Tabla 13.

Tabla 13. Resultados obtenidos de [2], utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, sin *stopwords* ni *stemming.*

Resultados PAN 2014 sin *stemming*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9054	0.9785	0.8425	1.0048
Aleatoria	0.8891	0.8515	0.9302	1.0000
Traducción	0.8624	0.8286	0.8992	1.0000
Resúmenes	0.4969	0.3450	0.9932	1.0429
Total	**0.8721**	0.8540	0.8934	1.0019

A continuación se mostrarán los resultados utilizando el trabajo de [1] en PAN 2015, los parámetros utilizados de la Tabla 14 a la 18 se encuentran descritos en la Tabla 6. Los resultados obtenidos utilizando la lista de las 50 *stopwords* más frecuentes en inglés, reportada por Stamatatos en [32], se muestran en la Tabla 14. Los resultados finales se enmarcan con negritas en cada tabla.

Tabla 14. Resultados obtenidos de [1], en alineamiento de texto PAN 2015 eliminando las 50 *stopwords* reportadas por Stamatatos [32].

Resultados PAN 2015 50 *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9812	0.9761	0.9933	1.0048
Aleatoria	0.8847	0.8701	0.8998	1.0000
Traducción	0.8791	0.9128	0.8477	1.0000
Resúmenes	0.6304	0.4862	0.9739	1.0404
Total	**0.9025**	0.8937	0.9163	1.0036

Con la lista de *stopwords* del paquete de Python NLTK, para el trabajo de [2] los resultados se muestran reportados en la Tabla 15.

Tabla 15. Resultados obtenidos de [1], en alineamiento de texto PAN 2015 eliminando las *stopwords* del corpus NLTK.

Resultados PAN 2015 NLTK *stopwords*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9812	0.9761	0.9933	1.0048
Aleatoria	0.8846	0.8701	0.8996	1.0000
Traducción	0.8794	0.9128	0.8484	1.0000
Resúmenes	0.6304	0.4862	0.9739	1.0404
Total	**0.9026**	0.8937	0.9165	1.0036

En la Tabla 16, mostramos los resultados utilizando la lista de *stopwords ShortList* y en la Tabla 17, reportamos los resultados obtenidos con la lista de *stopwords BigList.*

Tabla 16. Resultados obtenidos de [1], utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, con la lista de stopwords *ShortList*.

Resultados PAN 2015 con *stopwords ShortList*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9812	0.9761	0.9933	1.0048
Aleatoria	0.8851	0.8702	0.9006	1.0000
Traducción	0.8798	0.9125	0.8494	1.0000
Resúmenes	0.6291	0.4848	0.9737	1.0406
Total	**0.9028**	0.8936	0.9171	1.0036

Tabla 17. Resultados obtenidos de [1], utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, con la lista de stopwords *BigList*.

Resultados PAN 2015 con *stopwords BigList*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9812	0.9761	0.9933	1.0048
Aleatoria	0.8860	0.8705	0.9020	1.0000
Traducción	0.8799	0.9101	0.8517	1.0000
Resúmenes	0.6289	0.4846	0.9737	1.0406
Total	**0.9031**	0.8929	0.9183	1.0036

De igual forma como se hizo anteriormente, se realizó un experimento sin la implementación de *stemming* en el texto de entrada ni la eliminación de *stopwords*, los resultados obtenidos se encuentran reportados en la Tabla 18.

Tabla 18. Resultados obtenidos de [1], utilizando el corpus de entrenamiento PAN 2014 en alineamiento de texto, sin *stopwords* ni *stemming.*

Resultados PAN 2015 sin *stemming*				
Ofuscación	*Plagdet*	*Recall*	*Precision*	*Granul.*
Ninguna	0.9820	0.9758	0.9952	1.0048
Aleatoria	0.8879	0.8590	0.9188	1.0000
Traducción	0.8709	0.8641	0.8788	1.0008
Resúmenes	0.5726	0.4205	0.9775	1.0376
Total	**0.8990**	0.8710	0.9340	1.0037

6. Conclusiones

En este artículo, se experimentó la implementación de diferentes listas de *stopwords*, en la competición internacional de plagio PAN 2014 y PAN 2015, para la sub-tarea alineamiento de texto, tomando los enfoques de [1, 2], en la etapa de pre-procesamiento, para conocer el efecto de la ausencia de información en el corpus de entrenamiento, en este caso de *stopwords*, que es la técnica en donde se elimina más información de los documentos originales.

Las listas de *stopwords* utilizadas las llamamos *ShortList* y *BigList,* debido a la cantidad de *stopwords* que contiene cada lista, así como las propuestas en [12] que son dos listas: una es las 50 *stopwords* más comunes en inglés propuesta por Stamatatos [32] y la segunda es la que está incluida en la librería de Python NLTK. Tambien se reportaron resultados cuando no se implementó *stemming.*

En la Tabla 19 se muestran los resultados finales de nuestras experimentaciones de PAN 2014 y PAN 2015 con el corpus de entrenamiento de PAN 2014, comparando los resultados obtenidos en los trabajos de [1, 2], los mejores resultados se muestran resaltados en negritas.

Tabla 19. Resultados obtenidos en PAN 2014 y PAN 2015 en comparación con los reportados en los trabajos de [1] y [2].

Resultados PAN 2014		Resultados PAN 2015	
Experimentos	*Plagdet*	Experimentos	*Plagdet*
Sin eliminar *stopwords*[2]	**0.8773**	Sin eliminar *stopwords*[1]	0.9025
Eliminación de 50 *stopwords*	0.8662	Eliminación de 50 *stopwords*	0.9025
Eliminación *stopwords* NLTK	0.8563	Eliminación *stopwords* NLTK	0.9026
Stopwords ShortList	0.8541	*Stopwords ShortList*	0.9028
Stopwords BigList	0.8455	*Stopwords BigList*	**0.9031**
Sin *Stemming*	0.8721	Sin *Stemming*	0.8990

Como se puede observar, para PAN 2014 el mejor resultado es el reportado por [2] en el corpus de entrenamiento, para ese trabajo lo mejor es mantener la mayor cantidad de información, así como aplicar *stemming*, esto da resultados favorables. Para PAN 2015 el mejor resultado es con la eliminación de *stopwords* de la lista *BigList*. Con la eliminación de *stopwords*¸ se logra una mejor detección de pasajes similares, sin olvidar que la implementación de *stemming* también juega un papel importante en la preparación del texto, eliminando caracteres que no son importantes a lo largo del proceso de alineamiento de texto.

Como se describió en el estado del arte la mayoría de los equipos de detección de plagio mediante alineamiento de texto, utilizan técnicas muy similares de pre-procesamiento, ocho son las técnicas que se identificaron entre trabajos para alineamiento de texto entre PAN 2012 y PAN 2015 [33, 34, 35]. Es muy común el uso de técnicas de pre-procesamiento en tareas de detección de plagio, pero no hay un estudio sobre cómo estas técnicas afectan los resultados finales.

Como trabajo futuro se tiene enfocarse a las siguientes etapas del método y en conjunto con el estado del arte, saber que técnicas nos podrían ayudar a tener mejores resultados en comparación con los de [1]. La etapa de pre-selección y extensión son clave para la formación de pasajes plagiados en los pares de documentos.

La implementación de características de n-gramas sintácticos en el texto [36, 37] o distintos tipos de n-gramas [17, 38, 39] en la etapa de preselección, puede ayudar a tener una mejor generación de fragmentos para formar pasajes plagiados. La implementación de diversos agrupadores en la etapa de extensión en combinación con la etapa de pre-selección, es algo que a lo largo de las competiciones de PAN 2012 y PAN 2014 se ha presentado, y también es una opción implementar y analizar algunos agrupadores [40, 41], Si bien ya se han registrado resultados favorables, la experimentación con nuevas heurísticas, nos abrirá el panorama, en cuanto la utilidad de diferentes técnicas tanto para resolver este problema como algunos otros que se relacionen.

Referencias

1. Sánchez-Pérez, M.A., Gelbukh, A.F., Sidorov, G.: Dynamically adjustable approach through obfuscation type recognition. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR Workshop Proceedings, vol. 1391, CEUR-WS.org, http://ceur-ws.org/Vol-1391/92-CR.pdf (2015)
2. Sánchez-Pérez, M.A., Gelbukh, A.F., Sidorov, G.: The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014. In: L. Cappellato, N. Ferro, M. Halvey, W. Kraaij (eds.). Notebook for PAN at CLEF 2014. CLEF 2014. CLEF2014 Working Notes. Sheffield, UK, September 15-18. CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 1180, CEUR-WS.org, 2014, pp. 1004–1011 (2014)
3. Mateos, F. J., Ruiz, J. L.: Procesamiento del lenguaje natural. Dpto. Ciencias de la Computación e Inteligencia Artificial; Universidad de Sevilla. http://www.cs.us.es/cursos/ia2/temas/tema-06.pdf (2012)
4. Baeza-Yates, R., Ribeiro-Neto, B.: Modern information retrieval. Vol. 463, New York: ACM press (1999)
5. Salton, G., McGill, M. J.: Introduction to modern information retrieval. (1983)
6. Real Academia Española.: Diccionario de la lengua española. Vigésima tercera edición, 2014, http://www.rae.es/ Consultado Agosto (2016)
7. Armeaga, G.J.: Comparación de medidas de similitud en cadenas textuales, para la detección de plagio en tareas escolares. Unidad académica Profesional Tianguistenco, Universidad Autónoma del Estado de México, Tesis de licenciatura (2015)
8. Barrón Cedeño, L. A.: Detección automática de plagio en texto. Departamento de Sistemas Informáticos y Computación, Tesis desarrollada dentro del Máster en Inteligencia Artificial, Universidad Politécnica de Valencia (2008)
9. Instituto Nacional del Derecho de Autor.: Derechos de Autor; México; http://www.indautor.gob.mx/accesibilidad/accesibilidad_autor.html, Consultado Mayo (2015)
10. Sánchez Vega, J. F.: Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado. Tesis sometida como requisito parcial para obtener el grado de: Maestro en Ciencias en el Área de Ciencias Computacionales Instituto Nacional de Astrofísica, Óptica y Electrónica (2011)
11. Balaguer, E. V.: Putting ourselves in SME's shoes: Automatic detection of plagiarism by the WCopyFind tool. In: Proc. SEPLN, pp. 34–35 (2009)
12. Sánchez-Pérez, M.A.: Detección automática de plagio a través de formación de pasajes. Centro de Investigación en Computación, Instituto Politécnico Nacional, Tesis de Maestría (2014)
13. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection (2014)
14. Elizalde, V.: Estudio y desarrollo de nuevos algoritmos de detección de plagio. Doctoral dissertation, Tesis de Licenciatura en Ciencias de la Computación Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (2011)
15. Potthast, M., Gollub, T., Hgaen, M., GraBegger, J., Kiesel, J., Michel, M., Oberlander, A., Tippman, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International competition on Plagiarism Detection. In: Forner et al. [33].

16. Alvi, F., Stevenson, M., Clough, P. D.: Hashing and Merging Heuristics for Text Reuse Detection. In: Cappellato et al. [35].
17. Palkovskii, Y., Belov, A.: Developing High-Resolution Universal Multi-Type N-Gram Plagiarism Detector. In: Cappellato et al. [35].
18. Glinos, D. S.: A Hybrid Architecture for Plagiarism Detection. In: Cappellato et al. [35].
19. Leilei, K., Haoliang, Q., Cuixia, D., Mingxing, W., Zhongyuan, H.: Approaches for Source Retrieval and Text Alignment of Plagiarism Detection. In: Forner et al. [34]
20. Palkovskii, Y., Belov, A.: Using Hybrid Similarity Methods for Plagiarism Detection. In: Forner et al. [34].
21. Leilei, K., Haoliang, Q., Shuai, W., Cuixia, D., Suhong, W., Yong, H.: Approaches for candidate document retrieval and detailed comparison of plagiarism detection. In: Forner et al. [33].
22. Torrejón, D. A. R., Ramos, J. M. M.: Text Alignment Module in CoReMo 2.1 Plagiarism Detector. In: Forner et al. [34].
23. Abnar, S., Dehghani, M., Zamani, H., Shakery, A.: Expanded n-grams for semantic text alignment. In: Cappellato et al. [35].
24. Suchomel, Š., Kasprzak, J., Brandejs, M.: Diverse queries and feature type selection for plagiarism discovery. In: Forner et al. [34]
25. Shrestha, P., Solorio, T.: Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism. In: Forner et al. [34].
26. Palkovskii, Y., Belov, A.: Applying Specific Clusterization and Fingerprint Density Distribution with Genetic Algorithm Overall Tuning in External Plagiarism Detection. In: CLEF, (Online Working Notes/Labs/Workshop) (2012)
27. Ledeneva, Y.: Effect of preprocessing on extractive summarization with maximal frequent sequences. In: Mexican International Conference on Artificial Intelligence, pp. 123–132, Springer Berlin Heidelberg (2008)
28. Hassan, S., Mihalcea, R., Banea, C.: Random walk term weighting for improved text classification. International Journal of Semantic Computing, 1(04), pp. 421–439 (2007)
29. Leilei, K., Yong, H., Zhongyuan, H., Haihao, Y., Qibo, W., Tinglei, Z., Haoliang, Q.: Source Retrieval Based on Learning to Rank and Text Alignment Based on Plagiarism Type Recognition for Plagiarism Detection. In: Cappellato et al. [35].
30. Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: Forner et al. [34].
31. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An evaluation framework for plagiarism detection. In: Proceedings of the 23rd international conference on computational linguistics: Posters, Association for Computational Linguistics, pp. 997–1005 (2010)
32. Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology, 62(12), pp. 2512–2527 (2011)
33. Forner, P., Karlgren, J., Womser-Hacker, C.: CLEF 2012: Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy (2012).
34. Forner, P., Navigli, R., Tu s, D., Ferro, N.: Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26, 2013, CEUR Workshop Proceedings, vol. 1179. CEUR-WS.org (2013)
35. Cappellato, L., Ferro, N., Halvey, M., Kraaij, W.: Notebook for PAN at CLEF 2014. CLEF 2014. CLEF2014 Working Notes. Sheffield, UK, September 15-18, 2014. CEUR Workshop Proceedings, ISSN 1613-0073, Vol. 1180, CEUR-WS.org (2014)
36. Sidorov, G.: Construcción no lineal de n-gramas en la lingüística computacional: n-gramas sintácticos, filtrados y generalizados. Sociedad Mexicana de Inteligencia Artificial (2013)
37. Posadas-Durán, J. P., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., Pichardo-Lagunas, O.: Syntactic N-grams as Features for the Author Profiling Task. In: Proceedings of CLEF (2015)

38. Efstathios, S.: Plagiarism Detection Using Stopword n-Grams. JASIST, 62(12), pp. 2512–2527 (2011)
39. Barrón-Cedeño, A., Rosso, P.: On automatic plagiarism detection based on n-grams comparison. In: European Conference on Information Retrieval, Springer Berlin Heidelberg, pp. 696–700 (2009)
40. Kryszkiewicz, M., Lasek, P.: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. In: International Conference on Rough Sets and Current Trends in Computing, Springer Berlin Heidelberg, pp. 60–69 (2010)
41. Smiti, A., Eloudi, Z.: Soft dbscan: Improving dbscan clustering method using fuzzy set theory. In: 6th International Conference on Human System Interactions (HSI), IEEE, pp. 380–385 (2013)

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
diciembre de 2016
Printing 500 / Edición 500 ejemplares