Avances en la Ingeniería del Lenguaje y del Conocimiento

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 97, octubre de 2015. 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 97, October 2015. Printing 500. The authors are responsible for the contents of their articles.

Printed in Mexico City, in the IPN Graphic Workshop – Publication Office.

Avances en la Ingeniería del Lenguaje y del Conocimiento

David Pinto
Darnes Vilariño (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2015

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

Indexed in LATINDEX and Periodica / Indexada en LATINDEX y Periódica

Printing: 500 / Tiraje: 500

Printed in Mexico / Impreso en México

Editorial

En el presente volumen se publican una serie de artículos seleccionados, los cuales fueron previamente presentados en el simposio en Ingeniería del Lenguaje y del Conocimiento (LKE‘2015), en particular en la tercera edición de esta serie de eventos. Esta conferencia ha sido organizada en el seno de la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla (BUAP) por tres años consecutivos. Nace como una iniciativa del laboratorio de Ingeniería del Lenguaje y del Conocimiento con la finalidad de ofrecer un espacio académico y de investigación, en el cual sea posible reportar trabajos relacionados con el área. Este evento promueve la cooperación entre diferentes grupos de investigación, pues permite el intercambio de resultados científicos, prácticos y la generación de nuevo conocimiento.

Esta edición especial de la Revista Research in Computing Science contiene entonces una serie de contribuciones originales que han sido seleccionadas a partir de un proceso de evaluación ciega doble (double blind), lo cual significa que los nombres de los autores de los artículos y los nombres de los revisores son ambos desconocidos. Este procedimiento es ejecutado en aras de proveer una evaluación anónima, que derive en artículos de mayor calidad para este volumen; particularmente, en esta ocasión la tasa de rechazo fue del 22%, cuidando que en todos los casos, al menos dos especialistas del comité revisor hicieran una evaluación de la pertinencia, originalidad y calidad de cada artículo sometido.

Esperamos que este volumen sea de utilidad para el lector y los autores de los artículos seleccionados encuentren en esta edición especial un espacio de intercambio científico productivo que enriquezca la colaboración entre estudiantes y académicos en el ámbito de la ingeniería del lenguaje y del conocimiento.

El proceso de revisión y selección de artículos se llevó a cabo usando el sistema libremente disponible llamado EasyChair, http://www.easychair.org.

David Eduardo Pinto Avendaño
Darnes Vilariño Ayala

Octubre 2015

Table of Contents

Page

TEMANTEX: A Markup Language for Spanish Temporal Expressions and Indicators

Dina Wonsever, Aiala Rosá, Marisa Malcuori, Mathias Etcheverry

Universidad de la Republica, Instituto de Computacion, Facultad de Ingenieria, Uruguay

{wonsever, aialar, mathiase}@fing.edu.uy, marisamalcuori@gmail.com

Abstract. We describe the TEMANTEX annotation scheme for temporal expressions and other lexical indicators of temporality and we analyze a first annotation experience. TEMANTEX is mainly a revision of the markup language TIMEX3, but with some additions and a different treatment for relative expressions. Our alternative proposal is justified for two reasons. First, our system aims to cover other temporality-related lexical elements by defining annotations for what we call temporal indicators, which do not have an equivalent in the TimeML system. Second, regarding temporal expressions, our scheme has relevant differences that improve the annotation process and the interpretation potential. A first task of corpus annotation on a set of 2.300 words, comprising 33 temporal expressions and 35 temporal indicators, showed encouraging results.

Keywords: Markup language, temporal expressions, annotation.

1 Introduction

This document describes the TEMANTEX annotation scheme for temporal expressions and other lexical indicators of temporality and analyzes some annotation experiences. There exist several annotation schemes for temporal expressions, and specifically TIMEX3, included in the TimeML [2] annotation scheme has been widely used, having adaptations for several languages. There are annotated corpora, mainly for English [3], but other languages have been incorporated more recently: Chinese, French, Italian, Korean and Spanish data were incorporated at the TempEval 2 [4]; and enhanced English and Spanish [6] corpora were provided for TempEval 3 [7].

Our alternative proposal is justified for two reasons. First, our system aims to cover other temporality-related lexical elements by defining annotations for what we call temporal indicators, which do not have an equivalent in the TimeML system. For instance, we annotate as temporal indicators terms like *previo/previous*, *siguiente/subsequent* that are not annotated in the Spanish TimeBank, as stated in the annotation guidelines ([5], section 3.2.2). Second, regarding temporal expressions, our scheme has relevant differences that improve the annotation process and the interpretation potential.

Temporal expressions or eTemps (section 2) are linguistic expressions that refer to timeline allocated intervals (or sets of intervals) or to temporal durations. Temporal expressions may include various types of calendar units (parts of a day, days, months, years, etc.), which can have an accurate or vague, absolute or relative reference. They can also consist of terms denoting lengths of time which are usually presented as vague or generic.

Temporal indicators or mTemps, from *marca temporal* in Spanish, (section 3) are an heterogeneous set of elements that influence the temporal interpretation of the text and which are neither temporal expressions *per se* nor are they included within a temporal expression. Unlike eTemps, which refer to intervals or durations of varying lengths on the timeline, mTemps are relational elements or lexical temporality indicators. We have classified in several types the temporal indicators: relational, stages, ordinal numbering, duration, frequency and relative.

TEMANTEX attempts to capture and categorize all the information that might enhance the task of automatic learning of the expressions and the temporal relations in a text. The temporal anchoring of events has multiple applications in tasks such as multi-document summarization, question answering, information retrieval. One of the distinguishing features of our scheme is that it remains as close as possible to the text, excluding the calculation of absolute values made by the annotator through the attribute VALUE, as it happens with TIMEX3. Our scheme also includes a VALUE attribute for which no exact calculations are needed.

A first task of corpus annotation on a set of 2.300 words, comprising 33 eTemps and 35 mTemps, showed encouraging results (section 4).

2 Temporal Expressions

To annotate temporal expressions, we define an eTemp element with the following possible attributes and values, which are explained below:

ATTRIBUTE	VALUES
Type	location, duration, frequency
Mode of reference	absolute, relative
Degree of accuracy	accurate, vague
Value	expression in LDT language
Focus[1]	deictic, anaphoric, age reference, a different speaker
Granularity	second, minute, hour, day, etc.

To mark up an eTemp in the text we must first consider which text segment represents it. In this case we decided to include, as part of the eTemp, all the elements that might contribute to the expression interpretation, as is the case of prepositions and adverbs: <*en la Edad Media/in the middle Ages*>, <*durante dos horas/during two hours*>.

[1] The focus only appears for relative expressions.

For expressions like *el día del tsunami en Japón* (*the day of the tsunami in Japan*), the segment *el día* (*the day*) will be marked up as eTemp. This is a relative expression where the focus is *el Tsunami en Japón* (*the Tsunami in Japan*). This way of working enables us to exclude from eTemps all the elements that are not part of the temporal language and to which we cannot attribute a value.

1.1 TYPE, ACCURACY and MODE OF REFERENCE Attributes

A type is assigned to eTemps by selecting one of the three values of the attribute TYPE: location (chronological location on a timeline: *el 20 de abril / on 20th April*, *el 20 de abril de 1980 / on 20th April, 1980*, *los últimos diez años / the last ten years*, *hace un año / a year ago*), duration (length of time without anchoring to a given point in time: *Esperó durante una hora / She waited for an hour*), and frequency (location on a timeline for a repeated event: *los jueves de 2 a 4 / Thursdays from 2 to 4*, *a menudo / often*). For all cases, the expression refers to a temporal interval or set of temporal intervals (or point in time as an extreme case).

ACCURACY shows if the temporal interval is fully specified (*hoy/today*, *4 horas / 4 hours*, *todos los martes / every Tuesday*) or if the reference is imprecise (*en estos días / these days*, *durante mucho tiempo / for a long time*, *a veces / sometimes*).
MODE OF REFERENCE shows if the expression is absolute, i.e., it does not require additional elements to be interpreted; or relative, i.e., it depends on an additional element, in the text or context of utterance, for its interpretation. This other element, which we call focus, is one of the main distinguishing features between ours and other annotation schemes.

1.2 The FOCUS Attribute

The FOCUS shows how the relative expression is interpreted, its values are: Deictic, Anaphoric, Age-reference, A-different-speaker.

A relative expression anchored to the utterance of the author of the text is assigned a deictic focus, directing us to a constant that is always the date of the document. This is generally the decision adopted in different annotation models. In this case, no additional elements are needed to interpret the temporal expression correctly.

For the three remaining values, the focus will be a text element indicated through the tag focus, and a link between the eTemp and its focus will be added. The most usual scenario for the anaphoric anchor of an eTemp is the anaphoric-type focus. Such is the case in the following example, where *En 1815 / In 1815* (also an eTemp) is the anaphoric focus of the relative eTemp *ese año / that year*.

> ***En 1815*** *Artigas logró que los porteños devolvieran Montevideo a los orientales, y* ***ese año*** *pudo gobernar todo el país. / In 1815 Artigas convinced the porteños to return Montevideo to the orientales, and that year he was able to rule the whole country.*

The anaphoric anchor of an eTemp is an Age-reference focus in expressions like a los 20 años / at 20 years of age, en su primera infancia / in his early chidhood, which direct us to a focus that may be the person's date of birth, if included in the text, or even the person's name. There are previous studies focused on the syntax and auto-

matic recognition of adverbial expressions that are Age-references according to our scheme [1]. In the following example, the eTemp *a los cinco años / at the age of five* has *el 27 de enero de 1756 / 27th Januray, 1756* as Age-reference type focus.

> *Mozart nació el 27 de enero de 1756. A los cinco años ya componía pequeñas piezas musicales. / Mozart was born on 27th January 1756. At the age of five he was already composing short musical pieces.*

When there is a relative temporal expression with a deictic anchor in reported speech, i.e., with a change of speaker, this expression takes on the value A-different-speaker for the FOCUS attribute. A focus tag is assigned to the expression that introduces the reported speech. In the following example, the eTemp *Hoy / Today* has the reporting verb *dijo / said* as its focus.

> *El pasado jueves el presidente dijo: "Hoy iniciamos una nueva etapa en la política cultural" / Last Thursday the President said: "Today we start a new stage in cultural policy."*

1.3 The GRANULARITY Attribute

The GRANULARITY attribute enables us to consign the temporal magnitude to which the expression refers. A wide range of values has been set for this attribute: second, minute, hour, fraction-day, day, fraction-week, week, fortnight, fraction-month, month, fraction-year, year, decade, century, millennium, historical period.

1.4 The VALUE Attribute: The Temporal Description Language LDT

Temporal expressions are described in abbreviated form through the attribute VALUE, expressly naming elements that are implicit or need to be deduced from the context. A temporal description language (LDT) has been defined [8] to describe the expressions within the VALUE attribute. A literature antecedent for LDT is TCNL, Time Calculus for Natural Language, from Han and Kolhase [9].

The name LDT is an abbreviation for Spanish *Lenguaje de Descripción Temporal*. LDT objects are intervals, interval sequences and points. They are arranged in a time line, on which there is an order. Points and intervals are in fact interchangeable with each other, e.g., the temporal expression *December 21, 1980* is seen as an interval in the example a) while in b) it is preferably seen as a point.

a) On December 21st, 1980 I visited some museums and spent the afternoon in the park.
b) A tax moratorium until 21st December 1980 was granted yesterday.

We distinguish in LDT two types of items: basic objects and complex objects. The basic objects are defined by extension and the complex objects are constructed from basic ones and other elements by means of some operations.

1.4.1 Basic Objects

The basic objects correspond to the usual calendar units as well as to names of special events and historical periods. They are noted by abbreviations. Some special basic

objects, necessary for the definition of complex objects, are also defined. In what follows we define different kinds of basic objects:

Calendar units, culturally recognized units.

- usual elements in our calendar system: day (di), month (me), year (yy), Century (sg), Millennium (mi), etc.
- parts of the day (morning, afternoon, etc.) and seasons.
- units of time system (hour, minute, second, etc.)
- names of festivals, Christmas, Easter, New Year, etc.
- culturally recognized names of historical periods (Middle Age, Industrial Revolution, BC (Before Christ), etc.)

Special items: generic intervals.

- u: universal interval
- x: generic interval

Special items: referents for temporal anchoring.

- fd: deictic focus (unique, the utterance moment)
- fa: anaphoric focus (variable: different events and temporal expressions in text)

1.4.2 Operations in LDT

The basic objects rarely appear isolated in a temporal expression. They usually appear with name (e.g., *September* is the month with basic object with name me9 in our notation), quantified (*the last days of September*) or in more complex constructions (*September 21, 1908*, a combination of several basic named units). Notice that an expression such as *September 21, 1908* can be seen as a successive application of restrictions on basic units:

(a) the basic unit day is restricted to a specific day, 21 → di21,

(b) the basic unit month is restricted to a specific month, 9 → me9,

(c) the basic unit year is restricted to a specific year, 1908 → aa1908,

(d) the term di21 is restricted to a specific month → di21-me9,

(e) di21-me9 expression is restricted to a specific year → di21-me9-aa1908.

The applied operations include on one side naming, defining, quantifying, restriction or selection into larger units, and, on the other side union, definition of regions and shift.

1.4.2.1 Naming

It is used to select a specific unit within a class of calendar units. The possible names are ordinal units:

- day, abbreviated ds if it is referred as a day of the week and dm if it is referred as a day of the month. It has two sets of names: ds - 1,2, ..., 7 (Monday is 1) and dm - (1,2, ..., 31)
- month, abbreviation me- (1,2, ..., 12)
- year, abbreviation aa - 1, 2, ... (optional AC)
- century, abbreviation sg - 1, 2, ... (optional AC)
- millennium abbreviation mi - 1,2, ... (optional AC)

We write the abbreviation of the unit and then the name (e.g., me10), followed by AC if applicable.

1.4.2.2 Selection

The selection is an operation with 3 arguments: selected temporal object, selection type, and unit or temporal object on which the selection is made. This operation is mainly used to express quantification on temporal units.

Notation : sel (o1, selectMode, o2)

We select a subset of objects o1 from o2 objects according to selectMode mode of selection.

Example: *the last days of December* → sel (di, last, me12)

The naming is a particular case of selection where the second object o2 does not intervene.

There is a wide range of selection modes:

- Ordinal
- Cardinal (special case of quantifier)
- Quantifier (one, all, some, many, few, several, most, late, early, middle, averaging, etc.). We also have constructions like *almost all*, etc.

The selected object (o1) may be the special object x (time, "temporal substance") previously defined.

*In **early April** the leaves begin to fall* → sel (x, early, me4)

It may also be necessary to use the universal interval u

*In **some years** we get good crops.* → sel (aa, some, u)

It may also be necessary to use the two special items together.

For a long time it was believed that the sun moved around the earth. → sel(x, mucho, u)

1.4.2.3 Duration Expressions

The basic units mentioned in 2.4.1 serve both to build expressions of location and duration. The duration expressions are basically composed of a calendar unit and a magnitude. We use the syntax Magnitude.TemporalObject

Examples:

*Today I have studied for **4 hours**.* → **4.hh**

*It took **a few days** to paint the house.* → **algunos.di**

There may be more than one unit in the same temporal expression, in this case we construct a compound expression whose denotation is the union of both time amounts, separating by hyphens the subexpressions in the different units.
*The ride to Pando takes **1 hour and 20 minutes**.* → **1.hh-20.mt**

1.4.2.4 Selection by Proximity (closer, mp)

We use the notation mp(Object, P, Dir) to select the temporal unit of type Object closest to the point P, in the direction specified in Dir (previous (neg), posterior (pos) or matching (ig)).

It is primarily used to make explicit the position regarding the textual focus on constructions with an implicit focus.

Examples:

*We will meet **in November.*** → mp (me11, fd, pos)

*I saw him **last Thursday.*** → mp (ds4, fd, neg)

In general terms, the context determines whether there is a reference to a textual (anaphoric) or to the deictic focus (usually the creation date of the document). Notice that if you use the anaphoric focus a temporal link with an explicit focus has to be signaled. The direction (previous, subsequent) is also recovered from the context. The case of coincidence (direction = ig) is used for expressions where there is a direct reference to a temporal expression.

We met many times in that year. → mp (aa, fa, ig)

The **mp** operation includes the access to a temporal coordinate, a deictic or anaphoric focus in the previous examples. In other words, we recovered the year, month or day or other unit of a time point. This is a kind of projection operation, in terms of some temporal unit.

*They met on **April 80** and in **December of that same year** they married.*

→ **me4-aa1980**

→ **me12-mp(aa,fa,ig)**

And there exists also an anaphoric link between *April 80* and *that same year* (not shown in the previous expressions).

1.4.2.5 Regions Definition

Region are time intervals. We can build them specifying a point, a length and a direction (*rd operator*), or by specifying both extremes (*rr operator*).

Regions by point and directions

rd (Point, Duration, Direction)

Definition of a temporal interval (region) from Point, Length and direction (neg, anterior; pos, later; ent, environment-centered in P)

In the last 40 years *we observed some climate changes.* → rd (fd, 40.aa, neg)

Region as a range
rr (Point 1, Point 2)
Build the range from Point 1 to Point 2

(i) *I lived with my brother between April 10 and April 20, 2010.*
→ rr(dm10-me4-aa2010, dm20-me4-aa2010)

(ii) I went to the movies several times between April 10 and April 30, 2010.
→ sel (time, number, rr (me4-aa2010-DM10, DM20-me4-aa2010))

(iii) I went to the movies several times between April 10 and April 30 I this year.
→ *sel (time, number, rr (me4-aa2010-DM10, DM20-me4-aa2010)) rr (DM10-me4-mp (aa, fd, ig), DM20-me4- mp (aa, fd, ig))*

To solve (iii) we must consider a further complication. (iii) it is similar to (ii), except that instead of referring to 2010, we use the term *this year*, referring to the year of the deictic focus.

(iv) I lived with my brother between April 10 and April 30.

This example is again more complex than the example (i). We refer to the closest period between April 10 and 30 that precedes the utterance. The expression is a range expression, both extremes (April 10 and April 30) are defined relative to deictic focus.

→ rr(dm10-mp(me4,fd,neg), dm20-mp(me4,fd,neg))

(v) I am going to the movies since two months ago.

In the example (v) we have a region defined by a start point and a direction. That starting point is defined by a shift operation applied to the deictic focus, we see the example in section 4.2.6

1.4.2.6 Shift

The expression desp(P, distance, direction) defines a point P' at a distance Distance preceding (direction = neg) or succeeding (sirection = pos) point P.

Examples:

*(i) We met t****hree years ago****.* → despl (fd, 3.AA, neg)
(ii) ***Tomorrow*** *we're going to see each other.* → despl (fd, 1.di, pos)
(iii) It is ***three years since*** *we are seeing each other.* → rr (despl(fd, 3.aa, neg), fd)

Notice that the verbal aspect affects the interpretation of the temporal expression. The same temporal expression (*3 years*) was interpreted as a point in time in (i) whereas in (iii) it should be interpreted as a region.

1.4.2.7 Union

To represent expressions like *April 4 and 5* it is desirable to have an operator able to form the aggregate or group containing 4 and 5. We have then defined the **union operator**: **un (O1, O2)**, where O1 and O2 are objects temporary.

The course is held on Tuesdays and Thursdays in the second semester.

→ sel (un (ds2, ds4), todo,sm2)

2 Temporal Indicators, *mTemp*

As mentioned in the Introduction, temporal indicators (*mTemps*) are elements of a relational nature or lexical temporality indicators. Even though an expression such as *hasta el año 1925 / until 1925* will be considered an *eTemp* and the element *hasta / until* is part of it, in an expression such as *hasta la elección / until the election*, the element *hasta / until* will be considered as an *mTemp* that links events, since it is not included in any temporal expression. But the adverb *hasta* / *until* is, without a doubt, a relevant element for the temporal analysis of texts, for example, for the temporal sequencing of events within the text. The vocabulary linked to temporality includes several types of temporal indicators which are classified in one of the following classes: relational, stages, ordinal numbering, duration, frequency and relative.

2.1 Relational and Relative mTemps

mTemps such as *antes de que / before , después / after , mientras / while* take on the value Relational. These determine a relation between two events which could be anteriority, posteriority, simultaneity, inclusion, etc. They could be applied to localized temporal intervals (*antes del jueves / before Thursday*) integrating eTemps, or to events, in which case they would be annotated as relational indicators.

In addition, relative mTemps provide a time reference with regard to the moment of utterance or any other moment expressly or implicitly mentioned in the text. They are analogous to relative eTemps and, as such, their focus can take on the values Deictic, Anaphoric or A-different-speaker. Therefore, in *El mes pasado el presidente dijo: "la **próxima** elección será un éxito" / Last month the President said: "the next election will be a success", próxima /next* is a relative temporal indicator of the type different speaker, with focus on the moment of utterance (*el mes pasado / last month*). Note that *"la semana próxima / next week"* is an eTemp, and in this case *próxima / next* is not analyzed as a temporal indicator.

2.2 Stage and Order mTemps

These mTemps focus on a stage within the development of the event (***Al principio** la guerra fue muy cruenta / At first, the war was a bloodshed*), or express lexically, generally in an indirect manner, the temporal sequencing of the events (*Lo entendí recién en la **segunda** clase / I was able to understand it in the second class*).

2.3 Duration and Frequency mTemps

This class is basically comprised by some verbs and adjectives, such as *durar*/*last* and *asiduas/frequent* in the following examples: *La entrevista **duró** toda la tarde / The interview lasted all afternoon, Se acostumbró a sus **asiduas** visitas / He grew used to his frequent visits.*

3 Annotation: Characteristics and Problems

In order to validate the scheme proposed, two previously trained annotators annotated a corpus of 2,300 words containing 33 eTemps and 35 mTemps.

Regarding eTemps determination, we found consistency between the annotators: annotator 1 (A1) marked 32 eTemps and annotator 2 (A2) marked 38. A2 twice marked as eTemps expressions that correspond to temporal indicators; in one case, he annotated an eTemp as focus; the three remaining expressions, which only A2 marked, do not correspond to eTemps nor mTemps.

In addition, when determining the values of the attributes we did not find significant differences between the annotators.

Annotation of temporal indicators was more problematic. Out of the 35 mTemps present in the corpus, A1 marked 30 correctly and A2 only 23. Moreover, we detected a significant number of false positives: A1 annotated 5 and A2 annotated 10. These values show that it is necessary to adjust the definition of temporal indicator.

In spite of the apparent complexity of the LDT language, the annotation of the value attribute proved simple and errorless. We looked at more than 50 cases, in a corpus of journalistic texts, and we didn't notice any error. Interestingly, almost all cases used the mp (closest) operator with a deictic focus.

4 Discussion, Future Work

We worked on a proposal for modeling temporal expression and other lexical elements that convey temporal meaning. The model is mainly compatible with TimeML, extending it with temporal indicators and new classes for different types of relative expressions and their related focus. A first task of corpus annotation showed encouraging results, suggesting the pertinence of our model.

Our plan is to proceed to the automatic recognition and interpretation of temporal expressions and indicators, as an intermediate task for text understanding. As a first step, we are experimenting with the recognition of the extent of temporal expression, using neural networks over a vector based representation of texts. For the interpretation, our plan is to extract the relevant information form the value attribute, that is, the expression in LDT language.

References

1. Galicia-Haro, S.N., Gelbukh, A.F.: Supervised Recognition of Age-Related Spanish Temporal Phrases. In: Proceedings of the 8th Mexican International Conference on Artificial Intelligence, MICAI, pp. 145–156 (2009)

2. Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G.: TimeML: Robust Specification of Event and Temporal Expressions in Text. In: Proceedings of the Fifth International Workshop on Computational Semantics. IWCS-5 (2003)
3. Pustejovsky, J., Verhagen, M., Saurí, R., Littman, J., Gaizauskas, R., Katz, G., Mani, I., Knippen, R., Setzer, A.: TimeBank 1.2. LDC catalog ref. LDC2006T08 (2006)
4. Pustejovsky, J., Verhagen, M.: SemEval-2010 Task 13: Evaluating Events, Time Expressions, and Temporal Relations (TempEval-2). In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. Association for Computational Linguistics. SEW-2009 (2009)
5. Saurí, R., Saquete, E., Pustejovsky, J.: Annotating Time Expressions in Spanish TimeML. Annotation Guidelines (Version TempEval-2010). Barcelona Media Technical Report 2010-02 (2010)
6. Saurí, R., Badia, T.: Spanish TimeBank 1.0. LDC catalog ref. LDC2012T12 (2012)
7. UzZaman, N., Llorens, H., Derczynski, L., Verhagen, M., Allen, J.F., Pustejovsky, J.: SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In: Proceedings of the 7th International Workshop on Semantic Evaluation, pp. 1–9, ACL SemEval (2013)
8. Wonsever, D., Malcuori, M., Etcheverry, M.: Esquema de anotación de expresiones y marcas temporales. Reporte técnico, serie: 0797–6410, PEDECIBA-Informática, https://www.fing.edu.uy/inco/pedeciba/bibliote/reptec/TR1115.pdf (2011)
9. Han, B., Kohlhase, M.: A Time Calculus for Natural Language. In: Proceedings of the 4th Workshop on Inference in Computational Semantics, Nancy, France (2003)

Sentiment Analysis for Reviews in Spanish: Algorithm for Negation Handling

Samara Gretel Villalba-Osornio, Juan Andrés Pérez-Celis,
Luis Villaseñor-Pineda, Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica Óptica y Electrónica,
Puebla, Mexico

{svillalba, celis, villasen, mmontesg}@inaoep.mx

Abstract. Sentiment Analysis (SA) is a challenging area where several problems still are open, one of them is treating the negation. Most of the research focusses on reviews written in English. Therefore, algorithms to deal with the negation in other languages are needed. This work proposes an algorithm to determine the scope of negative particles and to treat the affected words on reviews in Spanish. Our algorithm includes a novel weight scheme which combines the relative frequency of a word with the vocabulary of the class. The experiments show that handling the negation improves the accuracy in Spanish reviews in most cases. Likewise, the proposed weight scheme reaches a maximum accuracy of 67% before treating the negation. Combining our algorithm with the proposed weight scheme yields an accuracy greater than the one reported in the state of the art.

Keywords: Sentiment analysis, negation treatment, Spanish, reviews, polarity classification, negative particles scope.

1 Introduction

Nowadays, the social web provides efficient tools to share opinions regarding social events, politics, products, and services. However, this information is not hasty to process automatically due to the lack of structure and the subjectivity included. The scientific community has shown a growing interest in the matter for the challenges it represents. Likewise, the business world share some interest in the matter for the possible market prediction. In order to cope with the subjectivity the field of Opinion Minning (OM), also known as Sentiment Analysis (SA), emerged [15].

The OM field tries to classify the documents by the polarity of the opinion expressed [10, 15]. Currently, there is a fair amount of research ongoing on that deal with several aspects of the SA [3]. This work focusses on one aspect: the negation; determining its scope and performing a suitable treatment.

The negation handling has a wide importance as people express opinions with negated words. For example, “La película no me gustó” (I did not like the

movie) is a negative opinion with a negated positive word. For the other case, "La historia no es mala" (The story is not bad) expresses a positive opinion with a negated negative word.

Even though, most of the work regarding the negation treatment focuses on English opinions, the increased presence of documents in other languages creates the need for developing research in languages different than English. This paper deals with the problem of negation in Spanish documents. The main contribution is an algorithm that establishes the reach of negative particles for opinions written in Spanish. The goal is to improve the correct prediction of the polarity in opinion texts.

In addition to the algorithm, a novel weight scheme based on a supervised approach is proposed. The scheme takes into account the relative frequency weight of terms in a class' document and the vocabulary of the classes. The novel weight technique was tested in two corpus of different domains with successful results.

The rest of the paper is organized as follows: Section 2 gives an overview of the related work: the negation treatment. Section 3 describes the classification method and presents the proposed algorithm. Section 4 shows a brief analysis of the two corpus used and details the experiments performed. Section 5 presents the results and a thorough analysis of them, highlighting the findings. Section 6 gives concluding remarks and suggests future work.

2 Related Work

There is previous work on the negation handling. Mostly the works are for English texts [1, 3, 8, 14]. There is also research ongoing on other languages such as Chinese, French and Spanish [2]. The most relevant works are detailed bellow, starting with the works for English reviews.

The first research dealing with the negation treatment, to the best of our knowledge, was performed for the English language in 2001. A simple approach is taken by adding a marker to the words following some negative particles until the end of the sentence [6]. The same approach is taken in [11] but only considering the negative particle *not* and its contraction. The authors consider that the effect of the negation is negligible. In [1] the algorithm adds a flag which changes its state when *not* is found in a document. Then the words after *not* are marked, if another *not* is found or a punctuation mark, the marker stops tagging words.

In [8] the model used modifies the polarity of a word next to a negative particle, an intensifier or a mitigator. The polarity of a word gets inverted for the case of a negative particle. The work states the treating the negation has a significant effect on classifying opinions. This model can be seen as if it uses a window of size 1, .i.e, the effect of negative particles presents only to the word next to them.

The window size is modified in some works. In [16] a window of size 4 with effect to the right is proposed. The windows can have different size and can have

an effect to the left too. In [3] the author used a window of size 3 to the right and 1 to the left.

To stop the effect of a negative particle not only a size of a window has been used. Nexus or words of a certain category can stop the effect [4, 14]. More sophisticated information from dependency trees has been used to stop the effect of negative particles [5, 7, 14].

For the Spanish language, the number of works is limited. In [14] English opinions where translated into Spanish for their further classification with a method proposed for reviews in English [4]. A more recent work considers the syntactic structure of the text in matter [12].

Lastly, in [2] the authors utilize the syntactic structure with grounds on the information provided by dependency trees for each sentence. Moreover, negative particles are included, such as "tampoco" (neither), "nadie" (none), "jamás" (never), "ni" (nor) y "nada" (nothing). This work reports a relative improvement of 3.02% when handling the negation. However, there is no analysis of the results to define the reasons for the improvement.

The work herein uses the same corpus as [2]. The difference with our work is that no dependency trees where used and the approach taken in this work is supervised; i.e., the list of words from eSOL, iSOL, SentiWordNet where not used. We inspired in the techniques previously reported for the English language and propose a much simpler approach. To establish the reach of the negative particles windows, nexus and punctuation marks were used.

3 Methodology

In this paper a Negation Algorithm (NA) that determine the scope of negative particles is proposed. The NA is incorporated into a classification method with a supervised approach. The method is described in the following subsections.

3.1 Negation Particles

The first step in the negation analysis is the identification of the negative particles in Spanish. According to the Real Academia Española (RAE)[1] several negative particles exist. In this work the particles considered are: "no", "sin" (whitout), "ni"(nor), "nada"(nothing), "nunca"(never) and "tampoco"(neither).

3.2 Negation Scope

The second step is to design the algorithm that will determine the words in a sentence affected by the negative particles. A fair number of approaches have been reported for the English language. This work takes ideas of the previous research in English and combine them to get an appropriate algorithm for negation handling in Spanish.

Namely, the algorithm takes into account the following aspects:

[1] "Nueva gramática de la lengua española: Fonética y fonología. Espasa Libros. (2011)"

- A window of size N affecting a N number of words following a negative particle. In this work windows of 2, 4, 6 and 8 were tested.
- The negative particle affects words between itself and a punctuation mark.
- When a negative particle is found a flag activates and so the effect of it. If another negative particle is found within the reach of the first one, then the flag is turn off and so the effect. Consider the sentence "No había visto nada tan bueno" (I have not seen anything that good), the particle *no* will affect *había* and *visto*, the particle *nada* will turn off the effect so the remaining words will be left without the negation effect.
- The effect of a negative particle is ended when an adversative nexus is within the reach of the negation. For example in "No estuvo romántica pero me gustó" (It was not romantic but I like it), the effect will be stopped by the word *pero* (but). The adversative nexus used in this work are: "pero"(but), "aunque"(even though), "sino" (if not, but), "sin embargo"(however), "no obstante"(nevertheless), "al contrario"(on the contrary) and "a pesar de" (despite).

After the algorithm is executed, the words that are affected by a negative particle are replaced with the negated forms in the document. That modified document is classified. For example if the original document contains "La película no me aburrió, está divertida." (The movie did not bore me, it is fun.). The modified document will contain "la película no_me no_aburrió está buena" (the movie not_bore not_me it is fun).

The algorithm is shown on Fig. 1.

Algorithm 1 Determinig the Negation Scope

```
1: Read Document.
2: for Each word in the document do
3:   if The word is a negative particle then
4:     for i = 0 until i = window size do
5:       if The next word is not a negative particle, and is not a puntuation
         mark, and is not an adversative nexus then
6:         Add "no_"+ Word.
7:       else
8:         Break and continue with the next word in the document.
9:       end if
10:    end for
11:  end if
12: end for
```

Fig. 1. Algorithm for the scope of negative particles.

3.3 Words' Weight Computation

The third step is to compute the weight of each word or term for the negative and positive class. For this step three different weights were calculated.

In the first weight scheme the weight of a term is related with the number of documents where the term appears. The weight scheme is shown in eq. 1. It will be referred as relative frequency scheme.

$$W(t_i) = \frac{D_c\ t}{|D_c|}\ ; \tag{1}$$

where:

- $D_c\ t$ are the documents of class c that contain the term t.
- $|D_c|$ are the total number of documents in the class c.

The second weight is described in eq. 2. It will be referred as Frequency scheme.

$$W(t_i) = \frac{F_c\ t}{\sum^d F_c\ t'}\ ; \tag{2}$$

where:

- $F_c\ t$ is the frequency of term t in documents of the class c.
- $F_c\ t'$ are the frequencies of all terms in the class c.

Lastly, the third weight scheme, shown in eq. 3 takes a novel approach. The core is the same as eq. 1, the relative frequency weight. The modification lies in $\frac{1}{Voc_c}$, this term was added with the aim of making a compensation as people use more positive words to express themselves in English and in other languages [13], including Spanish.

$$W(t_i) = \frac{D_c t}{|D_c|} \times \frac{1}{Voc_c}\ ; \tag{3}$$

where:

- $\frac{D_c t}{|D_c|}$ is the relative Frequency weight.
- Voc_c is the vocabulary of class c.

3.4 Polarity Modification

Considering that the weights are computed and the negation scope is defined, the next step is modifying the polarity of the terms affected by a negative particle. Starting from the fact that there are two lists, one with positive weights and one with negative weights for each word, the modification is as follows. For each word or term added to the vocabulary the negative form is also added, i.e., a word comprised by *"no_" + word* is added. This *no_* prefix acts as a mirror for the weights. The weights of a term in its negated form are the opposite weights of that same term in its normal state. For example: if the word *good* has a positive weight of 0.7 and a negative weight of 0.3, the word *no_good* will have a positive weight of 0.3 and a negative weight of 0.7.

With the weights for negative and positive words computed, the last step is the classification of the documents.

3.5 Classification Method

The decision whether a document D belongs to the positive class or negative class is taken based on the sum of the weights for all the words in D. It is noteworthy to recall that the document to be classified is the modified document, which is the output after executing the proposed algorithm for the scope of negative particles. The decision is based on comparing the result of eq. 4 for each of the two classes. The document will be classified as positive if the sum of the positive weights for all words in the document under test is greater than the sum of the negative weights for all the words in the document, and will be negative otherwise. In case the result of eq. 4 is equal for both classes, the document will be classified as negative.

$$Weigth(D_x) = \sum_{i=1}^{n} W(t_i) + \sum_{j=1}^{m} W(\overline{t_j}) \; ; \tag{4}$$

where:

- n is the total of non-negated terms in the document x.
- m is the total of negated terms in the document x.

The summation on the left expresses the sum of the normal weights for all non-negated terms. The summation on the right adds the mirrored weights for all negated terms.

4 Experimental Framework

Tests where conducted on two corpus in Spanish. The first corpus is of movie reviews [5] and the second corpus is the SFU [9]. The description of both corpus is detailed on the next subsections.

4.1 Corpus Details

The Corpus of Movie Reviews (CMR) is a Spanish corpus containing 3878 movie reviews obtained from the website MuchoCine [2]. The reviews are rated from 1 to 5, where 1 is a bad movie and 5 an excellent movie. For this work the reviews rated with 3 are considered neutral and are disregarded from the work. The opinions rated with 1 or 2 are negative opinions. The opinions with 4 or 5 are positive opinions.

The corpus SFU [3] is comprised by 400 reviews regarding eight topics (books, cars, computers, electro domestics, hotels, movies, music and cellphones). Each topic has 50 reviews, 25 negative and 25 positives. For the two corpus, the opinions were written by users of the web, there are no modifications on the writing.The opinions preserve spelling mistakes, grammar mistakes and colloquial expressions. The Table 1 shows details of the documents in the corpus.

[2] "http://www.lsi.us.es/ fermin/corpusCine.zip."
[3] "https://www.sfu.ca/ mtaboada/research/SFU_Review_Corpus.html."

Table 1. Details of CMR and SFU corpus.

Classes in the CMR.				
Class.	**Documents.**	**Vocabulary.**	**#of Words.**	**Ave. # of Words.**
Positive	1352	42245	736172	545
Negative	1274	37542	560639	440
Classes in the SFU corpus.				
Positive	200	14578	120227	601
Negative	200	10714	77159	386

The number of negative particles in each class was counted for this corpus. The results shown in Table 2 suggest that treating the negative particles will have a effect in the classification as the number of appearances is high. It is interesting to point out that the most used particle *no* is used more times in the positive class than in the negative class. This fact may be a consequence of having longer reviews and greater number of words in the positive class.

Table 2. Appearance of negative particles in both Corpus.

	CRM		**SFU Corpus.**	
	Positive	**Negative**	**Positive**	**Negative**
NO	7932	7606	4797	1528
SIN	2217	1600	256	169
NI	876	1233	161	199
NADA	719	866	183	211
NUNCA	387	284	81	39
TAMPOCO	201	315	33	45

4.2 Experiments Considerations

Several experiments were conducted with both corpus, however the main corpus is CMR due to the greater extension in the number of reviews and to the fact that it is only focused on one domain. For the CMR, the classes were evened by disregarding the last reviews. Both classes remain with 1270 reviews. The experiments on the SFU corpus were conducted mainly to test the novel weight scheme proposed in this paper in various corpus.

The experiments detailed in this paper were performed in Python using the Natural Language ToolKit (NLTK). For the tests in both corpus a preprocessing was performed. This step began by changing characters containing accents of tilde. Characters such as á, é, were changed to their forms without accents. The special character $\tilde{n}$ was changed by n. Symbols were changed for a space. Punctuation marks were substituted by special markers for the sake of the algorithm. These markers have no influence in the classification step.

During the experiments stop words have a weight of zero in both classes. This rule allows stop words to be counted for the window size of negative particles but they have no effect in the total weight. The list of stop words used can be found in the RANK NL website [4].

Finally, all the tests use cross validation. The results report the macro-average and the standard deviation. For the CMR ten folds were used. For the SFU corpus five folds were used. Folds have even classes. For each fold first 80% of the reviews was used to compute the weights and the remaining 20% was classified. Folds were chosen to ensure that all reviews were classified; once for the SFU corpus and twice for the CMR.

4.3 Experiments

The first experiment consists in classifying the documents in the CMR with the three schemes mentioned in the paper. The experiment was performed with the negation algorithm disabled and enabled. The results of Table 3 show that the greatest accuracy is obtained with the proposed weight scheme and the negation algorithm enabled. The standard deviation is shown bellow each macro-average. For the three weight scheme enabling the negation algorithm modifies the accuracy. In the case of the Relative Frequency Scheme (RFS) the accuracy drops. In the Proposed Weight Scheme (PWS) and in the Frequency Weight Scheme (FWS) the accuracy shows an improvement.

Table 3. Classification of CMR with the different weight schemes.

	Proposed Weight Scheme							
	Without Negation				**With Negation**			
	Prec.	**Recall.**	**F1.**	**Acc.**	**Prec.**	**Recall.**	**F1.**	**Acc.**
Positive	0.9177 (±0.0299)	0.3774 (±0.1111)	0.5276 (±0.1116)	0.6727 (±0.0549)	0.7947 (±0.0372)	0.5872 (±0.0645)	0.7096 (±0.0280)	**0.7373** (±0.0145)
Negative	0.6114 (±0.0442)	0.9680 (±0.0083)	0.7486 (±0.0323)		0.7029 (±0.0279)	0.8288 (±0.0546)	0.7589 (±0.0151)	
	Frequency Weight Scheme							
Positive	0.9884 (±0.0880)	0.0247 (±0.0094)	0.0481 (±0.0179)	0.5116 (±0.0048)	0.8780 (±0.0433)	0.3095 (±0.0508)	0.4893 (±0.0580)	**0.6470** (±0.0289)
Negative	0.5058 (±0.0024)	0.9984 (±0.0020)	0.6715 (±0.0023)		0.5916 (±0.0202)	0.9535 (±0.2878)	0.7300 (±0.0181)	
	Relative Frequency Scheme							
Positive	0.5249 (±0.0149)	0.9952 (±0.0061)	0.6871 (±0.0118)	**0.5465** (±0.0265)	0.5109 (±0.0096)	0.9083 (±0.0020)	0.6760 (±0.0081)	0.5210 (±0.0182)
Negative	0.9652 (±0.0358)	0.0977 (±0.0571)	0.1726 (±0.0927)		0.9928 (±0.0225)	0.0429 (±0.0381)	0.0800 (±0.0678)	

The next experiment consists on changing the size of the window. Sizes of 2, 4, 6 and 8 were chosen. The results are shown in Table 4. The best accuracy is reached with a window size of 4. Likewise, Table 4 reports the results of

[4] http://www.ranks.nl/stopwords/spanish.

another experiment conducted with the goal of identifying the impact of the adversative nexus as disablers of the negation effect. The necessary changes were performed in the negation algorithm to ignore the effect of adversative nexus. The experiment is based in the PWS and was conducted with 2, 4, 6 and 8 as the size of the window.

Table 4. Windows size variations.

		With Adversative Nexus.				**Without Adversative Nexus.**			
Size	**Class**	**Prec.**	**Recall.**	**F1.**	**Acc.**	**Precc.**	**Recall.**	**F1.**	**Acc.**
2	**Pos**	0.8008 (±0.0323)	0.5775 (±0.0698)	0.7054 (±0.0341)	0.7369 (±0.0174)	0.8625 (±0.0363)	0.4740 (±0.0948)	0.6439 (±0.1255)	0.7174 (±0.0324)
	Neg	0.6994 (±0.0320)	0.8385 (±0.2568)	0.7610 (±0.0120)		0.6593 (±0.0383)	0.9133 (±0.2777)	0.7642 (±0.0165)	
4	**Pos**	0.7945 (±0.0372)	0.5872 (±0.0645)	0.7096 (±0.0280)	**0.7373** (±0.0145)	0.8516 (±0.0271)	0.4880 (±0.0922)	0.6535 (±0.0644)	**0.7203** (±0.0339)
	Neg	0.7029 (±0.0279)	0.8288 (±0.0546)	0.7589 (±0.0151)		0.6640 (±0.0404)	0.9042 (±0.2742)	0.7643 (±0.0186)	
6	**Pos**	0.7862 (±0.0290)	0.5862 (±0.0602)	0.7063 (±0.0297)	0.7334 (±0.0169)	0.8456 (±0.0261)	0.4848 (±0.0863)	0.6498 (±0.0614)	0.7172 (±0.0325)
	Neg	0.7002 (±0.0276)	0.8220 (±0.2514)	0.7549 (±0.0144)		0.6614 (±0.0374)	0.9011 (±0.2730)	0.7817 (±0.0179)	
8	**Pos**	0.7821 (±0.0279	0.5882 (±0.0609)	0.7060 (±0.0301)	0.7321 (±0.0171)	0.8417 (±0.0308	0.4787 (±0.0802)	0.6437 (±0.0566)	0.7126 (±0.0287)
	Neg	0.7004 (±0.0286)	0.8172 (±0.2499)	0.7530 (±0.0140)		0.6572 (±0.0330)	0.8987 (±0.2727)	0.7581 (±0.0159)	

The following test was performed over the SFU corpus. The test consist in using the three weight schemes (PWS, FWS and FRS) to classify the reviews. The window size is 4 and the adversative nexus are taken into account. The test was conducted without the negation algorithm and with the algorithm enabled. The results are shown in Table 5.

5 Analysis of Results and Discussion

The PWS got the greatest accuracy for the CMR and a better accuracy than BWS in the SFU corpus. The fact suggest that considering the vocabulary length for the weight schemes can improve the accuracy when classifying opinions. As mentioned before the PWS was inspired by the work in [13] which states that the vocabulary of a positive class is greater then the negative one. In an attempt to verify this statement, an analysis of the vocabulary for each of the ten folds was performed. The results in Table 6 show that the vocabulary for each fold of the the positive class is greater than the negative one.

The results of the PWS with a window size of 4 applying cross fold validation, as in all the experiments, were compared to those in [2]. The same corpus was used. The reviews with three stars were also disregarded in [2]. The approach taken by the authors is unsupervised. The Table 7 shows the comparisson. Our

Table 5. Results with the different weight schemes in the SFU corpus.

	Proposed Weight Scheme							
	Without Negation				With Negation			
	Prec.	**Recall.**	**F1.**	**Acc.**	**Prec.**	**Recall.**	**F1.**	**Acc.**
Positive	0.7333 (±0.4346)	0.0400 (±0.0285)	0.0846 (±0.0385)	0.5162 (±0.0162)	0.7850 (±0.2233)	0.0916 (±0.0418)	0.1907 (±0.0674)	**0.5375** (±0.0318)
Negative	0.5077 (±0.0095)	0.9900 (±0.0136)	0.6712 (±0.0108)		0.5202 (±0.0174)	0.9650 (±0.3954)	0.6759 (±0.0092)	
	Frequency Scheme Weight							
Positive	0.8000 (±0.2738)	0.0297 (±0.0113)	0.0571 (±0.0213)	0.5087 (±0.0121)	0.6520 (±0.1307)	0.1541 (±0.0928)	0.2783 (±0.0990)	**0.5400** (±0.0298)
Negative	0.5038 (±0.0072)	0.9901 (±0.0135)	0.6678 (±0.0092)		0.5242 (±0.0196)	0.8950 (±0.3694)	0.6601 (±0.0199)	
	Relative Frequency Weight							
Positive	0.5395 (±0.0367)	1 (±0)	0.7003 (±0.0304)	**0.5700** (±0.0603)	0.5240 (±0.0187)	0.8333 (±0)	0.6875 (±0.0160)	0.5450 (±0.0387)
Negative	1 (±0)	0.1400 (±0.1206)	0.2307 (±0.1754)		1 (±0)	0.0900 (±0.0707)	0.1595 (±0.1117)	

Table 6. Average length of the vocabularies.

Vocabularies		
Positive	**Negative**	**Total**
36269 (±276)	33216 (±103)	49379 (±222)

approach is almost 4% more accurate without the Negation Algorithm (NA) and almost 9% more accurate with the NA enabled.

Table 7. Comparisson of PWS with [2].

Jiménez, Martínez, Martín & Molina.							
Without Negation				With Negation.			
Prec.	**Recall.**	**F1.**	**Acc.**	**Prec.**	**Recall.**	**F1.**	**Acc.**
0.6365	0.6276	0.6320	0.6312	0.6519	0.6430	0.6474	0.6475

This Work							
Prec.	**Recall.**	**F1.**	**Acc.**	**Prec.**	**Recall.**	**F1.**	**Acc.**
0.7645	0.6727	0.7156	0.6727	0.7488	0.7080	0.7278	0.7373

The difference of the accuracy between the results of the experiments where the adversative nexus were considered and then disregarded for several size of windows is rather small 0.0195. To the end of understanding the reason of this effect, the Table 8 counts the condition which made the negative particles lose their effect.

The number of times the negation is stopped by punctuation marks increases as the windows size do so. The trend for number of times the negation is stopped

by the window size decreases as the size increased. It is an interesting finding that the number of times the negation effect is stopped by an adversative nexus is only 146 times for a window size of 8.

Table 8. Determining the negation scope.

	Positive Class.				**Negative Class.**			
Window Size	**Size 2**	**Size 4**	**Size 6**	**Size 8**	**Size 2**	**Size 4**	**Size 6**	**Size 8**
Punctuation Mark	1511	3740	5592	6795	1392	3494	5132	6242
Window	9503	6505	4320	2901	8866	5834	3871	2587
Nexus	30	83	120	146	25	71	102	122
Negative Particle	481	839	987	1082	649	1091	1238	1320
NO	31	83	120	146	34	92	158	210
SIN	6	32	45	58	8	28	35	46
NI	182	329	383	409	264	448	501	520
NADA	216	335	345	350	301	457	474	474
NUNCA	23	35	36	39	22	39	41	41
TAMPOCO	23	26	27	27	20	27	29	29

Another remarkable finding is that after analyzing the content of the reviews, the impact of considering double negative particles to end the effect of negation is not reflecting the actual case of the nested negative particles. Plenty sentences where nested negative particles present seek to produce an effect of intensification. This is commonly found in negative reviews. For the following examples the structure of the sentences is shown and an English translation that may not be gramatically correct as the point is to show the effect of Negative Particles (NP) in the sentences. In the negative class we found expressions such as:

- "No tiene ni subtramas ni similares ..." (NP + Verb + NP + Noun + NP + Adjective).
- "Ni engancha el guión ni interesa la historia ..." (NP + Verb + Article + Noun + NP + Verb + Article + Noun).
- "La trama no me parece ni original ni demasiado interesante ..." (Article + Noun + NP + Pronoun + Verb + NP + Adjective + NP + Adverb + Adjective).

In the possitive class nested negative particles are to be found in coloquial expressions or neutral phrases such as:

- "No tiene ni ton ni son ..." (NP + Verb + NP + Noun + NP + Noun). The meaning of this expression is that it lacks of interesting moments.
- "Ni más ni menos ..." (NP + Adjective + NP + Adjective).
- "No es ni mala ni buena ..." (NP + Verb + NP + Adjective + NP + Adjective).

An analysis of the output of the classification was conducted. The cases where the Negation Algorithm (NA) was enable and disable are compared. Two interesting situations were noticed:

1. All the negative reviews classified as positive without the NA remain positive after enabling the NA.
2. All the positive reviews classified correctly without the NA remain positive after enabling the NA.

For the first case, the analysis shows that the length of those reviews (of point 1) are above the average of positive and negative reviews. Likewise, negative particles seldom appear on those reviews making the negation handling effect negligible. Review #8 is a clear example. It has 985 words and says:

"... y es que cuando los paisajes te parecen más bonitos e interesantes que la historia de amor que se está contanto, es algo que no funciona, falla la pasión, falla la emoción, falla lo conmovedor, ...". (... and when the landscapes seem prettier and more interesting to you than the love story that is being told, there is something not working right, the passion fails, the emotion fails, the touching fails, ...).

It is clear that this is a negative review, however, the word "fails" is not used by the NA as a negative particle. Using verbs for negating the meaning of words was not considered in the NA. Moreover, combining those verbs with strongly positive weighted words makes the NA to err in the classification.

The second case have the same characteristics. The length of those reviews are above the average of positive and negative reviews. Likewise, the reviews lack of negative particles making the effect of the NA negligible.

The case where the reviews were misclassified before the NA and after the NA was also presented. This reviews have a fair amount of negative particles too close to each other, breaking the effect of the negation and causing a small change in the total weight summation. The difference in the total sum with the NA and without the NA is not enough to correctly classify the review in matter.

The reviews that where misclassified before the NA but where correctly classified after applying the NA have a particular characteristic. Representative words of a certain class are negated, i.e., words with a weight significantly greater in a certain class than in the other are negated. After applying the NA the weights of those words are mirrored yielding a correct classification. The following fragment of a positive review serves as an example:

– "... no es que sea mala ..." (... it is not that it is bad ...).

In the classification state the only word to consider is *mala* (bad) as the other words are stop words. The word *mala* has a weight of 0.0008231 and 0.0054293 respectively. When classifying without the NA the review will be placed on the negative class. Nonetheless, after applying the NA the weights are mirrored and the opinion is classified correctly as a positive review.

6 Conclusions

The algorithm proposed for handling the negation has a direct impact in the classification of Spanish reviews. In the Proposed Weight Scheme (PWS) the

effect increases the accuracy from 0.6727 to 0.7373 and from 0.5162 to 0.5375 for the Corpus of Movie Reviews (CMR) and the SFU corpus respectively. The Frequency Weight Scheme (FWS) improves from 0.5116 to 0.6470 and from 0.5087 to 0.54 for the CMR and the SFU corpus respectively. The relative frequency Weight Scheme have a counterproductive effect the accuracy drops from 0.5465 to 0.5210 and from 0.57 to 0.5450 for the CMR and the SFU corpus respectively. The greatest accuracy is achieved with the PWS and applying the Negation Algorithm (NA).

The tests performed for different size of windows suggest that the best performance comes when using a size of 4, at least for the CMR. The result agrees with the ones reported for English reviews. More tests should be conducted to verify if the behavior is the same in other domains of reviews.

The use of nexus in the NA increases the classification accuracy in 1.9% for the best case: with a window size of 2. The analysis showed in 8 suggest that the increase is not greater because only a few times the effect of negative particles is stopped by adversative nexus. 90% of the time the negation loses its effect due to the window size or the punctuation marks.

The PWS has a satisfying performance for classifying Spanish reviews. Giving importance to the vocabulary of the classes translate into a maximum improvement of 21% over the BWS, which only differs in the factor $\frac{1}{Voc_c}$. The PWS outperforms in the CMR the FWS by 12% without handling the negation an by 9% when the NA is used.

For the SFU corpus, the PWS has a similar behavior as the FWS. The PWS is 1% better when the NA is off and 0.25% worse then the NA is on. However there is still an improvement in both weights when applying the NA if compared to the result when the NA is off. The highest accuracy is achieved by the BWS. Compared to our PWS the BWS have a 6% higher accuracy without the NA and 0.8% with the NA enabled. The results of the PWS may have dropped in the SFU corpus due to the fact that the corpus is small (400 reviews) and have 8 different categories. Further tests will be performed to identify the exact reason. Our guess is that the vocabulary for each category varies dramatically in number and in the words used, making our PWS not able to compensate the greater use of vocabulary in positive reviews competently.

The supervised approach has a better performance for classifying Spanish reviews. Even without using our NA, with the PWS the accuracy achieved is 4% higher than the accuracy reported in [2]. When enabling the NA our approach outperforms by 9% the same work.

In the future work the algorithm presented for negation handling will be applied in the training set to obtain a vocabulary that includes words with the prefix "no_" indicating that was affected by a negation. The words with the mark will be treated in the same way of the rest of the features, calculating their probabilities according to their frequencies in the documents.

References

1. Narayanan, V., Arora, I., Bhatia, A.: Fast and accurate sentiment classification using an enhanced Naive Bayes model. In: Intelligent Data Engineering and Automated Learning–IDEAL 2013, pp. 194–201, Springer, Heidelberg (2013)
2. Jiménez Zafra, S. M., Martínez Cámara, E., Martín Valdivia, M. T., Molina González, M. D.: Tratamiento de la Negación en el Análisis de Opiniones en español (2015)
3. Bakliwal, A., Arora, P., Patil, A., Varma, V.: Towards enhanced opinion classification using NLP techniques. In: Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology, pp. 101–107, IJCNLP (2011)
4. Brooke, J., Tofiloski, M., Taboada, M.: Cross-Linguistic Sentiment Analysis: From English to Spanish. In: RANLP, pp. 50–54 (2009)
5. Cruz, F. L., Troyano, J. A., Enriquez, F., Ortega, J.: Clasificación de documentos basada en la opinión: experimentos con un corpus de crıticas de cine en espanol. Procesamiento de Lenguaje Natural, vol. 41. (2008)
6. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In: Proceedings of the Asia Pacific finance association annual conference, Vol. 35, p. 43 (2001)
7. Jia, L., Yu, C., Meng, W.: The effect of negation on sentiment analysis and retrieval effectiveness. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1827–1830, ACM (2009)
8. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational intelligence, vol. 22, pp. 110–125 (2006)
9. Konstantinova, N., de Sousa, S. C., Díaz, N. P. C., López, M. J. M., Taboada, M., Mitkov, R.: A review corpus annotated for negation, speculation and their scope. In: LREC, pp. 3190–3195 (2012)
10. Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., Perea-Ortega, J. M.: Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, vol. 18, pp. 7250–7257 (2013)
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86, Association for Computational Linguistics (2002)
12. Vilares, D., Alonso, M. A., Gómez-Rodríguez, C.: Clasificación de polaridad en textos con opiniones en espanol mediante análisis sintáctico de dependencias. Procesamiento del lenguaje natural, vol. 50, pp. 13–20 (2013)
13. Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., Danforth, C. M.: Human language reveals a universal positivity bias. In: Proceedings of the National Academy of Sciences, vol. 8, pp. 2389–2394 (2015)
14. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Computational linguistics, vol. 37, pp. 267–307 (2011)
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and trends in information retrieval, vol. 2, pp. 1–135 (2008)
16. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, pp. 347–354 (2005)

A Concise Representation for Detailed Sentiment Analysis

Víctor M. Morales[1], Saúl León Silverio[1], Maya Carrillo[1], Aurelio López-López[2], Luis Enrique Colmenares-Guillen[1]

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

[2] INAOE, Computational Science Department, Tonantzintla, Puebla, Mexico

{victor.morales,cmaya,lecolme}@cs.buap.mx
saul.ls@live.com, allopez@inaoep.mx

Abstract. This paper presents initial results in sentiment analysis classification, as an attempt to go beyond categorizing texts only by 'positive' or 'negative' orientation, using fine-grained features for this purpose. We present a method for sentiment classification based on a concise representation built from analyzing appraisal groups such as "very good" or "not terrible". An appraisal group is represented as a set of attribute values anteceding an appraisal word (adjective). An appraisal lexicon is used to identify adjectives guiding the analysis. We performed experiments classifying movie reviews in Spanish using features based upon attitude taxonomy information, and report improvements on precision with eight dimensional vectors and a support vector machine algorithm.

Keywords: Consice representation, appraisal group, sentiment analysis.

1 Introduction

Many traditional data mining tasks in natural language processing focus on extracting data from documents and mining it according to topic. In recent years, the natural language processing community has recognized the value of analyzing opinions and emotions expressed in free text. Sentiment analysis is the task of having computers automatically extract and classify the opinions in a text. Sentiment analysis has become a growing field for commercial applications, with at least a dozen of companies offering products and services for sentiment analysis, with very different sets of goals and capabilities.

Some applications of sentiment analysis focused on classifying movie or product reviews as positive or negative, or identifying positive and negative sentences, but more complicated applications involve opinion mining in ways that require a more detailed analysis of the sentiment expressed in texts. One such application

is the use of opinion mining to determine areas of a product that need to be improved by summarizing product reviews to identify what parts of the product are generally considered good or bad by users. For instance, if the product is considered of good quality but its presentation is not very pleasant. Another application requiring a more detailed analysis of sentiment is to understand where political writers fall on a political spectrum, something that can only be done by looking at support or opposition to specific issues. There are others applications [1], which allow politicians to have a better understanding of how their constituents view different issues, or predicting stock prices based on opinions that people have about the companies.

These applications can be tackled with a better understanding of the expressed opinions. Sentiment analysis researchers are currently working on creating the techniques to handle these more complicated problems, defining the structure of opinions and the techniques to extract such structure. For instance, the typology described by Appraisal Theory is reviewed in [2], presenting a methodology for annotating appraisal expressions, and its use to annotate a corpus of book reviews. The authors discuss an inter-annotator agreement study, and consider instances of common disagreement that indicate areas in which appraisal can be refined.

EmotiNet is presented in [3], a knowledge base for representing and storing affective reaction to real-life contexts. The authors detail the methodology employed in designing, populating, and evaluating such knowledge base. The basic model was further extended from additional resources, such as VerbOcean, ConceptNet, and SentiWordNet, with the aim of generalizing the knowledge, and later evaluated.

Appraisal theory was applied in news articles, specifically in the domain of stock markets [4]. They implemented an approach of Chinese new words detection by using N-gram model and applied the result for Chinese word segmentation and sentiment classification, employing Naïve Bayes, K-nearest Neighbor and Support Vector Machine algorithms. The best accuracy result reaches 82.9% in their experiments.

A different kind of domain specific feature-based heuristic for aspect-level sentiment analysis of movie reviews is reported in [5]. In particular, they developed an aspect oriented scheme that analyses the reviews of a movie and assign it a sentiment label on each aspect. The scores on the aspects from multiple reviews are then aggregated and a net sentiment profile of the movie is generated on all parameters. The scheme used is based on SentiWordNet, with two different linguistic feature selections comprising of adjectives, adverbs and verbs and n-gram feature extraction. Such scheme is also employed to compute the document-level sentiment for each movie reviewed and compared the results with those obtained with an API. The sentiment profile of a movie is also compared with the document-level sentiment result. They claim that the scheme produces a more accurate and focused sentiment profile than a simple document-level sentiment analysis.

Text classification with fine-grained attitude labels is reported in [6]. The

system developed relies on the compositionality principle and a novel approach based on rules elaborated for semantically distinct verb classes. The evaluation of the method on 1000 sentences, describing personal experiences, showed promising results in terms of accuracy for 14, 7 and 3 labels.

In the direction of these effort to have fine-grained sentiment analysis, and specifically taking advantage of Appraisal Theory, the goal of our research is to analyze opinions at a fine grain level that will allow an improved understanding of the opinion expressed in the text.

We developed a method for sentiment classification based on a concise representation built from analyzing appraisal groups such as "very good" or "not terrible", where an appraisal group is represented as a set of attribute values anteceding an appraisal word. An appraisal lexicon is used to identify adjectives guiding the analysis. We performed experiments classifying movie reviews in Spanish using features based upon attitude taxonomy features.

This paper is organized as follows: Section 2 provides a brief overview of related works, Section 3 describes our proposed approach in detail, Section 4 reports experiments and results, and finally section 5 includes conclusions and further work.

2 Related Work

Appraisal Theory is a particular approach for exploring, describing and explaining the way language is used to evaluate, to adopt stances, to construct textual personas and to manage interpersonal positioning and relationships. Thus, this theory explores how speakers and writers transmit judgments on people and object, generally. The Appraisal Theory has emerged over a period of almost 20 years as a result of work conducted by a group of researchers lead by Prof. James Martin of the University of Sydney ([7], [8], [9]). The taxonomy of Appraisal Theory has four systems: affect, graduation, orientation and polarity, and three subsystems: affect, judgment and appreciation, as follows:

- Attitude provides the type of appraisal being expressed as affect, appreciation, or judgment.
- Affect is concerned with emotional response and disposition and is typically realized through mental processes of reaction ("This pleases me", "I hate chocolate", etc).
- Judgment encompasses meanings which serve to evaluate human behavior positively and negatively by reference to a set of institutionalized norms.
- Appreciation system by which evaluations are made of products and processes. It encompasses values which fall under the general heading of aesthetics, as well as a non-aesthetic category of 'social valuation' which includes meanings such as significant and harmful.
- Graduation describes the intensity of appraisal in terms of two independent dimensions of force (or 'intensity') and focus ('prototypicality')
- Orientation is whether the appraisal is positive or negative (often simply termed 'sentiment').

– Polarity of an appraisal marked if it is scoped in a polarity marker (such as 'not'), or unmarked otherwise.

The research to map this theory to automatic sentiment analysis is recent and has been made for English language. In this context, we have the work of [10] where they apply the appraisal linguistic theory and find that it could be helpful in distinguishing different types of subjective texts (e.g., movie reviews from consumer product reviews).

In [11], two representations of text are compared, within the same experimental setting for sentiment orientation analysis. The two representations are bag-of-words (BoW) and a nine dimensional vector (9Dim), i.e. a compact representation, as we are experimenting. The former represents text with a high dimensional feature vector as we also tested. But, the short representation encodes grammatical knowledge of clauses in sentences, which is lexicon-independent. The aim of their paper was to demonstrate how text sentiment orientation classifiers' performance is improved when the length of the sentence comprising a training vector is varied. Their results indicated that the classification accuracy benefits from increasing the text length, and the results also indicated that the compact method can provide comparable results to BoW under the same sentiment classification algorithm. They also employed the support vector machines (SVM) algorithm.

3 Proposed Approach

We now describe the method for sentiment classification that basically consist on getting a concise representation built from analyzing appraisal groups, using an appraisal lexicon to identify adjectives revealing opinions. We test two different representations that are described. The lexical resource needed for the analysis is described as well as the way the appraisal groups are handled.

3.1 Method

The proposed method counts the number of positive, negative, judgment, appreciation and affects words present in an opinion. Each opinion is represented as a five entry vector, each one corresponding to the entries that a word has in the appraisal lexicon.

To evaluate an opinion, each word is looked up in the appraisal lexicon, if the word is found, the opinion vector is increased according to the values that the word has in the lexicon. However, the presence of appraisal modifiers plays a key role in this process, because depending on the type of modifier; values may increase, decrease or even be inverted, so as a result the values of the opinion vector are affected.

Two approaches (models) are used to evaluate the opinions under analysis, which are described below:

Model-1 Weighted Terms: In this approach, each opinion is represented as a vector of five entries, each one corresponds to the registered values in the attitude lexicon (i.e. positive, negative, judgment, appreciation and affect). The opinion is reviewed word by word, when a term is found in the lexicon; all its five values are stored in a temporary vector. Therefore, a window of three words before the term is analyzed, if appraisal modifiers are detected, then following steps are followed:

- If the appraisal modifier is an inversion, temporary positive and negative values are exchanged.
- If the appraisal modifier is considered of high impact, all the values of the temporary vector are duplicated.
- If the appraisal modifier is of low impact, all the values of the temporal vector are halved.

Finally, the temporary vector is added to the opinion vector.

Model-2 Weighted Terms by Polarity: In the previous model, vector entries for affect, judgment and appreciation, do not help a lot to determine the opinion polarity. With the aim that these aspects contribute to clarify the opinion polarity, in the second model, each of these attributes was represented by two vector entries. Then each opinion is represented by a vector of eight entries (i.e. positive, negative, positive judgment, negative judgment, positive appreciation, negative appreciation, positive affect and negative affect).

To determine the polarity of each lexicon term, as there are words that are valued both positively and negatively, the positive and negative values are subtracted, if the remainder is greater than zero, the term is considered positive; if subtraction is less than zero, the term is considered negative, and if the remainder is zero we say that the word is neutral.

As in the previous model, this approach processes the opinion word by word, when a term is found in the lexicon, the actual polarity of the term is calculated, i.e. the negative value is subtracted from the positive value and then, in accordance to the criteria shown in Table 1, the value of the eight temporary vector entries are established.

Afterwards, a window of three preceding words, as in the previous approach, is considered, if attitude modifiers are found, the following steps are done:

- If the attitude modifier is an inversion, temporary positive and negative values are exchanged, i.e. positive for negative, positive judgment for negative judgment, and so on.
- If the attitude modifier is of high impact, all the values of the temporal entries are duplicated.
- If the attitude modifier is considered of low impact, all the values of the temporary entries are halved. Finally, temporary entries are added to the final opinion vector.

Table 1. Vector entries.

Vector Entry	***Description***
Positive	Correspond to the positive term value in the aptitude lexicon.
Negative	Correspond to the negative term value in the aptitude lexicon.
Positive Judgment	If the term actual polarity is positive, the term judgment value in the lexicon is assigned to this entry, otherwise 0 is assigned.
Negative Judgment	If the term actual polarity is negative, the term judgment value in the lexicon is assigned to this entry, otherwise 0 is assigned.
Positive Appreciation	If the term actual polarity is positive, the term appreciation value in the lexicon is assigned to this entry, otherwise 0 is assigned.
Negative Appreciation	If the term actual polarity is negative, the term appreciation value in the lexicon is assigned to this entry, otherwise 0 is assigned.
Positive Affect	If the term actual polarity is positive, the term affect value in the lexicon is assigned to this entry, otherwise 0 is assigned.
Negative Affect	If the term actual polarity is negative, the term affect value in the lexicon is assigned to this entry, otherwise 0 is assigned.

3.2 Attitude Lexicons

To identify the attitude words in the opinions, the attitude lexicon build by [12] was used. It has 3,005 word entries, where each word was manually evaluated, considering three values: 0, 1, and 2, to establish its polarity (positive, negative) and its correspondence to an attitude subsystem (i.e. judgment, appreciation, affect). Here, 0 indicates the lower level and 2 the highest level. Some examples are showed in Table 2, along the values for each entry:

Table 2. Examples of words in attitude dictionary.

	Attitude Feature Value				
Word	***POS***	***NEG***	***AFF***	***JUD***	***APP***
Audaz (*bold*)	2	0	0	2	0
Suicidio (*suicide*)	0	2	0	2	0
Fértil (*fertile*)	2	0	0	2	2
Añejo (*old*)	2	1	0	0	2

Analyzing the word *suicide* in Table 2, one can notice that it has a negative connotation, but at the same time denotes a judgment term with a high level. On the other hand, *old* does not have a clear orientation (more positive than negative) but has a high value in terms of appreciation (i.e. social or aesthetics valuation). This lexicon is further referred as HLM.

The original attitude lexicon was enlarged adding the words in two different lexicons gathered by [13] to have as many words as possible, however the new words only are categorized as positive or negative, leaving judgment, appreciation and affect without values. This enlarged lexicon is here on referred as PBM.

3.3 Appraisal Groups

An appraisal group is defined as a group of words with a main adjective that belongs to an attitude subsystem type and a list of preceding modifiers [14]. Each group denotes the transformation of one or more attitude attributes. For example sentences: "It is a very boring movie", "It is a little entertaining film" and "It is not an interesting movie." In this ongoing work, to analyze opinions we consider just three kind of modifiers to define an appraisal group:

- Inversion group are words denoting a polarity inversion. (e.g. "not")
- Low Impact words that decrement the polarity value. (e.g. "little")
- High Impact words that increase the value of the polarity. (e.g. "very")

The appraisal modifiers considered are depicted in Table 3 (Spanish and English translation):

Table 3. Modifiers in appraisal phrases.

Effect	**Spanish Terms**
Inversion	No
Low Impact	Poco (little), escaso (little), escasamente (barely), exiguo (meager), falto (lacking), insuficiente (insufficient), insuficientemente (insufficiently), falta (missing)
High Impact	Muy (very), demasiado (too), bastante (quite), bastantemente (sufficienly), mucho (much), sumo (most), sumamente (most), excesivo (excessive), asaz (exceedingly), excesivamente (excessively)

4 Experiments and Results

The results obtained by the proposed approach are described in this section. The learning process was carried out with Weka, using the support vector machine (SVM) algorithm with default parameters and 10-fold cross validation.

4.1 Corpus

Sentimental Analysis has been a topic of interest in the last years. However the available corpora in Spanish are few. Fortunately, [15] made a contribution to the community by creating a corpus for sentiment analysis in Spanish. Their corpus is about movie reviews. The authors prepared the corpus according to the following criteria:

- A high number of available reviews (more than two thousand).
- When the content is generated by users, a minimum texts quality has to be ensured.

– Each user who reviews a movie has to assign a review score to the movie, which allows distinguishing whether a critics contains a favorable or unfavorable opinion.
– The web publishing license should allow to use freely the contents.

According with the previous criteria, the authors selected the mucho-cine web site. They chose 3,878 reviews (opinions). Each one has associated a reviewer, a review summary, a body review and a score given by the reviewer to the movie. The movie score is a number between 1 and 5, where 1 means an unfavorable review (negative) and 5 a favorable one (positive).

In our experiments the reviews ranked as 3 were eliminated, because they are considered neutral reviews, then the total reviews considered for our experiments were 2,625. The reviews ranked 1 and 2 where considered negative and those ranked 4 and 5 as positive. Therefore, the experiments were carried out with a two class corpus: 1,351 positive review and 1,274 negative reviews.

4.2 Experiment Outcomes

To show the importance of attitude lexicon, three experiments were performed, the first using HLM lexicon, the second using the PBM lexicon, and the third combining both lexicons. Table 4 shows the results obtained where MD1 refers to the first model and MD2 to the second.

Table 4. Results with different lexicons.

	Measure Value		
Experiment	***Precision***	***Recall***	***F-Measure***
MD1-HLM Lex	**0.634**	**0.621**	**0.615**
MD1-PBM Lex	0.575	0.564	0.556
MD1-Both	0.625	0.614	0.608
MD2-HLM Lex	**0.652**	**0.643**	**0.641**
MD2-PBM Lex	0.575	0.564	0.556
MD2-Both	0.649	0.641	0.639

We can notice that combining polarity with the information of appraisal subsystems contributed to improve classification (MD1-HLM versus MD2-HLM or MD1-Both versus MD2-Both), in terms of recall and precision, using the HLM lexicon. However, increasing the lexicon did not help to improve classification (in fact, affecting the classification, see for instance MD2-HLM versus MD2-Both), given that the new entries did not include additional information about the appraisal subsystems.

Establishing the utility of appraisal groups has been difficult mainly because no all the movie reviews have an appraisal group. To evaluate their utility, the reviews where at least one appraisal group appears, were extracted. Then two experiments were executed: the first makes use of appraisal groups and the

second is limited to perform the count of lexicon values, without the use of appraisal groups. HLM lexicon is used, because is the only one built according to the Appraisal Theory (i.e. it has values in all the entries). Table 5 shows the results of both experiments.

Table 5. Results when using appraisal groups.

	Measure Value		
Experiment	***Precision***	***Recall***	***F-Measure***
MD1	0.668	0.642	0.605
MD1 w/AG	0.655	0.634	0.599
MD2	0.689	0.681	0.667
MD2 w/AG	**0.719**	**0.706**	**0.693**

First, we observe that in the first model (MD1, more limited, since it has less information of polarity) is affected by the addition of the appraisal information. However, we also notice that when representing separately the appraisal components (MD2), the classification improves by taking advantage of the polarity identified in the analysis, that seems to be a consistent significant improvements, but we still have to test such significance.

From results obtained by MD2 w/AG, it is relevant to analyze some cases where the classification was done correctly and incorrectly, as showed below:

In the movie reviews:

- Movie review 1)"*El guion fue un tanto* ***aburrido****, además de una* ***muy mala*** *interpretación del papel por parte de los actores principales, lo* ***aceptable*** *solo fue su* ***excelente*** *musicalización*".
  The attitude terms were detected, and the modifier *"muy"* preceding to the term *"mala"* was processed by duplicating the values that the word has in the attitude lexicon and then added to the corresponding vector entries, which increases finally the negative polarity review. Therefore, the review is correctly classified as negative
- Movie review 2)"*El mayor* ***fallo*** *de Fin, protagonizada por Maribel Verdú, Clara Lago y Daniel Grao, es que es casi* ***imposible sentir*** *algún tipo de* ***empatía*** *por sus personajes.*"
  Similarly, the attitude terms are detected and their values are taken from the attitude lexicon. However, the phrase *"casi imposible sentir algún tipo de empatía"*, turns out to be a complex appraisal group that modifies the *"empatía"* word intention, but it can not be detected. So, the values that the word has in the attitude lexicon are not modified and they are directly added to the respective entries in the final vector. This problem causes that the classification model incorrectly assigns a positive polarity to the review.

still making improvements on This situation shows that the proposed method has a good performance with reviews that contain basic appraisal groups. However,

those groups that have a more complex grammatical structure can affect partially the classification of the review polarity. Therefore, we are working on improving the detection of more complex appraisal groups.

Despite the work done, the proposed method could not overcome the baseline, which used a vector representation of the reviews; this had a precision of 0.77. However, it uses vectors of 58,474 entries in contrast with our approaches that used vectors of 5 and 8 entries, and provides a good approximation in quite a short time.

5 Conclusions and Future Work

We have described a method to evaluate the polarity of opinions using the attitude subsystem of the Appraisal Theory, together with simple appraisal groups. As it was shown in experiments, when at least one appraisal group is present in a review (opinion) and separating the positive and negative values for each attitude attribute, the precision improves, obtaining a precision value quite close to that obtained by a traditional vector representation, so we would expect that in a corpus where opinions have an adequate number of appraisal groups, the advantage of our proposal would be more evident.

On the other hand, our method depends on the attitude lexicon as also showed in experiments, and then if we enrich the lexicon, we would expect a precision improvement, also.

Our results are also consistent with previous explorations of the advantages of using adverbs in sentiment analysis [16]. In our case, the contribution of adverbs in the appraisal groups was noticeable. This also opens the opportunity to consider a similar detailed attention to adverbs during our analysis of appraisal groups.

An advantage of our method is the vector dimension reduction from 58,474 to 8, which has a significant impact on storage and time processing. This can be advantageous when dealing with big data collections.

Our ongoing work consists on completing the lexicon (that unfortunately requires expert knowledge), considering more elaborated appraisal groups, and using or building other corpus appropriate to test our proposed approach and execute a qualitative analysis of the results.

We also plan to experiment adding grammatical knowledge of clauses and on different sentences lengths, as previous related research. We are also considering taking into account other syntactic class information as features for classification.

References

1. Chen H., Zimbra D.: AI and opinion mining. Intelligent Systems, IEEE, 25(3), pp. 74–80 (2010)
2. Read J., Carroll J.: Annotating expressions of Appraisal in English. Lang Resources & Evaluation, vol. 46, pp. 421–447 (2012)

3. Balahur A., Hermida J. M., Montoyo A.: Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. IEEE Transactions on Affective Computing 3(1), pp. 88–101 (2012)
4. Gao Y., Zhou L., Zhang Y., Xing C., Sun Y., Zhu X.: Sentiment classification for stock news. In: 5th International Conference on Pervasive Computing and Applications (ICPCA), pp. 99–104, IEEE Press (2010)
5. Singh V. K., Piryani R., Uddin A., Waila P.: Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In: 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), pp. 712–717, IEEE Press (2013)
6. Neviarouskaya A., Prendinger H., Ishizuka M.: Recognition of affect, judgment, and appreciation in text. In: Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, pp. 806–814 (2010)
7. Martin J. R.:Reading Positions/Positioning Readers: JUDGEMENT in English. Prospect: a Journal of Australian TESOL, 10(2), pp. 27–37 (1995)
8. Martin J. R.: Beyond Exchange: APPRAISAL Systems in English. In: Evaluation in Text, S. Hunston & G. Thompson (Eds.), pp. 142–175, Oxford: Oxford University Press (2000)
9. Martin J. R., White P. R. R.: The Language of Evaluation: Appraisal in English. London, U.K.: Palgrave (2005)
10. Taboada M., Grieve J.: Analyzing Appraisal Automatically. In: Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text, pp. 158–161, AAAI Technical Report SS-04-07, Stanford University, CA, AAAI Press (2004)
11. Wang J., Dong A.: A comparison of two text representations for sentiment analysis. In: 2010 International Conference on Computer Application and System Modeling (ICCASM), vol. 11, pp. V11–35, IEEE Press (2010)
12. Hernández L., López-López A., Medina J. E.: Classification of Attitude Words for Opinion Mining. International Journal of Computational Linguistics and Applications 2(1-2), pp. 267–283 (2011)
13. Pérez-Rosas V., Banea C., Mihalcea R.: Learning Sentiment Lexicons in Spanish. In: Proceedings of the International Conference on Language Resources and Evaluations (LREC 2012), pp. 3077–3081 (2012)
14. Whitelaw C., Garg N., Argamon S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 625–631, ACM (2005)
15. Cruz F. L., Troyano J. A., Enríquez F., Ortega J.: Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. Procesamiento del Lenguaje Natural, vol. 41, pp. 73–80 (2008)
16. Benamara F., Cesarano C., Picariello A., Reforgiato D., Subrahmanian V. S.: Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM) (2007)

Clasificación de servicios Web mediante una red neuronal artificial usando n-gramas de palabras

José A. Reyes-Ortiz, Maricela Bravo

Universidad Autónoma Metropolitana, Unidad Azcapotzalco,
Departamento de Sistemas, Distrito Federal,
México

{jaro, mcbc}@correo.azc.uam.mx

Resumen. Este artículo presenta un enfoque basado en *n-gramas* de palabras para la clasificación automática de servicios Web utilizando una red neuronal artificial de tipo perceptrón multicapa. Los servicios Web contienen información de gran utilidad para lograr una clasificación basada en la funcionalidad del mismo. El enfoque se basa en *n-gramas* de palabras extraídas de la descripción del servicio web para determinar su pertenencia a una categoría. La experimentación realizada muestra resultados prometedores, logrando una clasificación con una *medida F*= 0.91 utilizando *unigramas (1-gramas)* de palabras (características compuestas por una unidad léxica) y una ponderación *TF-IDF*.

Palabras clave: Servicios web, modelo espacio vectorial, clasificación de servicio web, *n-gramas* de palabras, red neuronal perceptrón multicapa.

1. Introducción

Los servicios Web son componentes de software reutilizables a través de los cuales es posible construir e integrar nuevas aplicaciones sin tener que implementar todos los elementos de un sistema. Actualmente los servicios Web han cobrado mayor popularidad debido a su proliferación para el ofrecimiento de servicios de almacenamiento y administración de recursos en la nube. Los servicios Web se encuentran disponibles tanto en repositorios públicos como privados a través de las descripciones de los servicios. Existen varios repositorios de servicios Web públicos: a) el directorio de Servicios Web SOAP soportado por Membrane[1]; b) el repositorio de servicios Web Visual Web Service[2]; c) el repositorio de servicios Web XMethods[3]; d) ProgrammableWeb[4]; e) OWLS-TC[5] es una colección de servicios de prueba

[1] http://www.service-repository.com/
[2] http://www.visualwebservice.com
[3] http://www.xmethods.com
[4] http://www.programmableweb.com
[5] http://projects.semwebcentral.org/projects/owls-tc/

recuperados con sus respectivas anotaciones en OWL-S. Estas anotaciones expresan la semántica de los elementos de un servicio Web.

Las descripciones de servicios Web se realizan mediante el lenguaje estándar WSDL, dicha descripción consiste de un archivo de texto basado en XML, dentro del cual se definen los elementos necesarios para la invocación del servicio.

Para que los programadores y desarrolladores de aplicaciones puedan hacer uso de los servicios necesitan buscarlos dentro de los repositorios de servicios. Esta tarea se le conoce comúnmente como descubrimiento de servicios. Sin embargo, el descubrimiento de servicios sigue siendo una tarea ardua y propensa a errores, ya que la mayoría de los repositorios ofrecen mecanismos de búsqueda basados en coincidencia de palabras clave. Aunado a este problema está el hecho de que los repositorios de servicios están organizados principalmente mediante estructuras estáticas que no permiten una organización flexible y dinámica de los servicios.

Este trabajo de investigación tiene como objetivo principal mejorar la estructura organizacional de los repositorios de servicios Web de tal forma que se facilite el descubrimiento de servicios. La principal contribución de este artículo se centra en un algoritmo de clasificación de servicios Web utilizando *n-gramas* de palabras. Como resultado se obtienen colecciones de servicios Web organizados por temas, su búsqueda se agiliza, consumiendo menos recursos, debido a que se realiza entre servicios dentro de la misma categoría.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se presenta la descripción de servicio Web. La Sección 3 expone un acercamiento al estado de arte de los trabajos relacionados. La Sección 4 muestra el enfoque propuesto. La Sección 5 presenta la experimentación y los resultados obtenidos para la clasificación de servicios web. Finalmente, las conclusiones son presentadas en la Sección 6.

2. Descripción de servicios web

El lenguaje de descripción de servicios recomendado para la implementación de servicios Web es llamado *Web Service Description Language (WSDL),* el cual es, actualmente, un estándar aceptado por la *W3C*. *WSDL* define una gramática XML para describir servicios conectados como una colección de nodos de comunicación capaces de intercambiar mensajes[6].

En este trabajo son consideradas la versión *WSDL 2.0*, la más reciente y que incorpora cambios significados en la descripción del servicio con respecto a la versión *WSDL 1.1*. *WSDL 2.0* cambia la etiqueta <*definitions*> por la etiqueta <*description*> (ver Fig. 1). Las diferencias más significativas entre *WSDL 1.1* y WSDL 2.0 son: el *targetNameSpace* es un atributo requerido en *WSDL 2.0*; la construcción de mensajes es eliminada en *WSDL 2.0*; la sobrecarga de operadores no es soportada en *WSDL 2.0*; *PortType* ha cambiado de nombre a *Interface*; la herencia en *Interface* es soportada mediante el uso de atributos extendidos; y *Port* ha cambiado el nombre por *Endpoint*.

[6] http://www.w3.org

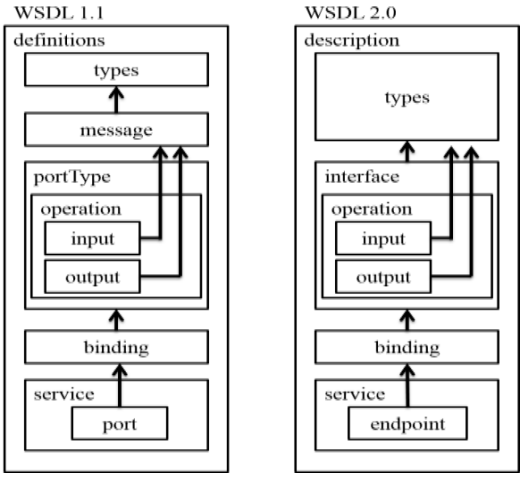

Fig. 1. WSDL 1.1 y WSDL 2.0.

3. Trabajos relacionados

Las descripciones de los servicios web, OWLS o WSDL, contienen texto libre (lenguaje natural), además de nombres de parámetros, nombres de datos, complejos y simples, nombres de operaciones y parámetros de entrada-salida. Diversos trabajos han utilizado esta información para una gran variedad de tareas relacionadas al procesamiento de textos. Una de estas tareas es calcular la similitud semántica entre servicios [2, 3, 8, 10, 15], clasificar los servicios basada en su contenido [4, 5, 11, 12, 14, 17, 19] y agrupamiento de servicios Web [16].

En el contexto de la clasificación de servicios Web, se han propuesto enfoques utilizando la colección de servicios OWLS-TC para una clasificación supervisada. El trabajo presentado en [19] utiliza las descripciones del servicio web, las descripciones de las operaciones y las descripciones de las entradas/salidas, con la finalidad de clasificar los servicios; en [14] se utiliza el algoritmo de máquinas de soporte vectorial, características de similitud entre los términos utilizando *WordNet* y una taxonomía del dominio, para clasificar, de manera eficiente, los servicios web de la colección; y en [9] se ha propuesto la clasificación automática de servicios web basada en la teoría de conjuntos aproximados, ellos proponen una representación de las descripciones con espacios vectoriales y una ponderación de los términos basada en la entropía.

No sólo existen trabajos que utilicen la colección mencionada, también, se han propuesto trabajos bajo colecciones privadas de servicios Web. En [11] utilizan una colección de 364 servicios web, obtenidos de repositorios públicos en la Web, para evaluar su metodología propuesta en la clasificación y predicción de la calidad de un servicio web basada en un conjunto de atributos de calidad, a saber: fiabilidad, documentación, rendimiento, respuesta y tiempo. En el contexto de la clasificación no supervisada, el trabajo presentado en [17] propone método para la clasificación automática a partir de las etiquetas presentes en el documento de descripción de cada

servicio, su método fue probado con 951 archivos WSDL distribuidos en 19 categorías. El trabajo propuesto en [12] presenta un enfoque de minería de textos para la clasificación automática de servicios web en dominios específicos y la identificación de conceptos clave en la documentación textual de los servicios. Su enfoque fue validado con un conjunto de 600 servicios web categorizados en 8 clases. Finalmente, el trabajo expuesto en [4] propone un mecanismo basado en reglas heurísticas que permite a los autores de servicios Web, clasificar sus servicios en una taxonomía manejada por un repositorio.

4. Enfoque propuesto

Este artículo presenta la clasificación de servicios Web utilizando sus descripciones textuales y nombres de parámetros. Esta problemática se aborda mediante el paradigma de clasificación de textos, tarea que consiste en asociar categorías predefinidas a partir del análisis de un texto. En la Fig. 2 se muestra la arquitectura general del enfoque propuesto.

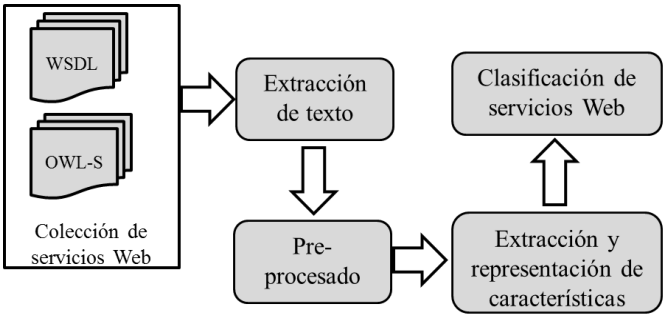

Fig. 2. Arquitectura para la clasificación de Servicios Web.

El enfoque analiza una colección de servicios Web para su clasificación. La colección se conforma de la descripción en WSDL y de su extensión semántica en OWL-S de cada servicio. Bajo esta arquitectura, la primera etapa es la extracción de textos a partir de los archivos de descripción de los servicios Web, para ello, se realiza un analizador de archivos WSDL y OWL-S con la finalidad de extraer el texto en lenguaje natural y los nombres de los elementos relevantes. Después, los textos extraídos son pre-procesados para obtener las unidades léxicas (palabras simples) que describen a cada servicio. El conjunto de unidades léxicas, características de cada servicio web, son representadas como *1-gramas*, *2-gramas* o *3-gramas*, mediante una ponderación de los términos, en un modelo espacio vectorial [1]. Finalmente, los vectores son utilizados para la clasificación de los servicios mediante una red neuronal de tipo perceptrón multicapa.

4.1. Extracción de textos

Los documentos de descripción (WSDL) y los documentos de descripción semántica de la funcionalidad (OWL-S), a partir de la colección de servicios Web,

son analizados con la finalidad de identificar y extraer la información textual, la cual es de utilidad para lograr la clasificación basada en el contenido. De esta manera, se extrae el contenido de las etiquetas *serviceName* y *textDescription* de la clase *Profile* a partir de las descripciones de la funcionalidad (OWL-S), estas etiquetas contienen texto en lenguaje natural. Por su parte, a partir del WSDL se extrae el nombre del servicio (*wsdl:service name*), el nombre de las operaciones (*wsdl:operation name*) y los nombres de los tipos de datos de los mensajes, ya sea simples o complejos (*xsd:simpleType name* y *xsd:complexType name*). Esta información en lenguaje natural y los nombres de los elementos del servicio se utilizan para representar y, posteriormente, clasificar los servicios basándose en su contenido.

4.2. Preprocesamiento de los textos

La primera tarea, después de tener el texto de los servicios, es la segmentación en palabras. En el caso de los servicios web, es común encontrar nombres de servicios, operaciones o tipos de datos con palabras compuestas. Para logar la segmentación de las palabras, se transforman los nombres compuestos en su forma canónica simple, es decir, en unidades léxicas, considerando el cambio de minúscula a mayúscula y el subguión como separadores de palabras. Por ejemplo, *getAddressLocation* o *get_address_location* se descomponen en las siguientes unidades léxicas: [*get*] [*Address*] [*Location*].

Adicionalmente, los textos de los servicios se normalizan, mediante la aplicación de una conversión a minúsculas, eliminación de signos de puntuación y supresión de las palabras vacías, es decir, palabras que no aportan significado y por lo tanto se consideran no funcionales para la clasificación de servicios basada en el contenido.

4.3. Extracción y representación de las características de los servicios

El conjunto de unidades léxicas (palabras) normalizadas y filtradas son representadas en el modelo espacio vectorial [1]. Este modelo se utiliza para representar los textos de una manera formal utilizando los términos como características, las cuales pueden ser unidades léxicas simples (*1-gramas*), secuencias de dos palabras (*2-gramas*) o cualquier secuencia de palabras (*n-gramas*).

En este artículo complementamos el modelo de espacio vectorial con la representación de bolsa de palabra, por sus siglas en inglés BOW [13], el cual consiste en una colección de textos y su vocabulario de términos (características). Cada texto de los servicios web es representado como un vector $S_j = (w_{1j}, w_{2j} \dots w_{nj})$, donde cada componente w_{ij} expresa la importancia o frecuencia ponderada que produce la característica i, unidad léxica (*1-grama*) o secuencia de palabras (n-grama), del vocabulario en el texto j del servicio Web. En este trabajo nos centramos en medir la influencia de *1-gramas*, *2-gramas* y *3-gramas* de palabras como características (términos del vocabulario) dentro del espacio vectorial para la clasificación de servicios Web.

Existen diferentes enfoques para obtener la importancia o ponderación de los términos del vocabulario sobre un texto. Este artículo se basa en un pesado *booleano*,

un pesado de frecuencia de aparición del termino (*TF*) y un pesado basado en la frecuencia de aparición del término en la colección de textos (*TF-IDF*).

La ponderación *booleana* de un término calcula el peso asignando un valor de 0 si el término del vocabulario no aparece en el texto y un valor de 1 si el término está presente. Mientras que, la ponderación basada en la frecuencia de aparición (*TF*) calcula el número de veces que un término del vocabulario aparece en un texto: $w_{ij} = TF(t_i, S_j)$. Esta ponderación otorga importancia a los términos más frecuentes, sin embargo no es una ponderación normalizada lo que puede provocar valores alejados entre los componentes del vector.

Finalmente, también se tiene la ponderación basada en la frecuencia de aparición del término en la colección de textos (*TF-IDF*), la cual captura la importancia que tiene un término para un texto de descripción de servicio Web. Esta ponderación utiliza la frecuencia de aparición de un término del vocabulario en un texto $TF(t_i, S_j)$ y la frecuencia inversa que determina si el término es común en la colección de textos $IDF(t_i, S_j) = \log \frac{|S|}{1+|s \in S\,:\, t_i \in s|}$. Entonces, la fórmula final para el cálculo de *TF-IDF* queda de la siguiente manera: $w_{ij} = TF(t_i, S_j) \times IDF(t_i, S_j)$.

4.4. Clasificación de servicios web

La clasificación de servicios web se basa en los vectores de cada servicio ponderado de manera *booleana*, *TF* o *TF-IDF*. La clasificación presentada en este artículo se considera una clasificación supervisada, ésta ha sido ampliamente utilizada en el campo del aprendizaje automático para estimar la función predictora de cada clase de nuestra colección. Por ello, se dividirá la colección de servicios web en un conjunto de entrenamiento y un conjunto para las pruebas.

El objetivo de esta etapa es construir un clasificador de servicios web considerando 9 categorías: *Communication* (Comunicación), *Economy* (Economía), *Education* (Educación), *Food* (Comida), *Geography* (Geografía), *Medical* (Medicina), *Simulation* (Simulación), *Travel* (Turismo) y *Weapon* (Armamento).

La tarea de clasificación se lleva a cabo mediante una red neuronal artificial de tipo perceptrón multicapa con aprendizaje supervisado. El perceptrón multicapa es una Red Neuronal Artificial (RNA) formada por múltiples capas de neuronas cuya finalidad es resolver problemas de múltiples clases. Este tipo de redes neuronales, presentadas en [7], [9] y [18], son considerados algoritmos de clasificación supervisada y están inspirados en las redes neuronales biológicas del cerebro humano. Bajo un esquema matemático, un perceptrón multicapa es una función no lineal compleja con un conjunto de unidades neuronales interconectadas, compuesta de una capa de entrada y una de salida, esta red es entrenada con un conjunto de servicios y, entonces, realiza los ajustes de parámetros correspondientes para arrojar salidas similares con los datos de prueba.

La idea es evaluar la tarea de clasificación de servicios con el perceptrón multicapa en combinación con las ponderaciones (*booleana*, *TF* o *TF-IDF*) y la formación de términos (*1-gramas*, *2-gramas* y *3-gramas*), con la finalidad de encontrar la mejor solución en términos de precisión. La implementación de la red neuronal artificial de tipo perceptrón multicapa con aprendizaje supervisado, para la clasificación de servicios Web, se ha llevado a cabo mediante la herramienta WEKA [6].

5. Experimentación y resultados

La evaluación del enfoque propuesto fue realizada con la versión 3.0 de la colección OLWS-TC[7], la cual consta de 1129 servicios descritos usando WSDL y OWL-S. Los servicios de esta colección está preclasificados en las siguientes categorías: *Communication* (Comunicación), *Economy* (Economía), *Education* (Educación), *Food* (Comida), *Geography* (Geografía), *Medical* (Medicina), *Simulation* (Simulación), *Travel* (Turismo) y *Weapon* (Armamento). Esta colección fue dividida en dos grupos: 899 para el aprendizaje del modelo de clasificación y 230 para las pruebas.

La experimentación consiste en evaluar el algoritmo de clasificación, red neuronal artificial de tipo perceptrón multicapa, combinándolo con la representación basada en *1-gramas*, *2-gramas* y *3-gramas*, además de evaluar los tres tipos de ponderaciones para los términos: *booleano*, *TF* y *TF-IDF*. Todas las configuraciones de los experimentos fueron ejecutadas sobre el conjunto de 899 descripciones de servicios para lograr el aprendizaje del modelo de predicción y, posteriormente, evaluadas con el conjunto 230 descripciones de prueba. Además, todos los experimentos se realizaron con los siguientes parámetros de la red neuronal de tipo perceptrón multicapa: *número de capas ocultas* = $\frac{\#\ atributos + \#\ clases}{2}$*; tasa de aprendizaje = 0.3; y, momentum=0.2*.

La evaluación de todos los experimentos se realizó utilizando las métricas de *Precisión (P)*, *Recuerdo (R)* y *medida F* ampliamente utilizadas en cualquier tarea de clasificación, en nuestro caso, clasificación de textos. Estas métricas comparan los resultados del clasificador a evaluar con los valores externos de confianza (servicios web preclasificados), utilizando los siguientes valores: a) *Verdadero Positivo (VP)* es el número de predicciones correctas del clasificador de servicios que corresponden al juicio externo de confianza (servicios preclasificados); *Verdadero Negativo (VN)* es el número de predicciones correctas del clasificador de servicios que no corresponden al juicio externo de confianza; *Falso Positivo (FP)* corresponde al número predicciones incorrectas del clasificador de servicios que corresponden al juicio externo de confianza; y, finalmente *Falso Negativo (FN)* es el número de predicciones incorrectas del clasificador de servicios que no corresponden al juicio externo de confianza.

Bajo estos criterios, se emplea la *Precisión (P)* para evaluar los algoritmos en términos de los valores de predicciones positivas, la cual se define como:

$$P = \frac{VP}{VP + FP} \tag{1}$$

También, se utiliza el *Recuerdo (R)* para expresar la tasa de correspondencias correctas con los servicios preclasificados de manera externa con una alta confianza, el cual se define como:

$$R = \frac{VP}{VP + FN} \tag{2}$$

[7] http://projects.semwebcentral.org/projects/owls-tc/

Finalmente, la *medida F* que representa la media armónica entre *Precisión* y *Recuerdo*, la cual tiene como fundamento el obtener un valor único ponderado entre ellas y se define como:

$$medida\ F = 2 * \frac{P * R}{P + R} \tag{3}$$

Los experimentos se han organizado para medir el impacto de los tipos de representación de los términos y las ponderaciones utilizadas en las características de las descripciones de los servicios Web. Bajo esta consideración, la Tabla 1 muestra los resultados, por categoría, de los experimentos utilizando la ponderación *booleana* para las características *1-gramas*, *2-gramas y 3-gramas*. También, se proporciona un peso promedio, resultado de considerar la *medida F* y un factor que refleja la importancia (número de servicios) de cada clase.

Tabla 1. Resultados de la clasificación considerando la ponderación *booleana.*

	1-grama			***2-grama***			***3-grama***		
Categoría	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Comunicación	0.98	0.96	0.97	0.96	0.81	0.88	0.87	0.82	0.84
Economía	0.98	0.97	0.97	0.98	0.97	0.97	0.96	0.95	0.95
Educación	0.90	0.97	0.94	0.91	0.96	0.93	0.89	0.94	0.91
Comida	0.34	0.30	0.32	0.26	0.24	0.25	0.25	0.23	0.24
Geografía	1.00	0.98	0.99	1.00	0.98	0.99	0.98	0.96	0.97
Medicina	0.36	0.40	0.38	0.30	0.33	0.31	0.3	0.31	0.30
Simulación	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.98	0.97
Turismo	0.99	0.92	0.95	0.99	0.96	0.97	0.97	0.94	0.95
Armamento	1.00	0.93	0.96	0.97	0.95	0.96	0.96	0.95	0.95
Peso promedio	0.881	0.881	**0.880**	0.873	0.869	**0.870**	0.851	0.852	**0.851**

En la Tabla 2 se muestran los resultados de la ponderación *TF* para los términos *1-gramas, 2-gramas y 3-gramas* para la clasificación de servicios con la red neuronal perceptrón multicapa.

Tabla 2. Resultados por categoría utilizando la ponderación *TF*.

	1-grama			***2-grama***			***3-grama***		
Categoría	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Comunicación	0.92	0.88	0.90	0.90	0.85	0.87	0.85	0.81	0.83
Economía	0.97	0.95	0.96	0.95	0.93	0.94	0.85	0.84	0.84
Educación	0.86	0.86	0.86	0.84	0.84	0.84	0.81	0.82	0.81
Comida	0.24	0.20	0.21	0.22	0.20	0.21	0.20	0.20	0.20
Geografía	0.91	0.85	0.88	0.88	0.82	0.85	0.83	0.78	0.80
Medicina	0.28	0.28	0.28	0.26	0.26	0.26	0.20	0.21	0.20
Simulación	0.85	0.69	0.76	0.82	0.67	0.74	0.76	0.61	0.68
Turismo	0.81	0.95	0.87	0.78	0.90	0.84	0.73	0.81	0.77
Armamento	0.86	0.78	0.82	0.80	0.72	0.76	0.74	0.65	0.69

	1-grama			***2-grama***			***3-grama***		
Categoría	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Peso promedio	0.815	0.816	**0.815**	0.792	0.791	**0.790**	0.732	0.733	**0.732**

Finalmente, la Tabla 3 presenta los resultados de la ponderación *TF-IDF* para los términos *1-gramas, 2-gramas y 3-gramas* en la clasificación de servicios con la red neuronal perceptrón multicapa.

Tabla 3. Resultados utilizando la ponderación *TF-IDF*.

	1-grama			***2-grama***			***3-grama***		
Categoría	***P***	***R***	***F***	***P***	***R***	***F***	***P***	***R***	***F***
Comunicación	0.98	0.97	0.97	0.97	0.96	0.96	0.92	0.91	0.91
Economía	0.98	0.98	0.98	0.97	0.98	0.97	0.91	0.92	0.91
Educación	0.91	0.97	0.94	0.91	0.97	0.94	0.85	0.89	0.87
Comida	0.59	0.55	0.57	0.41	0.35	0.38	0.31	0.25	0.28
Geografía	0.99	0.98	0.92	0.97	0.97	0.97	0.87	0.88	0.87
Medicina	0.56	0.57	0.56	0.39	0.41	0.40	0.29	0.39	0.33
Simulación	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.96	0.96
Turismo	0.94	0.93	0.93	0.99	0.92	0.95	0.89	0.85	0.87
Armamento	0.98	0.98	0.98	0.99	0.95	0.97	0.93	0.87	0.90
Peso promedio	0.904	0.915	**0.910**	0.885	0.888	**0.886**	0.813	0.820	**0.816**

Los resultados mostrados en las Tablas 1, 2 y 3 hacen notar que la mejor alternativa es considerar de formación de términos *1-gramas* y su ponderación *TF-IDF* para la clasificación de servicios Web mediante una red neuronal perceptrón multicapa. Con esta configuración se logra un 91 % de servicios clasificados correctamente.

Los resultados de la experimentación demuestran la efectividad de nuestro enfoque para la clasificación de servicios Web. A pesar de que los resultados no son alentadores para dos clases (Medicina y Comida), el enfoque puede ayudar a los desarrolladores de servicios web a localizar y reutilizar los componentes de software existentes en un repositorio desorganizado de servicios.

6. Conclusiones

Este artículo ha presentado un enfoque de clasificación de servicios web utilizando una red neuronal artificial de tipo perceptrón multicapa y mediante n-gramas de palabras. Se han presentado diversos experimentos, en los cuales se combinaron la composición de términos del vocabulario (*1-gramas*, *2-gramas* y *3-gramas*) con las ponderaciones de estos términos (*booleano*, frecuencia de aparición, y frecuencia de aparición inversamente proporcional a su aparición en toda la colección). A partir de esta experimentación, se hace notar que nuestra propuesta de clasificación de servicios web ha logrado un 91 % de eficiencia.

Las principales contribuciones de este trabajo son: a) las combinaciones de composiciones de términos y sus ponderaciones en la experimentación para la clasificación de servicios Web; b) la extracción de términos de la colección para formar el vocabulario; y c) el modelo de clasificación, basado en una red neuronal perceptrón multicapa para la clasificación de servicios Web mediante sus descripciones OWLS y WSDL.

Como trabajo futuro, la clase descubierta por el clasificador puede ser almacenada, mediante la instanciación en una ontología para tareas de extracción de información, descubrimiento y composición de servicios Web.

Referencias

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley, Boston (1999)
2. Bravo, M., Alvarado, M.: Similarity measures for substituting Web services. International Journal of Web Services Research, 7 (3), pp. 1–29 (2010)
3. Bravo, M., Rodríguez, J., Reyes, A.: Enriching Semantically Web Service Descriptions. In: On the Move to Meaningful Internet Systems: OTM 2014 Conferences, pp. 776–783, Springer Berlin Heidelberg (2014)
4. Corella, M.Á., Castells, P.: A heuristic approach to semantic web services classification. In: Knowledge-Based Intelligent Information and Engineering Systems, pp. 598–605 (2006)
5. Chen, L., Zhang, Y., Song, Z.L., Miao, Z.: Automatic web services classification based on rough set theory. Journal of Central South University, 20, pp. 2708–2714 (2013)
6. Garner, S.R.: Weka: The Waikato environment for knowledge analysis. In: Proc. of the New Zealand Computer Science Research Students Conference. pp. 57–64 (1995)
7. Hassoun, M.H.: Fundamentals of Artificial Neural Networks. Cambridge, MA: MIT press (1999)
8. Hau, J., Lee, W., Darlington, J.: A semantic similarity measure for semantic web services. In: Web Service Semantics Workshop at WWW, pp. 10–14 (2005)
9. Haykin, S.: Neural Networks and Learning Machines. 3rd Edition, Pearson Education, New Jersey (2008)
10. Liu, F., Shi, Y., Yu, J., Wang, T., Wu, J.: Measuring similarity of web services based on WSDL. In: 2010 IEEE International Conference on Web Services, pp. 155–162 (2010)
11. Mohanty, R., Ravi, V., Patra, M.R.: Web-services classification using intelligent techniques. Expert Systems with Applications, 37(7), pp. 5484–5490 (2010)
12. Nisa, R., Qamar, U.: A text mining based approach for web service classification. Information Systems and e-Business Management, pp. 1–18 (2014)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, 34(1), pp. 1–47 (2002)
14. Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., Bouguettaya, A.: Web service classification using support vector machine. In: 22nd IEEE International Conference on Tools with Artificial Intelligence, Vol. 1, pp. 3–6 (2010)
15. Wu, J., Wu, Z.: Similarity-based web service matchmaking. In: IEEE International Conference on Computing, Vol. 1, pp. 287–294 (2005)
16. Wu, J., Chen, L., Zheng, Z., Lyu, M. R., Wu, Z.: Clustering web services to facilitate service discovery. Knowledge and information systems, 38(1), pp. 207–229 (2014)

17. Yuan-jie, L., Jian, C.: Web service classification based on automatic semantic annotation and ensemble learning. In: 26th International on Parallel and Distributed Processing Symposium Workshops & PhD Forum, pp. 2274–2279 (2012)
18. Zak, S.H.: Systems and Control. University Press, New York (2003)
19. Zhang, J., Pan, D.: Web Service Classification. (2008)

Textos subjetivos y objetivos: un estudio sobre reconocimiento de subjetividad en corpus

Antonio Reyes, Sonia Morett

Instituto Superior de Intérpretes y Traductores,
Laboratorio de Tecnologías Lingüísticas, México DF,
México

{antonioreyes,soniamorett}@isit.edu.mx

Resumen. En este artículo se describe un estudio exploratorio en el que se analiza cómo la subjetividad es expresada a través de elementos léxicos (subjetivemas) en un corpus de artículos periodísticos en español. En particular, dos tipos de textos constituyen el corpus: la nota informativa (texto expositivo) y la columna de opinión (texto argumentativo). El objetivo es identificar elementos que, en tanto subjetivos, son utilizados por los hablantes para persuadir de manera subyacente a los lectores, independientemente del tipo de género al que pertenece el texto. Los resultados hasta el momento han permitido reconstruir un panorama amplio de los subjetivemas que, sin importar el tipo de discurso, los hablantes emplean al emitir valoraciones subjetivas. El reconocimiento de estos elementos constituye un recurso relevante para tareas tales como análisis de sentimientos, minería de opiniones o reconocimiento automático de lenguaje persuasivo.

Palabras clave: Subjetividad, subjetivema, reconocimiento automático de subjetividad, lenguaje persuasivo.

1. Introducción

De acuerdo con [6], el término subjetividad no debe entenderse necesariamente como lo opuesto a objetividad, sino como una estrategia lingüística que permite a los hablantes involucrarse y verse reflejados en las acciones que codifican por medio del lenguaje. En este escenario, el presente trabajo describe la identificación de elementos léxicos subjetivos en español que dan cuenta de cómo los hablantes imprimen en sus enunciados sus posturas éticas, estéticas y, en definitiva, ideológicas, sobre diferentes tópicos a partir de un corpus integrado por dos tipos de textos: notas informativas, que tradicionalmente son consideradas como textos objetivos, y columnas de opinión, que a su vez, son consideradas como textos subjetivos.

Conceptualmente, este trabajo se cimienta en la noción de subjetivema introducida por [10] en la teoría de la enunciación, así como en los aportes realizados al estudio del léxico valorativo desde la teoría de la valoración ([17,13]) inserta en

el marco de la Lingüística Sistémico Funcional. Ahora bien, por subjetivema se entiende aquella forma o elemento léxico que porta una carga de subjetividad a los enunciados respecto a lo narrado o descrito. En este sentido, los subjetivemas se emplean para externar estados emocionales o valoraciones y apreciaciones de tipo cognitivo hacia personas, objetos, hechos o situaciones. Así, estos pueden corresponderse con *entidades* (parsimonia, apremio, inteligencia, torpeza), *cualidades* (grande, lejano, limpio, sucio, correcto, deshonesto), *procesos* (amar, odiar, susurrar, vociferar) u *opiniones* (juiciosamente, afortunadamente, lastimosamente). Dada esta heterogeneidad, los subjetivemas no son dependientes de categoría gramatical alguna; hecho que los vuelve difíciles de sistematizar y reconocer automáticamente. Aunado a lo anterior, los subjetivemas (ya sean adjetivos, adverbios, sustantivos o verbos) pueden ser o bien afectivos o bien evaluativos. En lo que respecta a los subjetivemas afectivos, estos dan cuenta de las reacciones emocionales tales como enfado, miedo, tristeza, alegría o sorpresa; en tanto que los subjetivemas evaluativos se relacionan con la transmisión de juicios de valor y apreciaciones de otra índole.

Nuestro objetivo, por tanto, consiste en i) identificar elementos concretos que son utilizados por los hablantes para codificar lingüísticamente la subjetividad en texto; y ii) determinar si este conjunto de elementos que, en tanto subjetivemas, son usados para transmitir valoraciones y apreciaciones subjetivas de manera directa e indirecta. Para lograrlo, se ha construido un corpus de artículos periodísticos integrado por dos tipos de textos: notas informativas (texto expositivo) y columnas de opinión (texto argumentativo). Los resultados han sido evaluados con una tarea de reconocimiento automático de subjetividad.

El resto del artículo está organizado de la siguiente forma: en la Sección 2 se introduce el marco conceptual que sustenta este trabajo. En la Sección 3 se describe el proceso para la identificación de subjetivemas; en tanto que en la Sección 4 se presenta la evaluación que se realizó utilizando el corpus de artículos periodísticos. Los resultados se detallan en la Sección 5. Finalmente, en la Sección 6 se sintetizan las principales conclusiones y se delinea el trabajo futuro.

2. Marco conceptual

En términos léxicos, el Diccionario de la Real Academia Española (DRAE) define *objetivo* como aquello que es "perteneciente o relativo al objeto en sí mismo, con independencia de la propia manera de pensar o de sentir"; en tanto que lo *subjetivo* es aquello que es filtrado por el sujeto, es decir, lo "perteneciente o relativo a nuestro modo de pensar o de sentir, y no al objeto en sí mismo" (cf. DRAE [2001]). Dada esta división, lo objetivo siempre será distinto a lo subjetivo. Sin embargo, si concebimos el lenguaje no sólo como un sistema rígido y perfectamente definido (tal cual se infiere de las definiciones previas), sino como un sistema de opciones semánticas, donde el hablante elige, dentro de todas las posibilidades a su alcance, aquéllas que mejor satisfacen sus necesidades comunicativas, entonces la división que se propone en el DRAE no es funcional,

puesto que el proceso de producción de los enunciados se inicia con la elección de los significados para, posteriormente, desembocar en la selección de la estructura discursiva y el léxico más conveniente para la transmisión de aquello que desea comunicar, tanto explícita como implícitamente.

En el escenario que supone esta visión más funcional, la interacción social a través del lenguaje implica una constante toma de decisiones que dependen de varios factores lingüísticos y extralingüísticos. No obstante, caer en el relativismo de afirmar que todo en el lenguaje es subjetivo conduce a posiciones que niegan la posibilidad de acceder al análisis crítico del discurso o de un simple acto de habla. En tal sentido, [10] reconoce como simplista y reduccionista señalar que en el lenguaje, la subjetividad está en todas partes; antes bien, es más productivo y realista reconocer que todo acto comunicativo lleva la marca de su hablante en modos y grados diversos. Vista desde esta perspectiva, la subjetividad recae particularmente en el proceso de la enunciación y en el contexto en la que ésta se produce[1]. Por lo tanto, hablar de subjetividad implica el desarrollo de una metodología que permita reconocer y sistematizar qué elementos formales están presentes al momento de la enunciación.

2.1. Teoría de la enunciación

Al situarnos en el plano de la enunciación, la descripción lingüística deja de centrarse en el sistema para también considerar quién produce el acto comunicativo, con qué recursos lingüísticos y con qué intención. Respecto a los recursos lingüísticos de la enunciación, [1] señala como elementos indicadores de subjetividad la deixis personal, la espacial y la temporal. Asimismo, se refiere a recursos modales que dotan de este tipo de contenido a los enunciados y apunta que el significado léxico prototípico de algunas palabras lo transporta de manera implícita. A partir de estos indicadores de subjetividad, [10] desarrolla de manera detallada la teoría de la enunciación que focaliza los siguientes tipos de subjetividad: deíctica, modal y léxica.

Por un lado, de acuerdo con la autora, la manifestación más notoria de la subjetividad se observa por la presencia de unidades deícticas, las cuales dependen absolutamente del contexto de la enunciación, pues hacen patente la existencia de un hablante que se refiere a otro u otros y elabora su discurso desde un lugar y un tiempo determinados. En lingüística se acepta por consenso la existencia de tres categorías deícticas: la personal, la espacial y la temporal. [5] se refirió, además, a la existencia de otra categoría deíctica, la deixis social, que se expresa mediante marcas en los enunciados que actualizan el contexto social de la enunciación.

Por otro lado, la subjetividad modal suele describirse como la actitud del hablante respecto al contenido de su discurso. Tal actitud se expresa de modo transversal en diferentes niveles de análisis de la lengua: i) el estilo de la

[1] Es importante señalar que este contexto es bastante amplío pues no sólo se concibe como un contexto exclusivamente lingüístico, sino como un contexto que, además, se compone de factores sociales, culturales, económicos, etc.

enunciación (directo o indirecto); ii) los modos gramaticales (indicativo, subjuntivo e imperativo); iii) la intención ilocutiva de los actos de habla (asertiva, directiva, comisiva, expresiva, performativa); iv) el modo de aserción (constativo, hipotético, obligativo); v) la polaridad del acto de habla (atenuada, intensificada o neutra); etc.

Finalmente, la subjetividad léxica se define a partir de la presencia de palabras que en el plano discursivo funcionan como catalizadores de información valorativa y/o apreciativa. Particularmente, este tipo de subjetividad es la que le da sentido al concepto de subjetivema que se introdujo en la Sección1. En relación con este tipo de subjetividad y con el concepto mismo de subjetivema, [10] señala que si bien "toda unidad léxica es, en un cierto sentido, subjetiva, dado que las *palabras* de la lengua no son jamás otra cosa que símbolos sustitutivos e interpretativos de las *cosas*", existen palabras o expresiones que portan una mayor carga de subjetividad respecto a lo narrado o descrito, las cuales se emplean para externar estados emocionales o valoraciones y apreciaciones de tipo cognitivo hacia personas, objetos, hechos o situaciones.

A partir del concepto de subjetivema, expuesto en los párrafos previos, se aborda la problemática de la subjetividad en texto en este trabajo; puesto que el subjetivema, como entidad formal, permite delimitar la subjetividad en términos de elementos lingüísticos concretos. En este sentido, en la siguiente sección se describe la metodología que se siguió para identificar un conjunto de subjetivemas propios del español.

3. Identificación de subjetivemas

Como se describió en la Sección 1, no es posible relacionar los subjetivemas con una categoría gramatical en específico. Dada la información que los hablantes codifican a través de ellos, estos pueden pertenecer a la categorías sustantivo, adjetivo, adverbio y verbo. Aunque para el caso de los adverbios (en específico los terminados en -mente) y los adjetivos se han identificado ciertos patrones morfológicos, tales como los prefijos anti-, contra- y los sufijos -ucho, -ete, que al unirse a una base léxica neutra le aportan a su significado un componente evaluativo o afectivo, en general, no hay elementos formales que distingan a los subjetivemas de otras formas léxicas, lo que dificulta su reconocimiento tanto manual como automático. Así, [10] advierte que "los subjetivemas constituyen un conjunto de hechos demasiado heterogéneos para que se pueda esperar elaborar tan pronto un procedimiento de cálculo del porcentaje de subjetividad que implica un texto cualquiera". No obstante, "su descripción permite ya, en cierta medida, evaluar comparativa, cualitativa y cuantitativamente, el funcionamiento enunciativo de dos ítems, secuencias, unidades o conjuntos textuales determinados".

El desarrollo de una metodología que permita su identificación, por tanto, supone un acercamiento a la sistematización de aspectos de la lengua muy relevantes en el terreno del Procesamiento del Lenguaje Natural (PLN) en tareas tales como análisis de sentimientos y minería de opiniones (véanse los traba-

jos de [18,19,12]), procesamiento de lenguaje figurado ([15,3,8]), o persuasión automática ([9,7,16]). No obstante, desarrollar dicha metodología implica un proceso totalmente objetivo que permita el reconocimiento de patrones que puedan generalizarse, sobre todo, si se piensa en la automatización del proceso de identificación. En este sentido, a continuación se describe la metodología que se utilizó para la identificación y extracción de subjetivemas.

3.1. Extracción de subjetivemas

De acuerdo con lo descrito en las secciones previas, los subjetivemas pueden pertenecer a varias categorías gramaticales; por tanto, el primer paso para su extracción consistió en determinar una categoría de prueba que nos permitiera i) probar la eficacia de la metodología para identificar y extraer subjetivemas, y ii) evaluar de manera empírica que los subjetivemas extraídos funcionan realmente como catalizadores de información subjetiva. En este sentido, se optó por trabajar con la categoría **adjetivo**. Esta elección está en relación con el hecho de que los adjetivos son las piezas léxicas valorativas por excelencia y, en consecuencia, las que mayor carga de subjetividad pueden imprimir a los enunciados. Además, como apunta [2], se trata de una categoría semántica, puesto que hay significados exclusivos del dominio de los adjetivos, por ejemplo, relaciones de diferentes tipos, así como valoraciones y apreciaciones que precisan, explican, caracterizan, clasifican o subvierten el significado básico del sustantivo al que caracterizan; i.e. los adjetivos dan cuenta de propiedades semánticas que pueden ser aplicadas a todos los constituyentes de nuestra realidad. Ahora bien, de las diferentes clases de adjetivos que existen en nuestra lengua (calificativos, determinativos, subsectivos, intersectivos, etc.), se consideró solamente la de adjetivos calificativos puesto que esta clase contiene adjetivos subjetivos y no subjetivos relacionados con la caracterización de atributos humanos.

Una vez establecido que se trabajaría con adjetivos calificativos, el siguiente paso consistió en extraer un conjunto de adjetivos que, de acuerdo con información lexicográfica, pudieran ser considerados como subjetivemas[2]. Para ello, se seleccionaron dos diccionarios como base de conocimiento: el DRAE y el Diccionario [11]. En cada diccionario se realizó una búsqueda a partir de lo que en lexicografía se conoce como marcas, es decir, etiquetas que declaran los atributos de la palabra desde diferentes puntos de vista (categoría gramatical, etimología, extensión geográfica de uso, etc.). Para este trabajo nos centramos

[2] De acuerdo con los registros del DRAE, el español cuenta con más de 170 mil acepciones correspondientes a adjetivos y sustantivos con usos adjetivales. Si consideramos que estos registros corresponden fundamentalmente al habla peninsular y que la lengua es un sistema dinámico, donde interviene el factor del tiempo y el desplazamiento entre categorías gramaticales, podemos suponer que este número es una referencia conservadora. Por tanto, el hecho de focalizar el concepto de subjetivema en un conjunto nuclear de adjetivos calificativos responde al hecho de que es simplista señalar que todas esas acepciones y, en consecuencia, el resto de elementos léxicos de una lengua, son subjetivos.

en las marcas que se refieren a particularidades sociales de uso. A continuación se sintetizan los pasos para la extracción:

a) **DRAE**: Se realizó una búsqueda a partir de los criterios *adjetivo* (marca gramatical) y *despectivo* (marca pragmática). El resultado fue un conjunto de 222 adjetivos.
b) **DIRAE**: Es un desarrollo de Gabriel Rodríguez que aprovecha toda la información del DRAE para realizar búsquedas inversas[3]. En este versión se efectuó una búsqueda con los criterios *adjetivo* (marca gramatical) y *dicho de una persona* (marca de uso). El resultado fue un conjunto de 1644 adjetivos.
c) **LEMA**: Para este diccionario se hizo una búsqueda utilizando los criterios de *adjetivo* (marca gramatical) y *despectivo, malsonante u ofensivo* (marca pragmática). El listado resultante fue de 51 adjetivos.

Efectuado este proceso de búsqueda, el total de candidatos a subjetivemas fue de 1,197 adjetivos. De este conjunto se hizo una depuración al desestimar adjetivos repetidos y adjetivos que se emplean para calificar situaciones u objetos y no a personas; así como a la supresión de los adjetivos conocidos como intensionales (verdadero, certero, falso, supuesto, principal) debido a que se refieren a la modalidad o toma de posición del hablante frente al referente de su discurso, lo que en la teoría de la enunciación se corresponde con a la subjetividad modal (véase 2.1).El conjunto final de candidatos a subjetivemas fue de 752 adjetivos. Algunos ejemplos se enlistan en la Tabla 1.

Tabla 1. Ejemplo de candidatos a subjetivemas.

DRAE	DIRAE	LEMA
absuelto	santurrón	solterona
naco	paternal	subversivo
discapacitado	puritano	deschavetado
morboso	tacaño	incapaz
impúdico	caritativo	calenturiento
malhablado	humanitario	soplón

4. Evaluación

A partir del conjunto de 752 adjetivos, el siguiente paso consistió en evaluar qué tan bien funcionaban como subjetivemas. Esto implicó realizar una tarea de reconocimiento automático de subjetividad en texto. Para ello, se creó un corpus de artículos periodísticos que fueron representados en términos de presencia/ausencia de subjetivemas con el fin de recuperar los textos que más

[3] Disponible en http://dirae.es.

probabilidad tuvieran de ser subjetivos. Posteriormente, los resultados fueron validados en una tarea de reconocimiento de subjetividad realizada por lectores humanos. A continuación se detallan las tareas realizadas.

4.1. Corpus

El corpus de trabajo se creó a partir de la recolección de artículos periodísticos en español. Se optó por considerar el lenguaje periodístico debido a que los comunicadores no pueden sustraerse de la visión y exposición parcial sobre los hechos que documentan, propiciando así la presencia de elementos subjetivos. En este sentido, siguiendo a [4], "el sujeto informa, y en este mismo acto interpreta y transforma desde su visión del mundo la realidad social y concreta que es objeto de la información; y es esa misma realidad y la relación entre los sujetos la que condiciona dicha visión del mundo". Aunado a lo anterior, la prensa es muy apreciada como fuente de información léxica, en tanto que puede concebirse como poseedora de repertorios lingüísticos con actualización diaria que regulan las comunicaciones del sector de la sociedad a la que se destina cada publicación.

Ahora bien, en tanto que nuestro interés por el reconocimiento automático de subjetividad en texto tiene por propósito futuro confrontar el léxico subjetivo frente al objetivo, el corpus se conformó con textos que, en términos de géneros periodísticos, fueran representativos de ambos tipos de léxico: la **nota informativa** (como paradigma de texto objetivo puesto que pertenece al género expositivo) y la **columna de opinión** (que supone ejemplos prototípicos de textos subjetivos puesto que se clasifica dentro del género argumentativo).

Un total de 40 artículos constituyen el corpus: 20 corresponden a notas informativas y 20 a columnas de opinión. Los textos fueron recuperados automáticamente de los periódicos mexicanos El Universal y La Jornada en sus versiones online. Con el fin de asegurar una evaluación objetiva, se decidió que todos los artículos tuvieran como eje temático la misma noticia: el conflicto magisterial en México entre 2013 y 2014. Ningún otro criterio fue tomado en cuenta. En la Tabla 2 se presentan algunos datos estadísticos del corpus.

Tabla 2. Estadísticas generales del corpus.

	Nota informativa	Columna de opinión	Total
Textos Universal	10	10	20
Textos Jornada	10	10	20
Types	3,621	3,810	*6,103*
Tokens	15,729	15,646	31,375
Oraciones	530	562	1,092

4.2. Recuperación automática de textos subjetivos

El proceso de recuperación automática de textos subjetivos consistió en identificar qué textos tenían mayor probabilidad de ser subjetivos, independientemente del género al que pertenecieran (nota informativa o columna de opinión). La tarea, en este sentido, se realizó por medio de la caracterización de los textos del corpus por medio de los subjetivemas extraídos del DRAE y del diccionario Lema. En primer lugar, tanto el corpus como los subjetivemas fueron lematizados con el fin de que las marcas morfológicas de género, número y apreciativo no incidieran ni sesgaran la caracterización. Para ello, se utilizó el algoritmo de [14] en su adaptación al español. En segundo lugar, se eliminaron todas las stopwords. Finalmente, cada texto del corpus se caracterizó de acuerdo con la Fórmula 1:

$$subjetividad_{(t)} = \frac{\sum_{s_1}^{n}}{|t|} \tag{1}$$

donde s_1 hasta s_n es el conjunto de adjetivos que funcionan como subjetivemas y $|t|$ es la longitud de cada texto del corpus. Es importante destacar que en esta caracterización no importó si el subjetivema se repetía n veces o sólo aparecía una vez. Al final, todas las frecuencias se sumaron y normalizaron por la longitud del texto para asegurar que el tamaño no incidiera en el resultado. Para finalizar el proceso, los textos cuyo valor fue más alto se consideraron como subjetivos. Los resultados se describen en la Sección 5.

4.3. Validación humana

Con el fin de contrastar los resultados de la aproximación automática con las percepciones que los humanos tienen respecto a lo que se entiende por subjetividad, se realizó una validación piloto sobre los resultados. Para ello, se seleccionaron tres hablantes de español de México. Dos de ellos con formación académica de licenciatura y uno más de maestría. Todos hablantes nativos de español e informados sobre el conflicto magisterial. A cada uno se le dieron 2 textos: el más subjetivo y el menos subjetivo, según los resultados del proceso automático. Una columna de opinión y una nota informativa, respectivamente. Los títulos de cada artículo son *La cruzada moral contra el magisterio*, para el primero y *Nos tachan de revoltosos por defender la educación pública*, para el segundo.

La tarea para cada lector humano consistió en i) leer ambos textos y determinar cuál era subjetivo y cuál, objetivo; y ii) responder un cuestionario en el que aportaran los elementos que cada uno consideró para tomar su decisión respecto a lo que hicieron en i). A ninguno se le dijo qué debía entender por cada concepto ni se le informó acerca de la procedencia del artículo. Sólo se les dijo que decidieran en función de lo que ellos consideraban como subjetivo y objetivo.

Los resultados se presentan en la siguiente sección.

5. Resultados

A pesar de que el conjunto de subjetivemas que se extrajo parecía muy pequeño en tanto las muchas posibilidades que tiene una lengua para expresar información subjetiva (sólo 752 adjetivos), los resultados obtenidos con el proceso automático muestran datos interesantes. Así, de los 40 textos que integran el corpus de trabajo, todos ellos, según lo observado, contienen elementos con carga de información subjetiva. Obviamente, la presencia de subjetivemas no es constante en todos los textos. Por ejemplo, hay textos con alrededor de 900 tokens, de los cuales más de 80 corresponden a subjetivemas; i.e. casi el 10 % del contenido es subjetivo. Por el contrario, hay otros textos que en ese mismo promedio de tokens contienen menos de 10 subjetivemas.

Como síntesis del proceso automático es importante destacar que alrededor del 42 % de los textos del corpus (i.e. 17) tiene contenido suficiente para ser considerado como texto subjetivo. Esto se deduce del porcentaje de subjetividad que se obtiene al aplicar la Fórmula 1. Así, resulta que si definimos un umbral mínimo de subjetividad $\geq .3$, entonces 17 textos caen en la categoría de texto subjetivo dado su alto contenido de subjetivemas.[4] Ahora bien, cabe destacar que de estos 17 textos, 5 pertenecen a la categoría nota informativa y los 12 restantes a la columna de opinión (claro está, que estos últimos ocupan los primeros lugares de subjetividad).

En relación con la validación humana, en la Tabla 3 se sintetizan los resultados que arrojó el análisis de los lectores humanos. De acuerdo con estos datos, se nota que entre los humanos no hay consenso respecto a lo que cada uno percibe como subjetivo u objetivo. Para la columna de opinión, que en esencia supone la presencia de valoraciones personales muy claras que lo convierten en subjetivo, dos de los lectores lo consideran como un texto objetivo; mientras que la nota informativa, que supone lo contrario, es categorizada por dos lectores como texto subjetivo.

Tabla 3. Resultados de la validación humana.

Texto	Modelo	Lector 1	Lector 2	Lector 3
Opinión	*Subjetivo*	Subjetivo	Objetivo	Objetivo
Nota	*Objetivo*	Subjetivo	Objetivo	Subjetivo

5.1. Discusión

Para entender este comportamiento, se analizaron las evidencias que aportó cada lector en torno al porqué de sus decisiones. Así, para la pregunta explícita

[4] Este umbral se definió tomando en cuenta que lo único que reconoce el modelo, por ahora, son adjetivos. Por tanto, pensar en un umbral mayor supondría un texto incoherente que en su mayoría debería estar escrito con adjetivos.

sobre qué palabras del texto justifican la subjetividad u objetividad, estos indicaron algunos adjetivos que se correspondían con los que nosotros habíamos detectado como subjetivemas (disidente, revoltoso, indolente), además de otros que no se habían considerado (estandarizado, enraizado, gris); pero para la pregunta sobre lo que consideraban como recursos empleados por el periodista para mostrar dicha tendencia, los tres respondieron con argumentos que denotaban que su proceso de análisis no estuvo exento de valoraciones subjetivas. Por ejemplo, el lector 2 señaló que valoró la columna de opinión como objetiva porque "*el autor trata de explicar el origen y las razones de la protesta para mostrar su parte racional que es escondida en el vituperio y en el clima de crispación mediática que se ha creado interesadamente contra el magisterio y su protesta*"; en tanto, el lector 1 indicó que valoró la nota informativa como subjetiva porque "*hace ver a los maestros como mártires de la gente que está cansada de los inconvenientes que provocan sus marchas*".

Esta falta de acuerdo entre los humanos, además de demostrar lo complejo de la tarea, pone en evidencia los siguientes puntos:

1. El proceso para la selección de subjetivemas responde de manera adecuada al proceso de representación de subjetividad en texto. No obstante, es importante explorar con otras categorías y con otro tipo de recursos que puedan integrar menos información léxica y más de tipo pragmática. Al respecto, hay un dato muy importante que sustenta esto: la información de la Tabla 2, sobre todo la relativa a los types, da indicios de que tanto notas como opiniones comparten bastante léxico. Así, lo que propicia que un texto se perciba como subjetivo está en relación con factores tales como la estructura sintáctica o la percepción, totalmente pragmática, que los humanos tienen sobre un hecho o periódico determinado.
2. La forma de representar la subjetividad en texto es operativa pero tiende a ser muy general. Si bien partimos de elementos cuantitativos, hay que perfeccionar los elementos cualitativos. Por ejemplo, la presencia de un solo subjetivema puede ser suficiente para hablar de un texto subjetivo.
3. Los textos que elegimos para probar los subjetivemas demuestran que para implementar un modelo de subjetivemas en escenarios reales es necesario partir de información formal, como los géneros periodísticos, que permita entender cómo funciona la transmisión explícita e implícita de valoraciones subjetivas. En este sentido, por ejemplo, nuestra intuición apuntaba a que serían los textos pertenecientes a la columna de opinión los que, dadas sus características, tenderían a ser los únicos subjetivos. No obstante, es interesante notar que, tanto en el proceso automático como en el humano, esta intuición no se corresponde totalmente con la realidad.
4. Aunado al punto anterior, es claro que las evidencias de subjetividad en los textos dependen del género discursivo de que se trate. Así pues, el discurso expositivo se vale de un léxico técnico o del uso de citas textuales para dar credibilidad, igualmente, recurre a la tercera persona como marca implícita de distanciamiento del hecho y, en consecuencia, de objetividad. En este

sentido, es importante considerar información tipográfica y sintáctica para complementar un modelo de reconocimiento de textos subjetivos.

5. Es indudable que los humanos tienen que participar en la evaluación de todo sistema automático; sin embargo, a partir de nuestros resultados nos queda claro que debemos explorar más opciones para que la valoración humana sea realmente efectiva. Por ejemplo, probar con lectores que desconozcan la noticia para no sesgar su análisis. Además, claro está, de tener un conjunto más amplio de evaluadores y de textos, posiblemente más cortos, que permitan perfeccionar el modelo.

Finalmente, es importante destacar que un modelo que permita reconocer documentos subjetivos basado en información fina, como es el caso de los subjetivemas, permitirá enfrentar problemas relacionados con el PLN desde un ángulo más complejo y abarcador.

6. Conclusiones y trabajo futuro

En este artículo se ha presentado una aproximación al procesamiento automático de subjetividad en texto. A partir del concepto de subjetivema se ha tratado de representar el proceso que los hablantes, a través de las palabras, realizan para emitir valoraciones que pueden clasificarse como subjetivas. En este sentido, se ha descrito una metodología para reconocer y extraer subjetivemas por medio de patrones lexicográficos. En particular, la metodología se probó con la categoría gramatical adjetivo. La relevancia del conjunto de subjetivemas fue evaluado en términos de una tarea de recuperación automática de textos subjetivos. Para ello, se creó un corpus de artículos periodísticos constituido por textos considerados como subjetivos y objetivos. Los resultados de la tarea, además, fueron valorados por tres lectores humanos.

Las conclusiones a las que llegamos apuntan a que el método de representación de subjetividad en texto a partir de subjetivemas es operativo. No obstante, para crear un modelo más robusto, es necesario considerar sustantivos y verbos, expandir la base de adjetivos con sinónimos y, posiblemente, antónimos, así como elementos tipográficos e información sintáctica con el fin de asegurar que el modelo no recaiga exclusivamente en información léxica sino que, además, incorpore información pragmática.

Estos aspectos serán abordados a detalle en trabajos futuros. Asimismo, dado nuestro interés por tareas y aplicaciones a las que subyazca el tratamiento de contenido subjetivo, será interesante estudiar el reconocimiento automático de estrategias pragmáticas complejas que se manifiestan en el nivel discursivo; tales como la identificación de usos tendenciosos del lenguaje o el reconocimiento automático de léxico persuasivo.

Referencias

1. Benveniste, E.: De la subjetividad en el lenguaje. In: Problemas de lingüística general I, pp. 179–187 (1982)

2. Catena, A.: Contribución a la formalización del adjetivo para la traducción automática español-francés. Ph.D. thesis, Universitat Autònoma de Barcelona (2006)
3. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning. pp. 107–116. CoNLL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
4. Espeche, C.: Periodismo objetivo o subjetivo. una falsa dicotomía. In: Congreso de periodismo y medios de comunicación. Buenos Aires, Argentina (2012)
5. Fillmore, C.: Lectures on deixis (1997)
6. Fournet, A.: Michel Bréal (1832–1915), a forgotten precursor of enunciation and subjectivity. ReVEL 9(16) (2011)
7. Gatti, L., Guerini, M., Stock, O., Strapparava, C.: Sentiment variations in text for persuasion technology. In: Proceedings of the 9th International Conference on Persuasive Technology (PERSUASIVE 2014). Padua, Italy (2014)
8. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in Twitter: A closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Vol. 2. pp. 581–586. The Association for Computer Linguistics (2011)
9. Guerini, M., Ozbal, G., Strapparava, C.: Echoes of persuasion: The effect of euphony in persuasive communication. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2015). Denver, Colorado (2015)
10. Kerbrat-Orecchioni, C.: L'énonciation. De la subjectivité dans le langage. Armand Colin, Paris (1980)
11. Lema: Diccionario de la lengua española. versión en cd room (2001)
12. Liu, B.: Sentiment analysis and subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group (2010)
13. Martin, J., White, P.: The Language of Evaluation. Appraisal in English. Palgrave Macmillan, London (2005)
14. Porter, M.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
15. Reyes, A., Rosso, P., Veale, T.: A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation 47(1), 239–268 (2013)
16. Strapparava, C., Guerini, M., Stock, O.: Predicting persuasiveness in political discourses. In: Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta (2010)
17. White, P.: Un recorrido por la teoría de la valoración (teoría de la valoración) (2000)
18. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics 35(3), 1–35 (2009)
19. Zhang, L., Barnden, J., Hendley, R., Wallington, A.: Exploitation in affect detection in open-ended improvisational text. In: SST '06: Proceedings of the Workshop on Sentiment and Subjectivity in Text. pp. 47–54. Association for Computational Linguistics (2006)

Plataforma web para la identificación y el análisis de eventos en Twitter

Antonio Juárez-González[1], Griselda Velázquez-Villar[2], Esaú Villatoro-Tello[2], Gabriela Ramírez-de-la-Rosa[2]

[1] Universidad Politécnica de Tlaxcala,
Tecnologías de la Información,
Tlaxcala, México

[2] Universidad Autónoma Metropolitana Unidad Cuajimalpa,
Departamento de Tecnologías de la Información,
México, D.F.

antonio.juarez@uptlax.edu.mx, grisvillar@yahoo.com.mx,
{evillatoro,gramirez}@correo.cua.uam.mx

Resumen. Debido a la gran popularidad que han adquirido actualmente las redes sociales entre personas, empresas, figuras públicas etc., surge la necesidad de contar con métodos automáticos que faciliten la búsqueda, recuperación y análisis de grandes cantidades de información. Ante esto, el Analista de Reputación en Línea (ARL) enfrenta el reto de identificar temas relevantes alrededor de un evento, producto y/o figura pública; a partir de lo cual puede proponer diferentes estrategias para fortalecer y/o revertir tendencias. Por lo tanto, en este trabajo se propone y describe una herramienta web que tiene como objetivo principal apoyar en las tareas desempeñadas por un ARL. Las técnicas de visualización propuestas permiten identificar de manera inmediata la relevancia y el alcance de las opiniones generadas sobre un evento sucedido en Twitter.

Palabras clave: Agrupamiento, medidas de similitud, visualización de información.

1. Introducción

El surgimiento de las redes sociales en Internet han propiciado que un mayor número de personas tenga la posibilidad de publicar libremente opiniones y comentarios acerca de una gran variedad de temas sociales, culturales, deportivos, científicos e incluso opiniones sobre productos y servicios.

Gracias a la popularidad que han adquirido estas redes sociales, actualmente es de gran interés para muchas entidades conocer lo que se dice de ellas dentro de este mundo digital. Al mismo tiempo, a través de estos medios de comunicación, es posible tener un acercamiento con distintos usuarios, mismos que aprovechan esta interacción para dar a conocer de forma específica su opinión sobre determinados temas, productos o servicios. Una de las redes sociales más

utilizadas para este fin es Twitter que permite enviar mensajes cortos llamados tuits (*tweets* en Inglés), con una longitud máxima de 140 caracteres[1]. De acuerdo a Statisticbrain[2] a enero del 2014, el número total de usuarios activos en Twitter asciende a 645,750,000 a nivel mundial y el número de tuits al día es de 58 millones. Por su parte, México ocupa el séptimo lugar entre los países más tuiteros del mundo con casi 15 millones de usuarios.

Con la finalidad de aprovechar la gran cantidad de información obtenida de las interacciones (directas e indirectas) entre usuarios y empresas, estas últimas han creado la figura de un Analista de Reputación en Línea (ARL). El trabajo de este profesional pasa por tres fases: la primera fase consiste en el monitoreo, éste permite conocer en todo momento lo que se está publicando con relación a la empresa, producto o figura pública de interés. La segunda fase es la identificación de temas relevantes dentro de la comunidad de usuarios de Twitter, priorizando los mensajes con las implicaciones más importantes, negativas o positivas, hacia la entidad en cuestión. Finalmente, la tercera fase consiste en proponer estrategias de mercado que permitan revertir los efectos negativos previamente identificados o incluso fortalecer los aspectos positivos de la entidad en revisión.

Debido a que las publicaciones crecen de manera acelerada, el análisis manual de esta información resulta complicado y desgastante para el ARL. En consecuencia, surge la necesidad de contar con sistemas automáticos que permitan realizar este análisis de forma más sencilla y oportuna. Recientemente diversos grupos de investigación están interesados en esta problemática [1] y se han dado a la tarea de desarrollar sistemas enfocados al análisis de opiniones generadas en Twitter. En este contexto surge Replab[3] como un foro internacional en el cual se han propuestos y evaluado distintos sistemas automáticos enfocados en el análisis de la reputación en línea, específicamente de la información producida en Twitter.

Hasta el momento, los diferentes grupos que han participado en Replab, se han enfocado en desarrollar métodos automáticos para tareas como: *i)* selección de tuits relevantes para una entidad[4] [4], *ii)* identificación de implicaciones negativas, positivas o neutrales hacia una entidad [4], *iii)* agrupamiento de opiniones por temática similar [4,3], y *iv)* la identificación de líderes de opinión dentro de una comunidad [8]. Sin embrago, el problema de cómo mostrar el resultado del análisis automático a un ARL de forma que se le facilite la toma de decisiones, ha sido poco explorado. Es por esto, que surge la necesidad de desarrollar sistemas que aprovechen los resultados de estos métodos automáticos y permitan generar representaciones visuales.

A partir de lo anterior, dentro de este trabajo se explora una alternativa de visualización de los resultados producidos por sistemas de análisis de contenido en Twitter, en particular sistemas desarrollados en el marco de la competencia de RepLab. El sistema propuesto se enfoca específicamente en la visualización

[1] http://about.twitter.com/
[2] http://www.statisticbrain.com/twitter-statistics/
[3] http://www.limosine-project.eu/events/
[4] Por *entidad* nos referimos al nombre de una figura pública y/o el de una organización.

de resultados del agrupamiento de opiniones por temática similar y la forma en cómo éstos se relacionan entre sí.

El resto del artículo se encuentra organizado de la siguiente manera. La sección 2 describe algunas de las plataformas existentes más cercanas a los objetivos del presente trabajo. La sección 3 describe detalladamente la arquitectura del sistema desarrollado. Posteriormente, la sección 4 muestra el funcionamiento y principales características de la plataforma web. Finalmente en la sección 5 se mencionan las principales conclusiones e ideas de trabajo futuro derivadas del presente proyecto.

2. Trabajo relacionado

Actualmente, existen disponibles en Internet variadas aplicaciones enfocadas al agrupamiento, análisis y visualización de información. Estas aplicaciones las podemos dividir en dos grandes categorías: *i)* herramientas especializadas en el agrupamiento y clasificación de grandes volúmenes de datos; y *ii)* herramientas no especializadas, que permiten realizar análisis cualitativo de datos a usuarios estándar.

Por un lado, entre las herramientas especializadas vale la pena mencionar que tienden a emplear un lenguaje muy técnico y como consecuencia se dificulta la interpretación de los resultados para usuarios no especialistas. Como ejemplo de éste tipo de herramientas podemos mencionar a Cluto [7] y Weka [6], herramientas multi-plataforma que tienen implementados gran variedad de métodos automáticos para el análisis de datos. Por otro lado, en la segunda categoría están las herramientas que se dedican al análisis de datos que se publican exclusivamente en redes sociales, y que buscan proporcionar a usuarios, expertos e inexpertos, con los elementos suficientes para realizar un análisis fácil e intuitivo de los resultados proporcionados por sus métodos de identificación de temáticas, polaridad, etc. Como ejemplo de tales herramientas podemos mencionar a Spot[5], AnaliticPro[6] y Socialmention*[7]; las cuales proporcionan al usuario variados esquemas de visualización de datos, los cuales buscan resaltar ciertos indicadores que permiten al analista evaluar y determinar la reputación que tiene un producto o tema en particular dentro de una comunidad específica.
Spot. Es una aplicación que permite la visualización interactiva de lo que se está publicando en Twitter en tiempo real. La idea principal es mostrar rápidamente las opiniones que se generan sobre un tema en particular. La forma de presentar los tuits, es a través de grupos contenidos en burbujas, las cuales se organizan y visualizan de diferentes maneras para resaltar distintos tipos de información sobre el tema de interés. Al realizar la búsqueda del tema, solo se recuperaran los últimos 200 tuits para la visualización. Hay que tomar en cuenta que los resultados de búsqueda de Twitter sólo se remontan alrededor de una semana.

[5] `http://neoformix.com/2012/IntroducingSpot.html`
[6] `http://www.analiticpro.cl/caracteristicas.php`
[7] `http://www.socialmention.com`

Por lo que la búsqueda y el análisis sólo se puede realizar sobre un conjunto muy limitado de tuits.

AnaliticPro. Aplicación que permite procesar grandes volúmenes de información producida en distintas redes sociales. Se pueden realizar mediciones con criterios personalizados, además de esto incorpora técnicas que permiten identificar el sentido (*i.e.,* positivo, negativo o neutral) de los comentarios, se pueden relacionar y construir frases para conocer la opinión generalizada dentro de una comunidad respecto a uno o varios temas. Una de las principales desventajas de esta herramienta es que para la construcción de sus informes se basa en técnicas semi-automáticas, es decir, hay informes que son generados con ayuda de expertos. Por tal motivo, para poder explotar al 100 % la infraestructura ofrecida por AnaliticPro se requiere del pago de licencias y/o servicios de análisis de reputación.

Socialmention*. Es una aplicación que monitorea y analiza la información que se está generando en distintas redes sociales de Internet en tiempo real. Además, permite seguir y medir fácilmente lo que se opina sobre alguna persona, empresa, producto, etc. Las búsquedas se realizan en más de 80 medios sociales incluyendo las más visitadas como son: Twitter, Facebook, friendFeed, YouTube, Digg, Google, etc.. A diferencia de las herramientas previas, Socialmention* propone diferentes medidas que facilitan la interpretación de los datos analizados, por ejemplo: *fuerza, sentimientos, pasión y alcance.* Intuitivamente, estas medidas proporcionan al usuario una idea de la importancia y el alcance del tema en revisión.

En general, las herramientas mencionadas anteriormente proponen distintos métodos para el análisis y visualización la información producida en redes sociales respecto a una entidad específica. Principalmente se han enfocado en identificar la polaridad de los comentarios, el origen de los mismos (*i.e.,* red social, dispositivo, tipo de usuario), y las posibles tendencias. Al contrario de las herramientas analizadas, el trabajo desarrollado en este artículo busca proporcionar herramientas que faciliten la identificación de temáticas relevantes y al mismo tiempo la relación que éstas pueden tener con otros aspectos secundarios sucedidos al rededor del mismo evento, aspecto que no es considerado por ninguna de las herramientas revisadas. De esta forma, la herramienta propuesta permitirá al ARL identificar de manera inmediata tanto los temas y sub-temas que suceden al rededor de un evento, así como la relevancia de los mismos. En las siguientes secciones se describe en más detalle el sistema desarrollado.

3. Sistema propuesto

El sistema propuesto se compone de tres grandes módulos, de los cuales el primero se encarga de la búsqueda y recuperación de tuits, posteriormente se hace un proceso de agrupamiento, el cual puede ser temático o no-temático, y finalmente se produce una salida gráfica la cual es mostrada al usuario. En las sub-secciones siguientes describimos los componentes principales de cada uno de estos módulos.

3.1. Recuperación de tuits

La recuperación de los tuits comienza con una consulta,en donde se especifica el tema a buscar. La consulta consiste en una cadena de caracteres que pueden ser uno o varios términos (*i.e.,* consultas compuestas), y el número de tuits que se desean recuperar. Para poder realizar este proceso se utilizó la librería twitter4j[8], la cual permite la conexión con la plataforma de Twitter. Es importante mencionar que para poder usar esta librería es necesario realizar un registro en la página de desarrolladores de Twitter. Este proceso permite la obtención de las llaves de acceso que permiten a sistemas automáticos hacer uso de la información que se genera en esta red social.

Así entonces, si la conexión a Twitter es exitosa, se recuperan los tuits, se almacenan para procesos posteriores, y además se muestran en la interfaz gráfica del sistema. Si en determinado momento sucede algún error de conexión, el usuario es notificado. Es importante mencionar que por medio de almacenar los tuits en una base de datos se permite a los usuarios acceder a su historial de búsquedas, lo cual es una opción importante para los ARL, pues permite analizar tendencias y/o hacer comparaciones de forma inmediata. Otro aspecto importante del proceso de recuperación de tuits es que está diseñado para obtener la mayor cantidad de meta-datos relacionados a cada tuit descargado, por ejemplo: el nombre de usuario, imágenes compartidas, información del perfil de usuario, fecha, hora, plataforma, idioma, etc.

Nótese que un paso previo al almacenamiento de los tuits en la base de datos es un módulo de *preprocesamiento*. Este paso es común en muchas tareas de procesamiento de lenguaje natural, y tiene como principal objetivo eliminar información que se considera sin carga temática. Para este caso se consideraron las siguientes operaciones de preprocesamiento: el texto es llevado a minúsculas, se eliminan símbolos de puntuación, se eliminan las URLs y se quitan palabras funcionales[9]. Finalmente, los tuits originales y pre-procesados quedan almacenados en la base de datos, listos para ser procesados por el módulo de agrupamiento.

3.2. Agrupamiento temático

Un primer paso necesario para realizar a tarea de agrupamiento temático es el *indexado* de los documentos a analizar, actividad que denota hacer el mapeo de un documento d_j en una forma compacta de su contenido. La representación más comúnmente utilizada es un vector con términos[10] ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información [2]. Es decir, un texto d_j es representado como el vector $\overrightarrow{d}_j = \langle w_{kj}, \ldots, w_{|\tau|j} \rangle$, donde τ es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento, mientras que w_{kj} representa la importancia del término t_k dentro del contenido del documento d_j.

[8] `http://twitter4j.org/en/index.html`

[9] También conocidas como palabras vacías o *stopwords* en Inglés

[10] Entiéndase por términos ya sea palabras y/o n-gramas de palabras o caracteres.

Esta técnica, conocida como bolsa de palabras (BOW[11]), es la forma tradicionalmente utilizada para representar los documentos [5]. Dentro de la herramienta desarrollada se consideraron sólo palabras simples como los elementos del vector. El peso w_{kj} puede ser calculado utilizando diferentes enfoques, el más simple de estos enfoques es el *booleano* que consiste en asignar un valor de 1 al término si éste aparece en el documento, y 0 en caso contrario. Agregado a éste, otros esquemas de pesado muy comunes son el conocido como frecuencia (*TF*) y frecuencia relativa (*TF-IDF*) [9]. Es conveniente mencionar que la herramienta desarrollada incluye estos tres esquemas de pesado.

Una vez que se tiene una representación apropiada de los documentos (*i.e.,* tuits) podemos proceder con el proceso de agrupamiento. Los grupos deben cumplir una serie de propiedades, *e.g.,* los documentos pertenecientes al mismo grupo deben ser muy similares, mientras que al mismo tiempo los documentos que pertenecen a grupos distintos deben ser tan diferentes como sea posible; a estas propiedades se les conoce como *homogeneidad* y *heterogeneidad* respectivamente. En general, para lograr aproximarse a dichas propiedades, en necesario determinar similitudes entre los objetos a partir de los valores de sus atributos; para nuestro caso se utilizó la medida del coseno.

En el sistema descrito en este artículo se trabajó con dos distintas técnicas de agrupamiento, específicamente se implementó un algoritmo de partición (*k-means*) y un algoritmo jerárquico (*Hierarchical Clustering*).

Por un lado, los algoritmos de partición agrupan los elementos entorno a elementos centrales llamados *centroides*. El algoritmo de *k-means* es un método iterativo que tiene como parámetro importante el valor de k (el número de grupos a formar), a pesar de lo cual es un algoritmo muy efectivo [10]. Por otro lado, los algoritmos jerárquicos se caracterizan por generar una estructura de árbol, llamada dendograma, en la que cada nivel del árbol es un posible agrupamiento de los objetos de la colección. El método de *Hierarchical Clustering* es un algoritmo jerárquico de tipo aglomerativo, es decir parte de las hojas del árbol, considerando a cada elemento como un grupo. Posteriormente y de forma iterativa va uniendo elementos en grupos más cercanos hasta que todos los documentos se encuentran dentro de un grupo [10].

Es importante mencionar que en el sistema desarrollado se incorporaron las implementaciones hechas en Weka [6] de los algoritmos de agrupamiento descritos con sus configuraciones por defecto.

3.3. Visualización

Para la representación visual de los resultados de agrupamiento temático y no-temático, se hace uso de la librería D3js[12], específicamente del tipo de gráfica denominada Bubble Chart.

[11] Bag Of Words por sus siglas en Inglés.

[12] D3JS (Data-Driven Documents) es una librería JavaScript que permite manipular y visualizar distintos tipos de datos (`http://d3js.org`).

Cada burbuja de la gráfica representa un grupo, resultado del agrupamiento temático realizado en la etapa anterior. Además, el tamaño de la burbujas representa, hasta cierto punto, la importancia del sub-grupo identificado. Así entonces, dentro de las burbujas que representan a los diferentes sub-grupos se congregan los tuits que corresponden a ese grupo como una serie de burbujas más pequeña. Una de las ventajas de la visualización propuesta es que es posible, mediante el posicionamiento del curso sobre un tuit (*i.e.,* las burbujas más pequeñas), ver el contenido de éste. Adicionalmente, mediante esta gráfica es posible ver los n términos más representativos el grupo, *i.e.*, los términos más frecuentes en los tuits del grupo en cuestión.

Como se mencionó en secciones anteriores, una de las ventajas de nuestro sistema es que permite al ARL, además de identificar los diferentes temas y sub-temas que suceden alrededor de un evento dado, muestra las relaciones temáticas entre los distintos sub-temas. Para lograr esto hacemos los siguiente: *1)* se identifican los conceptos más representativos de cada sub-tema, *2)* para cada par de sub-grupos se buscan los conceptos contenidos en la intersección, y *3)* finalmente, los conceptos compartidos entre cada par de sub-grupos son mostrados al usuario en forma textual. Con esto, en analista puede identificar rápidamente conceptos (palabras) clave que la comunidad de usuarios está empleando para referiste al evento de interés.

Adicionalmente, el sistema desarrollado también permite generar gráficas con información extraída de los meta-datos de los tuits que se han recuperado. Particularmente, es posible construir gráficas agrupando los tuits por plataforma empleada para leer y escribir en Twitter, *e.g.*, Android, IOS, web, etc; por el número de favoritos, número de retuits. En conjunto, esta información resulta de utilidad para el ARL debido a que le permite identificar la relevancia y el alcance del evento que este siendo estudiado.

4. La plataforma en funcionamiento

Para ilustrar el módulo de visualización implementada en la plataforma propuesta, se realizó una búsqueda de tuits sobre el tema *Ayotzinapa*[13]. La consulta recuperó 3,000 tuits, número de tuits que se especificó mediante las opciones de la plataforma web. Con el objetivo de mostrar algunas de las características de la visualización, la Figura 1 muestra dos distintos resultados después de hacer un agrupamiento por temática similar. En la imagen de la izquierda se puede observar que el resultado del agrupamiento generó dos grupos, mientras que la imagen de la derecha se muestra una salida que resultó en tres sub-temas. Es importante recordar que uno de los parámetros requeridos por la plataforma es el valor de k, el cual indica la cantidad de sub-temas que queremos identificar. Intuitivamente, entre mayor sea el valor de k estaremos exigiendo mayor nivel de especialidad en los sub-temas generados, mientras que un valor muy pequeño permite mayor generalidad en los sub-temas.

[13] La consulta se realizó el día 26 de septiembre a las 12:27 horas, el día del aniversario de las desapariciones de 43 normalistas en el estado de Guerrero, México.

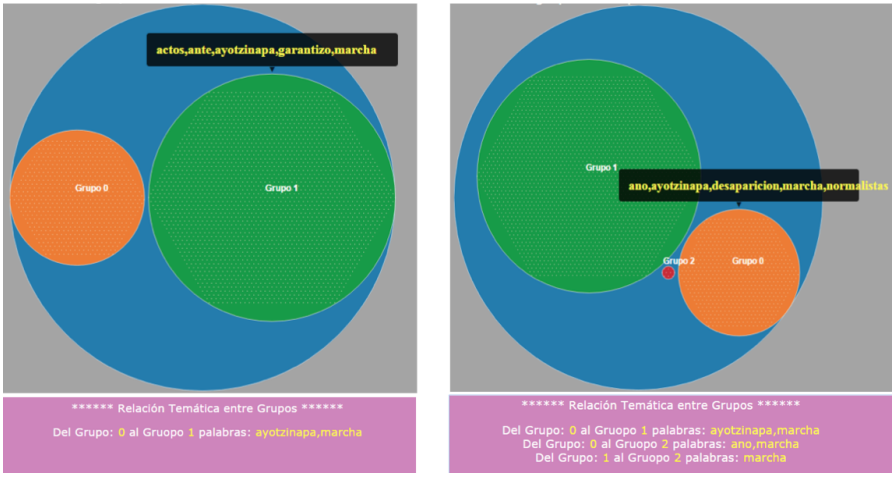

Fig. 1. Visualización del agrupamiento temático para 3000 tuits del tema Ayotzinapa. En la izquierda se muestra el resultado de generar dos grupos, mientras que en la derecha se muestra el resultado de la generación de tres grupos.

Como puede verse en la Figura 1, es posible obtener el conjunto de palabras más representativas de cada grupo, las cuales indican, hasta cierto punto, el contenido semántico de cada sub-grupo. En el caso de la imagen a la izquierda, el Grupo 1 se puede describir con las palabras *actos*, *ante*, *Ayotzinapa*, *garantizo*, y *marcha*, en contraste con el contenido semántico del Grupo 0 que se puede definir por las palabras *año*, *Ayotzinapa*, *desaparición*, *marcha*, *normalistas*. Ante este ejemplo, un ARL podría discernir que mientras todos los tuits de ambos grupos hablan sobre la marcha que se realizó sobre el caso Ayotzinapa, un subgrupo hace referencia al aniversario de la desaparición de normalistas en Ayotzinapa, mientras que el otro subgrupo hace mención sobre las garantías que se prometieron para los actos realizados en el contexto de la marcha. Este tipo de información podría fácilmente corroborase al posicionar el cursor sobre un tuit particular y ver su contenido.

Por otro lado en la imagen de la derecha de la Figura 1, el tercer grupo (Grupo 2) hace mención de los términos descriptivos: *año*, *marcha*, *México*, *normal*, *tragedia*; mientras que el grupo es mucho menor en relación a los grupos 0 y 1, es claro el sub-tema que éste describe, generalizando el problema a nivel país y describiendo el evento como una tragedia.

Agregado a lo anterior, en la Figura 1 también se pueden ver, de manera muy simple, las palabras que los grupos comparten. Por ejemplo, en el caso del agrupamiento de la imagen en la izquierda, dado que sólo existen dos grupos, los términos comunes entre ellos son *Ayotzinapa* y *marcha*. Con esta información sintetizada, el ARL podría tener un panorama general del tema en análisis.

Como se mencionó en la Sección 3.3, además de mostrar información sobre el agrupamiento temático, también es posible ver gráficas de los metadatos de

los tuits. En la Figura 2 se pueden observar las gráficas generadas por tipo de plataforma usada para enviar el tuit, por número de retuits y por número de favoritos (de izquierda a derecha en la Figura 2). De la gráfica que agrupa los tuits por plataforma de publicación podemos ver que de los 3000 tuits recuperados, 928 fueron enviados desde la aplicación de Twitter para Android; de la gráfica que agrupa los tuits por el número de retuits que éstos tienen, es posible ver que unos de los tuits mas retuiteados (436 veces) es un tuit que contiene las palabras temáticas "No dejarse engañar por la telenovela de PGR-Televisa, pide sobreviviente de Ayotzinapa"[14]. Finalmente, la gráfica que visualiza los tuits agrupados por favoritos muestra que el tuit marcado más veces como favorito es el video de la postura de la figura pública Fher (integrante del grupo musical Maná).

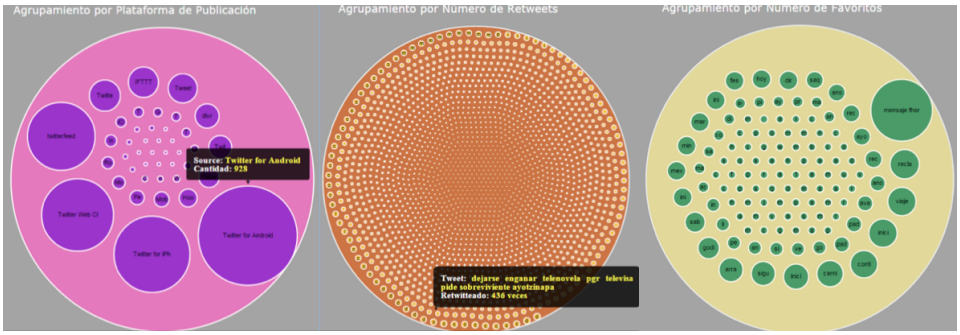

Fig. 2. Visualización por metadatos de una colección de tuits. De izquierda a derecha: agrupamiento por plataforma de publicación, agrupamiento por número de retuits, y agrupamiento por número de favoritos.

5. Conclusiones y trabajo futuro

En este artículo se describió el trabajo realizado para la construcción de una herramienta web diseñada para apoyar en las actividades desempeñadas por un ARL. Específicamente, la herramienta propuesta permite hacer la identificación automática de temas y sub-temas (y la relación temática entre éstos) sucedidos alrededor de un evento ocurrido en Twitter. Una de las ventajas ofrecidas por la aplicación desarrollada es que gracias a su propuesta de visualización de resultados, un analista puede, de manera sencilla e inmediata, identificar la relevancia y el alcance de las opiniones expresadas entorno a un evento de interés.

El uso de técnicas tradicionales de agrupamiento nos permitió definir una estrategia para lograr la identificación de sub-temas dentro de un conjunto de tuits. En particular, se emplearon dos tipos de algoritmos de agrupamiento, de

[14] `https://twitter.com/Revolucion3_0/status/645042141064925184`

partición y jerárquico, los cuales han mostrado ser efectivos en diversas tareas de agrupamiento de documentos. Como forma de representación de los textos se empleó la técnica conocida como bolsa de palabras (BOW), así como varios esquemas de pesado. En general, los métodos y técnicas empleadas son métodos muy intuitivos y en consecuencia fáciles de entender. Sin embargo, es necesario que el ARL conozca el significado de los parámetros que requieren estos métodos para poder hacer un uso eficiente de la herramienta desarrollada.

Como trabajo futuro se pretende adaptar la herramienta de manera que proporcione mayores facilidades a los usuarios para hacer compliación de corpus de tuiter.

Agradecimientos. Los autores agradecen a CONACyT por el apoyo otorgado a través del programa de redes temáticas (Red Temática de Tecnologías del Lenguaje, proyecto no. 260178). Adicionalmente, los autores agradecen también a UPTlax, UAM-C y SNI-CONACyT por todas las facilidades proporcionadas.

Referencias

1. Amigó, E., de Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M., Spina, D.: Overview of replab 2014: Author profiling and reputation dimensions for online reputation management. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings. pp. 307–322 (2014)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrival. Addison Wesley (1999)
3. Berrocal, J.L.A., Figuerola, C.G., Ángel Zazo Rodríguez: Reina at replab2013 topic detection task: Community detection. In: Proceedings of the Fourth International Conference of the CLEF initiative (2013)
4. Cossu, J.V., Bigot, B., Bonnefoy, L., Morchid, M., Bost, X., Senay, G., Dufour, R., Bouvier, V., Torres-Moreno, J.M., El-Beze, M.: Lia@replab 2013. In: Proceedings of the Fourth International Conference of the CLEF initiative (2013)
5. F., S.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (Nov 2009)
7. Karypis, G.: Cluto a clustering toolkit. Tech. Rep. Technical Report 02-017, Dept. of Computer Science, University of Minnesota (2002)
8. Ramírez-de-la Rosa, G., Villatoro-Tello, E., Jiménez-Salazar, H., Sánchez-Sánchez, C.: Towards automatic detection of user influence in twitter by means of stylistic and behavioral features. In: Gelbukh, A., Espinoza, F., Galicia-Haro, S. (eds.) Human-Inspired Computing and Its Applications, Lecture Notes in Computer Science, vol. 8856, pp. 245–256. Springer International Publishing (2014)
9. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5), 513–523 (Aug 1988)
10. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: In KDD Workshop on Text Mining (2000)

Anotación automática de datos acerca de la reputación de los políticos en redes sociales

Rocío Abascal-Mena[1], Jean-Valère Cossu[2], Alejandro Molina[3], Juan-Manuel Torres-Moreno[2]

[1] Universidad Autónoma Metropolitana, Unidad Cuajimalpa, México, D.F.

[2] Université d'Avignon et des Pays de Vaucluse, Laboratoire Informatique d'Avignon, Avignon, Francia

[3] Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), México, D.F.

rocio.abascal@gmail.com, alejandro.molina@etd.univ-avignon.fr, {jean-valere.cossu,juan-manuel.torres}@univ-avignon.fr

Resumen. Recientemente, la minería de opinión en Twitter, por medio de la Recuperación de Información (RI) y del Procesamiento de Lenguage Natural (PLN) atrajo interés en la investigación política. Sin embargo, la obtención de datos anotados de dominio específico sigue siendo un paso manual muy costoso. Además, la cantidad y calidad de estas anotaciones pueden ser críticas en relación al rendimiento de los métodos de PLN basados en técnicas de Aprendizaje de Máquina (AM). Una solución alternativa es utilizar conjuntos de datos cruzados por idioma (cross-language) y cruzados por dominio (cross-domain) para simular los datos de entrenamiento. Este artículo describe algunos enfoques de AM para la anotación automática de tuits en español que tratan acerca de la reputación de los políticos. Nuestra principal conclusión es que un clasificador estadístico sin entrenamiento específico del dominio político puede simular una anotación tan confiable como la provista por anotadores humanos. Este tipo de clasificadores, también puede resultar equivalente al uso de recursos específicos tales como lexicones de polaridad o datos de dominio traducidos manualmente.

Palabras clave: Procesamiento de lenguaje natural, aprendizaje de máquina, minería de opinión, análisis político automático.

1. Introducción

Los medios de comunicación modernos están cambiando nuestra visión sobre la sociedad en muchos aspectos. Las Ciencias Humanas, Sociales y Políticas deben evolucionar para tener todas las herramientas metodológicas con el fin de comprender las tendencias sociales y políticas tan rápido como lo requiere la sociedad moderna. Particularmente, Twitter[1] se ha utilizado no solo para hacer públicas las opiniones sobre

[1] http://www.twitter.com

diferentes eventos o personas, sino también para implicarse y participar en movimientos sociales.

Así, el uso de las redes sociales durante la campaña presidencial de 2012 en México adquirió mucha importancia y se convirtió el instrumento principal para ejercer la opinión pública, especialmente entre la gente joven. El movimiento juvenil yosoy132 nació durante esa campaña, por medio las redes sociales, permitiendo a jóvenes de diversas las universidades y condición social, compartir un tema tendencia (trending topic) común. En efecto, los estudios conducidos por el *Instituto Nacional de Estadística, Geografía e Informática* (INEGI) afirman que el 40.3 % de los usuarios de Tecnologías de la Comunicación (TIC) en México son jóvenes que se comunican vía las redes sociales y los télefonos móviles, y que permanecen conectados la mayor parte del día. Este porcentaje equivale a 15.3 millones de personas entre los 18 y 34 años de edad que son votantes potenciales [30]. La participación juvenil usando Twitter en México se ha incrementado creando significativas comunidades sociales y políticas alrededor de temas electorales. Por otra parte, los jóvenes mexicanos no son los únicos que envían tuits de manera masiva. Twitter, recientemente, ha captado una gran atención por parte de los principales candidatos que promueven su discurso en línea [27].

Como consecuencia, Twitter ofrece la oportunidad de recolectar, en tiempo real, grandes cantidades de datos, directamente de los usuarios; por lo que luego los tuits pueden ser analizados con el fin de realizar un seguimiento de las reacciones generadas en los distintos eventos. Debido a que Twitter ofrece la posibilidad de extraer tuits, a partir de los cuales se puede obtener un corpus, ha habido una gran cantidad de investigación lingüística aplicada en los tuits.

El uso de datos públicos en línea, para realizar estudios de sentimientos, reduce sigfinicativamente los costos, los esfuerzos y el tiempo necesario para administrar encuestas y cuestionarios de gran escala [5]. El Análisis de Sentimientos (AS) es un área útil en el estudio de la comunicación en línea debido a que le permite a los investigadores contar con la capacidad de medir automáticamente la emoción en textos en línea [31]. Los estudios políticos en español desde un punto de vista del Aprendizaje Automático o el Aprendizaje de Máquinas (AM) son todavía escasos [36]. Sin embargo, esto podría cambiar utilizando métodos de AM para simular las anotaciones humanas y ayudar a los expertos (en trabajos como [28]) a etiquetar una gran colección de datos.

Estudios usuales en el dominio asumen que se requiere un gran esfuerzo para la adquisición de tuits y su posterior etiquetado manual. Además, se necesita un proceso de validación para corregir los errores introducidos en el etiquetado manual. También es importante considerar que los acontecimientos políticos siempre ocurrirán mucho más rápido que nuestra capacidad para obtener los datos anotados manualmente en varios idiomas. En este contexto, proponemos un enfoque que puede proporcionar una pre-anotación confiable a partir de datos fuera del dominio que necesitan supervisión superficial antes de la validación con el fin de obtener un corpus fiable que se pueda usar para estudios políticos más complejos como la detección de la tendencia política de un usuario o el monitoreo de la reputación de un político.

El resto del artículo está organizado de la siguiente manera: la Sección 2 da una visión general de los trabajos relacionados y establece, además, la motivación de nuestro trabajo. En la Sección 3, ofrecemos detalles acerca de los conjuntos de datos utilizados.

En la Sección 4 proponemos nuestros enfoques mientras que la Sección 5 está dedicada a la evaluación. Finalmente, la Sección 6 presenta algunas conclusiones sobre nuestro trabajo y abre ciertas perspectivas de trabajo a futuro.

2. Estado del arte

2.1. Minería de tuits y análisis de sentimientos en contextos políticos

El análisis de datos políticos ha sido abordado en trabajos anteriores pero generalmente en inglés (véase [20] y [37]). En [24], se utiliza un léxicon subjetivo proveniente del Opinion Finder con el fin de determinar puntuaciones positivas y negativas para cada uno de los conjuntos de datos correspondientes a un tuit. En este caso, los números crudos de tuits positivos y negativos acerca de un tema dado son usados para calcular una puntuación de confianza (la relación entre el número de tuits positivos y negativos). Los autores indican que con una simple inspección manual de los tuits se pueden encontrar ejemplos que han sido clasificados incorrectamente. Sin embargo, los autores utilizan este método para medir la "confianza de los consumidores" (la aprobación presidencial en este caso) durante las elecciones presidenciales de 2008 en Estados Unidos.

Un enfoque diferente se empleó en [5] para analizar las preferencias políticas mediante el estudio del humor contenido en los tuits. Un instrumento psicométrico llamado Perfil de Estados de Ánimo (Profile of Mood States, POMS) se utilizó para destilar seis diferentes atributos emocionales: tensión, depresión, cólera, vigor, fatiga y confusión. Originalmente, POMS proporciona una lista de adjetivos para la cual un paciente tiene que indicar el nivel de aprobación. Cada adjetivo se relaciona con un estado de ánimo y, por lo tanto, la lista puede ser explotada como base para un "analizador de ánimo" a partir de datos textuales.

En [34] se presenta un trabajo en dos partes: en la primera se introduce el concepto de LIWC (Linguistic Inquiry and Word Count) que se utiliza para realizar un análisis superficial de los tuits relacionados con los diferentes partidos políticos que compitieron en la elección federal de Alemania en 2009. En la segunda parte, los autores afirman que el conteo de tuits con referencias a una de los partidos, refleja con precisión los resultados de las elecciones. También, se establece que el Error Absoluto Medio (Mean Absolute Error, MAE) de la "predicción" basada en los datos de Twitter estuvo muy cerca de las encuestas reales que se llevaron a cabo .

No obstante, también los métodos puramente estadísticos han tenido impacto en el tema. Un número creciente de análisis empíricos de sentimientos y humor basados en las colecciones de Twitter han sido utilizados junto con sofisticados algoritmos de preprocesamiento de texto, usando clasificadores basados en léxico, Máquinas de Soporte Vectorial (SVM) y métodos Bayesianos. La idea principal, de estos métodos, es la de entrenar un clasificador con las palabras clave de los tuits para determinar el estado de ánimo (ver [39,21]). También, estos métodos se han propuesto para la explotación de los tuits con el fin de detectar los cambios del estado de ánimo de la gente durante todo el día [21,19]. En [7], los autores midieron los cambios en el estado de ánimo de la población de Estados Unidos, durante tres años, a partir de tuits que proporcionaban indicadores relevantes acerca de la política.

Estudios más cualitativos proponen nuevas revelaciones sobre el comportamiento humano y como resultado se muestra que hay una enorme necesidad en el desarrollo de herramientas de minería de opinión acerca los medios sociales [22,8,11,14,18]. Sin embargo, la gran mayoría de estos trabajos utilizan corpus anotados en inglés para la experimentación, y hasta donde sabemos, no hay estudios en español ni en francés sobre el análisis de sentimientos políticos. Pero, ¿cómo hacer frente a la falta de conjuntos de datos para entrenar? En este artículo, se explora experimentalmente el uso de datos no específicos del dominio de la politica combinados con datos en otro idioma, lo que nos permite transmitir la experiencia aprendida de un clasificador de dominio en francés a otro en español.

2.2. Procesamiento de idiomas cruzados

La traducción de la información expresada en diferentes idiomas en una representación común es, en general, muy compleja. Sistemas de Recuperación de Información en Varios Idiomas (CLIR, por sus siglas en inglés) ayudan a recuperar documentos en diferentes idiomas al plantear una consulta en un lenguaje fuente. Luego, la consulta se hace corresponder a una representación común con el fin de recuperar los documentos pertinentes en un idioma destino. La traducción de documentos, o incluso frases, en el lenguaje de consulta requiere de enormes recursos. Por lo general: (1) textos paralelos, (2) sistemas de traducción automática y (3) diccionarios bilingües en formatos legibles por una máquina.

Recientemente, el estudio del procesamiento de varios idiomas ya ha sido abordado en el Cross Language Evaluation Forum (CLEF). En [15] se describen los experimentos y resultados del uso del conjunto de datos de CLEF 2007 en un sistema de recuperación de información de una lengua hindú al inglés. Mediante el uso de una simple traducción de una palabra a otra y el aprendizaje de una tabla de alineación de palabras, obtuvieron 73 % de rendimiento de un sistema monolingüe. Especialmente, el resultado más importante de este trabajo es el descubrimiento de considerar las 4 traducciones de palabras más probables, sin umbral en la probabilidad de la traducción, lo que dio mejores resultados que la traducción de una palabra a otra.

En [6] se presenta un sistema de apoyo para la recuperación de información en varios idiomas, llamado Mulinex, el cual recupera documentos de la Web mediante el empleo de un diccionario basado en la traducción de consultas. Mulinex soporta francés, alemán e inglés mediante el uso de grandes cantidades de datos para la traducción y diferentes algoritmos de categorización de documentos: clasificadores de n-gramas para las entrada con ruido, el algoritmo del k-vecinos más cercanos para los documentos normales y clasificadores basados en patrones para documentos muy cortos. Además de la funcionalidad del cruzamiento de idiomas, Mulinex proporciona la traducción automática de documentos y sus resúmenes. Mulinex usa un asistente de consulta que ofrece la oportunidad de realizar desambiguación de la traducción de consultas interactivas. La traducción de los términos de la consulta son nuevamente traducidos en el lenguaje de consulta inicial. Sin embargo, este enfoque tiene algunas limitaciones claras debido a la falta de uso de sinónimos en el diccionario y porque los homónimos significativos en el idioma de destino pueden dar lugar a la confusión de las traducciones.

En el trabajo [32] se traduce del tamil al inglés utilizando traducción automática estadística. Los autores describen un sistema de Recuperación de Información (RI) en el dominio de la Agricultura para los agricultores de Tamil Nadu el cual les ayuda a especificar sus necesidades de información en tamil y recuperar los documentos en inglés. El sistema está diseñado con aprendizaje dinámico por lo que cualquier palabra nueva que se encuentra en el proceso de traducción se actualiza en el diccionario bilingüe.

En [38], en lugar de utilizar representaciones de documentos existentes, con información adicional en un entorno de agrupación de mútiples vistas, los autores utilizan un enfoque alternativo para codificar la información adicional como restricciones. Los resultados muestran que con los datos reales este enfoque es eficaz en la mejora de la agrupación únicamente a través del uso de los documentos originales.

Un trabajo interesante que utiliza un modelo de mezcla en varios idiomas para la clasificación de sentimientos se presenta en [23] en el cual se utiliza la alineación para reducir el sesgo hacia la lengua inicial de partida en el aprendizaje de transferencia. El modelo propuesto puede aprender sentimientos nunca antes vistos a partir de grandes volúmenes de datos sin etiquetar, que no están cubiertos por el limitado vocabulario en la traducción automática de los datos etiquetados. Con este método se utilizan datos en paralelo sin etiquetar; independientemente de si se utilizan los datos etiquetados en el idioma de destino o no.

2.3. Anotaciones automáticas de tuits

Recientemente, varias investigaciones dentro del proyecto Limosine[2] [1,3] llevan a considerar la anotación automática para el análisis de la reputación en línea (e-reputation) de entidades corporativas, principalmente en inglés. Pero la reputación en línea de figuras públicas también interesa a los investigadores del frances y el español. En el marco de los proyectos Imagiweb[3] [35] y TASS[4] [36], respectivamente para ambos idiomas. En dichos proyectos, se ha observado que la interpretación humana de tuits, es propensa a tener discrepancias. Sigue siendo difícil obtener una anotación fiable y verdadera ya que hechos y opiniones tienen que ser considerados independientemente de si el contenido es de opinión o no. A menudo es difícil saber todas las implicaciones que un mensaje puede tener en la reputación en línea de una entidad determinada. Por último, cabe mencionar que el contexto político hace que la tarea sea aún más difícil. En este trabajo se investiga qué tanto las técnicas de Aprendizaje de Máquinas (Machine Learning) sin datos de entrenamiento correctos pueden ejecutar un trabajo semejante al que hacen los anotadores humanos.

[2] `http://www.limosine-project.eu`

[3] `http://mediamining.univ-lyon2.fr/velcin/imagiweb/`

[4] *Taller de Análisis de Sentimientos en la SEPLN* Ver: `http://www.daedalus.es/TASS2013/corpus.php`

3. Acerca de los conjuntos de datos

3.1. Datos de AMLO durante la última elección presidencial en México

El corpus analizado es el mismo que el utilizado por [16]. Consiste en 800 tuits que contienen el hashtag #AMLO y que fueron extraídos entre el 9 y el 11 de junio de 2012. AMLO es el acrónimo de Andrés Manuel López Obrador, quien fue candidato a las elecciones presidenciales en México. Estos tuits han sido anotados manualmente de acuerdo a la polaridad de la reputación desde el punto de vista del autor[5]. Los desacuerdos en la anotación se han resuelto con el uso de un anotador adicional. El conjunto de datos utilizado es pequeño debido a que la anotación de un gran volúmen de tuits especializados es un proceso que requiere mucho tiempo y dinero. Además, el número de anotadores para esta tarea era muy limitado. Sin embargo, estudios semejantes [27,28] acerca de Twitter y los políticos mexicanos han sido conducidos con la misma cantidad de datos.

Las clases del conjunto de datos de AMLO están bien equilibradas con solo una ligera diferencia con los tuits negativos con respecto a las clases de la colección de datos traducidos en fracés como se muestra en la Tabla 1

Tabla 1: Distribución de las clases en los datos para español de México.

	Distribución de la clase (Español)	**Distribución de la clase** (Francés)
Negativo	0.41	0.37
Neutral	0.29	0.30
Positivo	0.30	0.33

3.2. ImagiWeb: datos de los candidatos durante la última elección presidencial en Francia

La colección utilizada en francés proviene del proyecto ImagiWeb, incluyendo todas las evaluaciones hechas por los anotadores. Esta colección recientemente se ha puesto a disposición del público[6] y fue utilizada en [16,35]. Consiste en 3,184 tuits anotados manualmente[7] para ambos candidatos presidenciales: François Hollande y Nicolas Sarkozy. Los tuits se extrajeron entre marzo y diciembre de 2012 y se refieren a los dos candidatos principales, que coincide casi el mismo período que el conjunto de política en español y el de RepLab, que se describirá más adelante.

La Tabla 2 muestra que la tendencia principal es negativa con un número muy limitado de tuits neutros. Según [35] la razón principal es que la política en Francia desata pasiones entre las personas. Para un análisis razonable solamente consideramos 3 niveles de polaridad sobre los 6 disponibles en el conjunto de datos.

[5] Es decir, si el autor tiene una opinión Positiva, Negativa o Neutral acerca de AMLO.

[6] `http://mediamining.univ-lyon2.fr/velcin/imagiweb/dataset.html`

[7] La anotación fue hecha por treinta personas (con educación superior), en cuanto a la polaridad (están disponibles más detalles estadísticos sobre el proceso de anotación en [35])

Tabla 2: Distribución de clases en la colección política francesa.

	Distribución de la clase
Negativo	0.60
Neutral	0.12
Positivo	0.28

3.3. TASS: datos de la Politica Española

Una parte de la evaluación de TASS 2013 [36] abarca el análisis de los sentimientos sobre mensajes políticos. El corpus proporcionado es una selección de 2,500 tuits (2,150 están todavía disponibles en línea), extraídos de Twitter durante la campaña electoral de las elecciones generales de 2011 en España. Los tuits mencionan a cualquiera de los cuatro principales partidos políticos a nivel nacional: *Partido Popular (PP), Partido Socialista Obrero Español (PSOE), Izquierda Unida (IU) y Unión, Progreso y Democracia (UPyD).* Los tuits fueron anotados manualmente de acuerdo a la polaridad global y la polaridad a nivel de entidad (3 niveles más la etiqueta de ninguno). Esta polaridad a nivel de la entidad es similar a la anotación de polaridad para la reputación de RepLab y a la definición de polaridad en el conjunto de datos Imagiweb. Más detalles sobre el procedimiento conjunto de datos y la anotación se pueden encontrar en [36].

Tabla 3: Distribución de la clase de la colección política española de TASS 2013.

	Distribución de la clase
Negativo	0.27
Neutral	0.38
Positivo	0.26
Ninguno	0.09

La Tabla 3 muestra que la tendencia principal es neutral con una ligera diferencia entre los valores positivos y negativos. Hemos eliminado de nuestros experimentos los tuits marcados como sin polaridad (la etiqueta ninguno).

3.4. RepLab: reputación de entidades

También usamos las anotaciones "polaridad de la reputación" española (23,100 tuits que representan cerca del 20 % de la colección) de la colección de datos de RepLab 2013 [3]. En RepLab el objetivo de la anotación fue decidir si el contenido del tuit tenía consecuencias positivas o negativas para la reputación de la empresa cuando el contenido de las palabras contenían sentimientos explícitos o solo informaban acerca de los hechos. Las anotaciones manuales son: positivo, negativo y neutro.
Como se muestra en la Tabla 4, la tendencia principal en el conjunto de RepLab es positiva.

Tabla 4: Distribución de la clase en el subconjunto español de la colección de RepLab'2013.

	Distribución de la clase
Negativo	0.24
Neutral	0.28
Positivo	0.48

4. Protocolo Experimental

En este trabajo, nos concentramos en cómo mejorar la portabilidad de los datos anotados. Nos enfocamos en determinar las características discriminatorias, que no dependen de las entidades, del dominio o del idioma y que nos sirven para entrenar modelos estadísticos. La elección de dichas variables está motivada por muchas de las limitaciones identificadas en los enfoques de análisis de sentimientos basados en léxico. En primer lugar, debido a que requieren el desarrollo de lexicones de sentimiento y anotación específicos del lenguaje, que son caros, ya que dependen de la mano de obra humana. En segundo lugar, debido a la brevedad, el ruido y la poca edición del texto proveniente de las actualizaciones de las redes sociales donde se limita la cobertura de los lexicones y que resultan menos eficaces que en los textos estándar [12,25]. En tercer lugar, y lo más importante, porque en la opinión política, la polaridad de la reputación rara vez se codifica en palabras que devengan sentimientos; también están integradas en otras palabras y en contextos cortos. Por ejemplo, menciones de asuntos o de organizaciones financieras o escándalos están altamente correlacionados con opiniones negativas, así como las menciones de intención de voto al lado opuesto.

A continuación, se describe a través de nuestros enfoques el problema de detectar la polaridad de reputación utilizando varios métodos.

4.1. Enfoque léxico

Los enfoques léxicos comienzan con una lista de palabras positivas y negativas, que están ya pre-codificadas. En nuestra colección de datos primero se analizó mediante un enfoque léxico combinado con un análisis lingüístico los tuits sociales y políticos, durante un período de tiempo, con el fin de detectar los sentimientos. Comenzamos con un léxico en español y uno traducido del inglés con el fin de contar para cada tuit y para cada corpus el número de palabras positivas y negativas contenidas en cada tuit. Todo el proceso se realiza de forma automática mediante el uso de R[8]. Las palabras contenidas en un tuit se clasificaron en positivas o negativas a partir del uso del léxico sin tomar en cuenta el sarcasmo que transforma la polaridad de un enunciado aparentemente positivo o negativo en su opuesto [13].

4.2. Preprocesamiento de los datos

Ignoramos todos los tuits duplicados (elegimos considerar solo el primero de acuerdo a la fecha). Cada idioma es tratado por igual. El texto se escribe con minúsculas y

[8] http://www.r-project.org/

es limpiado a partir de la eliminación de enlaces de hipertexto, stop-words y signos de puntuación. Los hashtags no se eliminaron.

4.3. Aprendizaje de máquina

Se utilizó el mismo método de aprendizaje de máquina, descrito en [1] y [36] para proponer las anotaciones. El método de bootstrapping [10], ha mostrado que un pequeño conjunto anotado acoplado a la máquina de aprendizaje podría ser competitivo en comparación con los anotadores a la hora de ejecutar tareas de minería de texto. La anotación se abordó como un problema de clasificación que consistió en determinar la polaridad de cada tuit. La elección de nuestros clasificadores está motivada por su buen desempeño en muchas tareas de clasificación en investigaciones previas sobre la detección de polaridad y el análisis de los sentimientos [17,3].

Las características utilizadas por nuestras propuestas son palabras, bi-gramas y tri-gramas. Estas componen la representación de tuits codificados.

El primer modelo explorado fue frecuencia de término – frecuencia inversa de documento (TF-IDF) combinado con los criterios de pureza de Gini [33]. Este último trabajo reporta mejoras en el valor de TF-IDF sobre el uso de n-gramas (n<=3). Con este modelo, estimamos la similitud de un tuit dado comparándolo con cada clase de polaridad y clasificándolo de acuerdo al índice de similitud .
La pureza de una palabra i está definida con el criterio de Gini como sigue (1):

$$gini_i = \sum_{c \in \mathbb{C}} \mathbb{P}^2(i|c) = \sum_{c \in \mathbb{C}} \left(\frac{DF_i(c)}{DF_{\mathbb{T}}(i)} \right)^2 \tag{1}$$

donde C es un conjunto de clases, $DF_{\mathbb{T}}(i)$ es el número de tuits del conjunto entrenado $\mathbb{T}$ conteniendo la palabra i y $DF_c(i)$ es el número de tuits del conjunto de datos entrenados con la clase c conteniendo la palabra i. El factor es utilizado para asignar un valos de peso a la contribución $\omega_{i,d}$ de cada término *i* en el documento *d*, como en la ecuación (2):

$$\omega_{i,d} = TF_{i,d} \times \log(\frac{N}{DF_{\mathbb{C}}(i)}) \times gini_i \tag{2}$$

Donde N es el número de tuits en conjunto entrenado y $\omega_{i,c}$ la contribución de cada término *i* en la clase *c* (3).

$$\omega_{i,c} = DF_{i,c} \times \log(\frac{N}{DF_{\mathbb{C}}(i)}) \times gini_i \tag{3}$$

Distancia coseno. Esta distancia se calcula para comparar las similitudes entre la bolsa de palabras de tuits y cada clase de la bolsa de palabras como se indica en la ecuación (4).

$$cos(d, c) = \frac{\sum\limits_{i \in d \cap c} \omega_{i,d} \times \omega_{i,c}}{\sqrt{\sum\limits_{i \in d} \omega_{i,d}^2 \times \sum\limits_{i \in c} \omega_{i,c}^2}} \tag{4}$$

Máquinas de Soporte Vectorial Los algoritmos de soporte vectorial han demostrado su capacidad para manejar grandes espacios vectoriales y determinar aquellas dimensiones que son relevantes para clasificación [17]. Nosotros elegimos entrenar Máquinas Lineales Multiclase[9] [9] con el objetivo de obtener la polaridad en un solo paso. La representación vectorial de la "bolsa-de-términos-peso"de cada tuit d está dada por la ecuación (5).

$$\omega_i = DF_{\mathbb{T}}(i) \times log(\frac{N}{DF_{\mathbb{T}}(i)}) \times gini_i \tag{5}$$

Baseline El algoritmo de base (baseline) se calcula etiquetando cada tuit d_1 de prueba con el tuit más similar d_2 en el conjunto de entrenamiento (según el índice de Jaccard). Esta similitud se calcula como se indica en la ecuación 6.

$$sim(d_1, d_2) = \frac{\sum\limits_{i \in d_1 \cap d_2} \omega_{i,d}}{\sum\limits_{i \in d_1 \cup d_2} \omega_{i,d}} \tag{6}$$

4.4. Información general del procesamiento

Las figuras 1, 2, 3 y 4 describen nuestra evaluación experimental. A continuación se explica cada evaluación.

4.5. Traducción del léxicon y los documentos

Como primer paso en la experimentación elegimos traducir manualmente una muestra de 220 tuits de nuestro conjunto de datos no etiquetados de México. Esto con el fin de realizar una clasificación automática utilizando el conjunto anotado francés como conjunto de entrenamiento, como se muestra en la figura 1. La muestra elegida para la traducción presenta un mejor equilibrio de clases (como se muestra en la tabla 1). El principal objetivo de este experimento es verificar la aplicabilidad de los mismos modelos a otro conjunto una prueba con otras entidades como se hizo durante RepLab 2012 [2]. Asimismo, separamos los dos candidatos del conjunto de Imagiweb en un conjunto separado de entrenamiento para evaluar si un candidato puede ser más similar a AMLO que otro.

Elegimos traducir, del inglés al español, un recurso léxico muy conocido utilizado en muchas tareas de Análisis de Sentimientos (AS) descrito en [4]. Hemos utilizado los datos de Bing Liu que se compone de alrededor de 6,800 palabras en inglés. Sin embargo, después de hacer una traducción automática utilizando el traductor de Google hemos recuperado solo 2,284 palabras positivas y 1,644 palabras negativas. En este caso, dejamos, de forma manual, solo las palabras que en español expresan un sentimiento. También se compararon los resultados de este enfoque con una clasificación utilizando un recurso léxico en español y construido específicamente para AS en Twitter y para el análisis de la política (Lexicon ElhPolar [29]). Luego, se evaluaron ambos

[9] Multi-Class Support Vector Machine `http://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html`

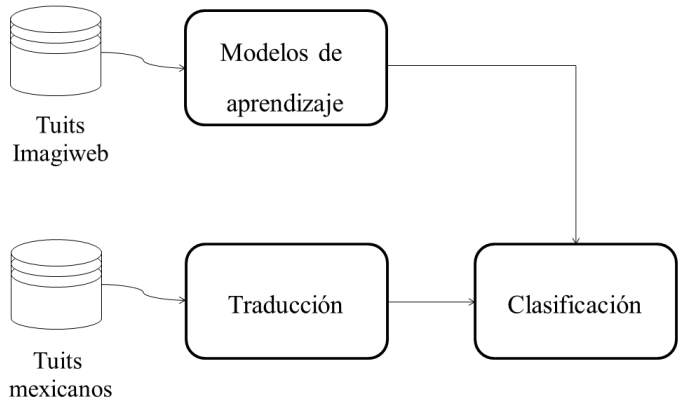

Figura 1: Proceso de clasificación usando los documentos traducidos

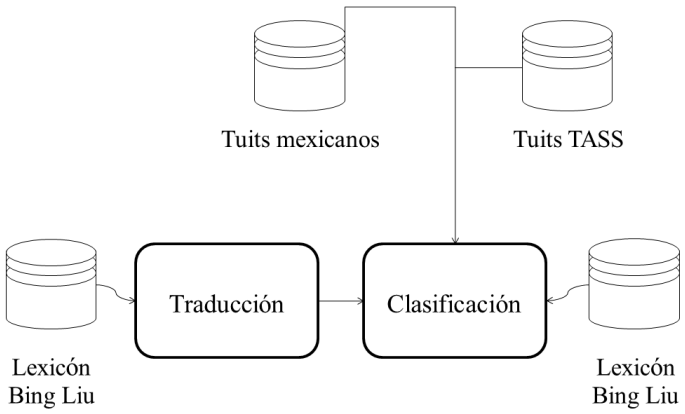

Figura 2: Clasificación usando el enfoque léxico

recursos léxicos usando el conjunto de datos TASS'2013 para comparar estos enfoques con respecto al estado del arte (véase la figura 2 y la Tabla 8).

A continuación, experimentamos en la clasificación utilizando el conjunto de reputación RepLab 2013 como conjunto de entrenamiento. Nos cuestionamos aquí acerca del

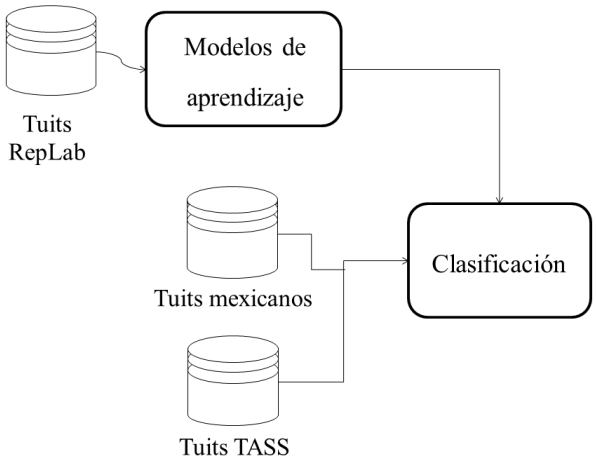

Figura 3: Proceso de clasificación usando el conjunto de datos RepLab'13 como conjunto de entrenamiento

desempeño de un mismo idioma en un conjunto masivamente etiquetado compartiendo el mismo, pequeño y ruidoso, vocabulario específico en textos provenientes de la red social. Se realizó la clasificación sobre los tuits mexicanos y el conjunto de datos políticos de TASS'2013 como se muestra en la figura 2.
En un último experimento usamos el conjunto de datos de TASS, como el mismo lenguaje, en el conjunto de entrenamiento para anotar automáticamente los tuits mexicanos (veáse figura 4).

5. Evaluación y resultados

5.1. Métricas

Para cada clasificador evaluado reportamos la precisión global y si la clase no está bien balanceada en cada conjunto de datos proponemos reportar el cálculo promedio de F-Score (Macro Averaged F-Score, anotada como F-Score en las tablas) así como la media de cada clase basádonos en la precisión y la exhaustividad con el fin de dar la misma importancia a cada clase. Por ejemplo, tomando en cuenta que la clase negativa representa el 60 % del conjunto de Imagiweb devolviendo todos los tuits como negativos daríamos una exactitud (accuracy) y un F-Score global del 60 % y una Macro F-Score Promedio F-Score del 33 %, ya que solamente detecta una clase, lo que no representa un rendimiento significativo y eficiente del sistema.

La fórmula para calcular la exactitud se presenta en la ecuación (7).

$$Exactitud = \frac{\text{Num. de documentos correctos}}{\text{Num. de documentos en referencia}} \tag{7}$$

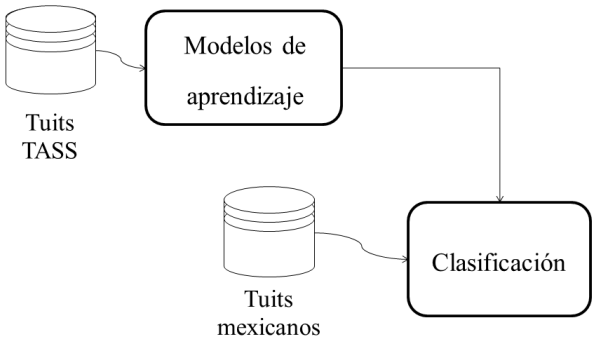

Figura 4: Proceso de clasificación usando el conjunto TASS 2013 como conjunto de entrenamiento

Y la Macro F-Score Promedio como en (8).

$$\text{F\_Score} = \frac{\sum_c \frac{2 \times (\text{Precision}_c \times \text{Exhaustividad}_c)}{\text{Precision}_c + \text{Exhaustividad}_c}}{\text{Num. de clases}} \tag{8}$$

La Precisión P_c para la clase c se determino por (9).

$$P_c = \frac{\text{Num. de documentos correctamente asignados en la clase}_c}{\text{Num. de documentos asignados en la clase}_c} \tag{9}$$

Y la Exhaustividad R_c para la clase c según (10)

$$R_c = \frac{\text{Num. de documentos correctamente asignados en la clase}_c}{\text{Num. de documentos pertenecientes a la clase}_c} \tag{10}$$

5.2. Aprendizaje de máquina usando datos traducidos

El conjunto Imagiweb ofrece anotación de sentimientos a nivel de persona, lo cual es más similar con respecto a nuestros tuits mexicanos proporcionando un experimento interesante. De acuerdo con la Tabla 5 el desempeño de los sistemas usando los mismos datos de contexto es realmente bajo. Las principales razones son que el vocabulario utilizado para describir a los dos principales candidatos franceses no es el mismo que el utilizado para AMLO además de que ambas distribuciones de clase parecen ser demasiado diferentes. A medida que el desempeño de los sistemas no aumentan mientras

que el tamaño del conjunto de entrenamiento sí aumenta (al considerar el conjunto Imagiweb secc. 6), se puede considerar que el rendimiento de los sistemas no está limitado a la cantidad de los datos de entrenamiento disponibles. Sin embargo, con el tamaño tan limitado de este conjunto de pruebas no somos capaces de concluir sobre la diferencia significativa entre el desempeño de los sistemas.

Tabla 5: Desempeño en la clasificación de tuits traducidos.

	FH sub-conjunto		NS sub-conjunto	
	F-Score	Exhaustividad	F-Score	Exhausitividad
Baseline	0.29	0.33	**0.38**	0.39
Coseno	0.17	0.33	0.34	**0.42**
MVS	**0.34**	**0.36**	**0.38**	0.40

Como los resultados son ligeramente mejores cuando consideramos tuits de NS como conjunto de entrenamiento podemos considerar que los usuarios de Twitter hablan de AMLO en forma similar a la forma en que hablan de NS .

Tabla 6: Desempeño en la clasificación de tuits traducidos.

	Imagiweb	
	F-Score	Exhaustividad
Baseline	0.30	**0.33**
Coseno	0.26	**0.38**
MVS	**0.35**	0.37

5.3. Léxicos traducidos y especializados en español

Ambos enfoques léxicos, Elh-polar y Bing Liu traducido, parecen no encajar con nuestro vocabulario del conjunto de datos o este tipo de análisis no se desempeña bien como se observa en la Tabla 7.

Tabla 7: Desempeño en la clasificación usando lexicon en datos de México.

	F-Score	**Exhaustividad**
Lexicon ElhPolar	0.25	0.32
Lexicon traducido (Bing Liu)	0.21	0.33

El enfoque léxico muestra también limitaciones sobre el conjunto de datos de TASS'2013 como se reporta en la Tabla 8. Queríamos comparar los mejores sistemas participantes en *TASS'2013, Tarea 3: Análisis de Sentimientos en Nivel Entidad* pero como los

organizadores de TASS no informaron sobre el Macro Promedio F-Score no somos capaces de evaluar el ”mejorrendimiento TASS en términos de F- Score y exactitud. Sin embargo, como los organizadores TASS reportan un F- Score global basado en la precisión y exhaustividad suponemos que los resultados que obtuvimos son en su mayoría similares al promedio de los participantes TASS'2013 (lo cual no se informa en este artículo).

Tabla 8: Desempeño en la clasificación usando un lexicon en el conjunto TASS'2013.

	F-Score	**Exhaustividad**
Lexicon ElhPolar	0.30	0.41
Lexicon traducido (Bing Liu)	0.12	0.29

5.4. Aprendizaje de máquina en datos en español fuera del dominio político

En este experimento se realizó la clasificación utilizando el conjunto de RepLab 2013 como conjunto de entrenamiento. Sin embargo, RepLab también proporciona anotación a nivel persona para algunos cantantes que pueden ser similares a los de los candidatos. La Tabla 9 muestra el rendimiento de clasificación sobre contenidos en español de acuerdo con F-Score y la exactitud. MVS es inferior al método baseline cuyo rendimiento es también más alto de lo esperado. El método por distancia coseno muestra buen desempeño, utilizando los tuits traducidos. Aquí se puede concluir que es posible obtener muy buenos resultados de la clasificación, con valores de confianza cercanos a los valores de acuerdo inter-anotador observados en la literatura [3,36,26]. Se puede deducir que con este rendimiento de los clasificadores ya se puede proporcionar una buena aproximación a la anotación manual pero de manera más rápida y más económica que lo que se puede esperar con recursos humanos.

Tabla 9: Desempeño en la clasificación en tuits mexicanos.

	F-Score	**Exhaustividad**
Baseline	0.50	0.51
Coseno	**0.74**	**0.74**
Lexicon ElhPolar	0.25	0.32
Lexicon traducido (Bing Liu)	0.21	0.33
MVS	0.17	0.31

La anotación a nivel de entidad de RepLab ofrece una granularidad similar a la anotación de “Party/Entity” de TASS. Sin embargo, cuando consideramos TASS'2013 como conjunto de prueba se obtiene un bajo nivel de rendimiento como se muestra en la Tabla 10.

Tabla 10: Desempeño en la clasificación usando el conjunto de TASS'2013.

	F-Score	Exhaustividad
Baseline	0.32	0.33
Coseno	0.32	0.33
MVS	0.33	0.33

5.5. Aprendizaje de máquina usando datos de dominio

En este experimento se realizó la clasificación utilizando el conjunto de datos de TASS'2013 para entrenamiento. La Tabla 11 muestra el rendimiento de la clasificación.

Tabla 11: Desempeño en la clasificación usando los tuits mexicanos.

	F-Score	Exhaustividad
Baseline	0.33	0.32
Coseno	**0.32**	**0.31**
MVS	0.31	0.29

El nivel más bajo de los resultados puede explicarse por el pequeño tamaño del conjunto de entrenamiento (en comparación con RepLab). Sin embargo, es bastante cercano con respecto a otros clasificadores, aunque el tamaño de la prueba es limitada. Por último, el principal hallazgo de estos experimentos es mostrar que las palabras utilizadas, este año, en este contexto son muy diferentes a las utilizadas en los dos conjuntos de México y RepLab .

5.6. Análisis cualitativo

Tratar con contenidos ambiguos a menudo conduce a notar errores interesantes. Algunos tuits tales como:
"RT 1.Naces 2.Eres AMLO 3. Creces 4. No eres presidente. 5. No eres presidente. 6. No eres presidente. 7. No eres presidente. 8. Mueres. JAJA" es etiquetado positivo por los sistemas cuando en realidad es un tuit negativo acerca del candidato AMLO.

Aquí hay otro ejemplo:
"AMLO gran orador cada vez que abre la boca sueña #elpejeaburrehastaalospejezombies"

El cual, en este caso se trata de una ironía debido a que la gente no está soñando con un país mejor, por el contrario se está aburriendo y casi duermiendo cada vez que AMLO habla.

Los sistemas automáticos también se beneficiarían de la división del hashtag ya que no son capaces de entender las palabras unidas en una sola como *"#esunhonortuitearporobrador o #alpejenolesalenlascuentas"*. Sin embargo, además de las reglas lingüísticas se requerirá un procesamiento más profundo incluyendo los conocimientos de lenguage como en el caso de *"#elpejeaburrehastaalospejezombies #elpejeaburrehastaasuspejezombies y #elpejeaburrehastalospejezombies* donde todos son la misma declaración.

Estos son ejemplos típicos de contenidos humorísticos que los sistemas no son capaces de manejar adecuadamente. Los léxicos probablemente nunca serán capaces de considerar correctamente este tipo de mensajes. Mientras que los enfoques de AM podrían manejar estos una vez que hayan visto ejemplos similares en el conjunto de entrenamiento o en un procedimiento de aprendizaje activo.

6. Conclusiones

En este artículo describimos y comparamos varios enfoques para la clasificación automática de tuits de opinión popular en política en español concernientes a la última elección presidencial en México. Este tipo de contenido es a menudo difícil de entender y la anotación es propensa a tener errores humanos. Nuestra evaluación experimental (a pesar de tener un conjunto de prueba limitado) establece que sin un entrenamiento específico del material, los enfoques de aprendizaje de máquina pueden lograr los resultados del estado del arte, mientras que en la literatura se insiste en la necesidad de datos de entrenamiento anotados manualmente.

La anotación de tuits relacionados con la política es conocida por ser una tarea difícil. Nuestros experimentos han demostrado que la necesidad de una anotación costosa por parte de un experto puede ser reconsiderada. Los enfoques de AM presentados son independientes del idioma y del dominio. Así, solo con un mínimo esfuerzo se pueden adaptar estos métodos a otro dominio, como por ejemplo, la popularidad de productos o entidades corporativas, y asimismo, se pueden manejar una gran cantidad de datos. Otra de nuestras aportaciones es el conjunto de datos anotados que pueden ser utilizados en investigaciones futuras: `https://github.com/some/url`

Referencias

1. Carrillo de Albornoz, J., Amigó, E., Spina, D., Gonzalo, J.: ORMA: A semi-automatic tool for online reputation monitoring in twitter. In: Advances in Information Retrieval - 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. pp. 742–745. Springer International Publishing (2014)
2. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., De Rijke, M.: Overview of replab 2012: Evaluating online reputation monitoring systems. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 333–352. Springer (2012)
3. Amigó, E., De Albornoz, J.C., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., De Rijke, M., Spina, D.: Overview of replab 2013: Evaluating online reputation monitoring systems. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 333–352. Springer (2013)
4. Bing, L.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2012)
5. Bollen, J., Mao, H., P, A.: Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
6. Capstick, J., Diagne, A.K., Erbach, G., Uszkoreit, H., Leisenberg, A., Leisenberg, M.: A system for supporting cross-lingual information retrieval. Information Processing and Management 36(2), 275–289 (2000)

7. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in twitter: The million follower fallacy. In: ICWSM'10: Proceedings of international AAAI Conference on Weblogs and Social (2010)
8. Chung, J.E., Mustafaraj, E.: Can collective sentiment expressed on twitter predict political elections? In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011 (2011)
9. Crammer, K., Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research 2, 265–292 (2002)
10. Di Fabbrizio, G., Tur, G., Hakkani-Tur, D.: Bootstrapping spoken dialog systems with data reuse. In: Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. pp. 72–80. Association for Computational Linguistics, Cambridge, Massachusetts, USA (2004)
11. Dodds, P., Danforth, C.: Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. Journal of Happiness Studies 11(4), 441–456 (2010)
12. Feczko, M., Schaye, A., Marcus, M., Nenkova, A.: Sentisummary: Sentiment summarization for user product reviews. In: proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. vol. 1, pp. 265–271 (2008)
13. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: A closer look. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers. pp. 581–586 (2011)
14. Gruzd, A.A., Doiron, S., Mai, P.: Is happiness contagious online? a case of twitter and the 2010 winter olympics. In: HICSS. pp. 1–9 (2011)
15. Jagarlamudi, J., Kumaran, A.: Cross-lingual information retrieval system for indian languages. In: Advances in Multilingual and Multimodal Information Retrieval. pp. 80–87. Springer Berlin Heidelberg (2008)
16. Jean-Valere Cossu, R.A., Mena, A.M., Torres-Moreno, J.M., SanJuan, E.: Bilingual and cross domain politics analysis. Avances en la Ingeniería del Lenguaje y del Conocimiento p. 9 (2014)
17. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. Springer (1998)
18. Kramer, A.D.I.: An unobtrusive behavioral model of "gross national happiness". In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Atlanta, Georgia, USA, April 10-15, 2010. pp. 287–290 (2010)
19. Lampos, V., Preoţiuc-Pietro, D., Cohn, T.: A user-centric model of voting intention from social media. In: ACL '13. p. 993–1003. Association for Computational Linguistics, Association for Computational Linguistics, Sofia, Bulgaria (08/2013 2013)
20. Malouf, R., Mullen, T.: Taking sides: User classification for informal online political discourse. Internet Research 18, 177–190 (2008)
21. Martínez, V., González, V.M.: Sentiment characterization of an urban environment via twitter. In: Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction - 7th International Conference, UCAmI 2013, Carrillo, Costa Rica, December 2-6, 2013, Proceedings. pp. 394–397 (2013)
22. Maynard, D., Bontcheva, K., Rout, D.: Challenges in developing opinion mining tools for social media. In: Proceedings of NLP can u tag # user generated content (2012)
23. Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., Wang, H.: Cross-lingual mixture model for sentiment classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. pp. 572–581. Association for Computational Linguistics, Stroudsburg, PA, USA (2012)
24. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the Fourth Inter-

national Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010 (2010)
25. Ohana, B., Tierney, B.: Sentiment classification of reviews using sentiwordnet. In: 9th. IT & T Conference. p. 13 (2009)
26. Pla, F., Hurtado, L.: Political tendency identification in twitter using sentiment analysis techniques. In: COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 183–192 (2014)
27. Sandoval, R., Matus, R.T., Rogel, R.N.: Twitter in mexican politics: Messages to people or candidates? In: 18th Americas Conference on Information Systems, AMCIS 2012, Seattle, Washington August 9-11, 2012 (2012)
28. Sandoval-Almazán, R.: Using twitter in political campaigns: The case of the PRI candidate in mexico. IJEP 6(1), 1–15 (2015)
29. Saralegi, X., Vicente, I.S.: Elhuyar at tass 2013. In: XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS2013). pp. 143–150 (2013)
30. Tello-Leal, E., Tello-Leal, D.A., Sosa Reyna, C.M.: Reflexiones sobre el uso de las tecnologías de información y comunicación en las campañas electorales en méxico: e-campañas. Revista Virtual Universidad Católica del Norte 36, 33–47 (2012)
31. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in twitter events. Journal of the American Society for Information Science and Technology 2(62), 406–418 (2011)
32. Thenmozhi, D., Aravindan, C.: Cross lingual information retrieval system for agriculture society. In: International Forum for Information Technology in Tamil Conference (INFITT) (2009)
33. Torres-Moreno, J., El-Beze, M., Bellot, P.: Bechet, opinion detection as a topic classification problem in in textual information access. chapter 9 (2013)
34. Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM. pp. 178–185 (2010)
35. Velcin, J., Brun, C., Dormagen, J.Y., Kim, Y.M., Roux, C., Boyadjian, J., Bonnevay, S., Neihouser, M., SanJuan, E., Khouas, L., A., M., M, N.: Investigating the image of entities in social media: Dataset design and first results. In: Language Resources and Evaluation Conference (LREC) (2014)
36. Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., Cristóbal, J.C.G.: TASS - workshop on sentiment analysis at SEPLN. Procesamiento del Lenguaje Natural 50, 37–44 (2013)
37. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea. pp. 115–120 (2012)
38. Wang, X., Qian, B., Davidson, I.: Improving document clustering using automated machine translation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 645–653 (2012)
39. Wijaya, V., Erwin, A., Galinium, M., Muliady, W.: Automatic mood classification of indonesian tweets using linguistic approach. In: International Conference on Information Technology and Electrical Engineering (ICITEE). pp. 41–46. IEEE (2013)

Tratamiento de la implicación textual a través de patrones semánticos

Saul León, Darnes Vilariño, David Pinto, Mireya Tovar, Beatriz Beltrán

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, Puebla, México

{saul.leon, darnes, dpinto, mtovar, bbeltran}@cs.buap.mx

Resumen. En el presente trabajo se presenta una nueva metodología para el Problema de Implicación Textual. Esta metodología propone el empate de patrones sintácticos, los cuales están construidos por la cantidad de términos. Esta aproximación es una variante del modelo de aproximación por empatamiento. Esta metodología fue probada y validada con tres corpus ofrecidos por varias conferencias internacionales (RTE1, CLTE y SICK). Los resultados obtenidos son comparables con los mejores resultados internacionales.

Palabras clave: Implicación textual, similitud semántica, taxonomía de WordNet, medidas semánticas.

1. Introducción

Encontrar de manera automática cuando dos textos comparten la misma información y en el mismo sentido, es uno de los problemas que en la actualidad está siendo muy estudiado por la comunidad de investigadores en Procesamiento del Lenguaje Natural (PLN). Lograr resolver de manera eficiente este problema beneficiaría a varias tareas de PLN tales como: Recuperación de Información, Traducción Automática, Resúmenes Automáticos, Preguntas y Repuestas, Pruebas de Lectura Comprensiva, entre otras. Anualmente se propone la conferencia International Workshop on Semantic Evaluation (SemEval), en este sentido desde hace cuatro años se ha planteado la tarea de Implicación Textual (IT). Los modelos desarrollados hasta el momento no superan más del 80% de precisión y dependen mucho de las características del corpus de entrenamiento y del corpus de prueba.

El problema a resolver se formula de la siguiente forma: Sean H y T, dos fragmentos de textos en un mismo idioma, el problema de implicación textual consiste en determinar si se puede inferir el significado de H a partir del significado de T. Cuando esto ocurre se dice que existe una implicación textual entre ambos fragmentos. De la misma forma se puede detectar la dirección de la implicación, es decir, si T contiene más información que H, y sostienen implicación textual, se dice que la dirección de la implicación es de T → H, en caso contrario, si H contiene más información que T y sostienen implicación textual, la dirección es T ← H.

También en muchas situaciones de la vida real es necesario detectar de manera automática cuando dos segmentos de texto, son contradictorios, esta situación es un caso particular de la implicación textual.

Para resolver el problema de la implicación textual, algunos autores han recurrido a modelos que utilizan fundamentalmente conteo estadístico, donde explotan diversas medidas de similitud a partir de un par de oraciones y como resultado confeccionan vectores característicos, los vectores representativos obtenidos alimentan a un algoritmo de aprendizaje supervisado y utilizan algún modelo de clasificación. Uno de los mejores trabajos reportados que emplea este tipo de metodología, es el presentado en [2]. En este trabajo se utilizan tres tipos de características; dentro de las características del primer tipo se emplean métricas basadas en conjuntos, suponiendo que A es el conjunto de palabras que conforman al texto 1 y B el conjunto de palabras del texto 2, entonces se calcula: |A|, |B|, |A - B|, |B - A|, |A ∩ B|, |A ∪ B|, |A / B|, |B / A|. El segundo tipo de características es calculado a partir de métricas de similitud, algunas de estas son: Coeficiente de Jaccard, Coeficiente Dice, Coeficiente de Sobrelapamiento, Pesos de Coeficientes sobrelapados, Similitud Coseno, Distancia de Manhattan, Distancia Euclideana, Distancia de Edit, Distancia de Jaro-Winker, por mencionar algunas. Como tercer grupo de características, se cuantifican las Relaciones Gramaticales a través de dependencias funcionales, Relaciones Parciales y Número de Entidades en común.

En general, realizar un conteo estadístico de características que comparten ambos textos, ofrece un buen comportamiento siempre y cuando en los datos de prueba aparezca información que se encuentra en los datos de entrenamiento, esto se ve reflejado en que estas propuestas no logran detectar de manera correcta el juicio de implicación con una exactitud mayor al 60%.

En la actualidad se está trabajando en la construcción de modelos no supervisados, es decir que no dependan de los datos de entrenamiento. Por ejemplo, en el trabajo desarrollado en [3] busca la construcción de lo que se conoce como patrones de empatamiento. Un patrón de empatamiento se basa en la aparición de términos iguales o similares que comparten un par de oraciones, por ejemplo:

Oración 1: Leonardo Da Vinci ***painted the*** Mona Lisa.

Oración 2: Mona Lisa ***is the work of*** Leonardo da Vinci.

Plantilla 1: X ***paint*** Y.

Plantilla 2: Y ***is the work of*** X.

Patrón de empatamiento: X ***paint*** Y → Y ***is the work of*** X.

El uso de plantillas garantiza que la implicación textual siempre es descubierta. En la actualidad se ha propuesto una metodología no supervisada para generar patrones de empatamiento sobre un large-scale-corpus [4], sin embargo construir dichos patrones de empatamiento es una tarea compleja y nuevamente depende de las características que tengan los textos con los que se está trabajando.

Una vertiente nueva intenta generalizar los patrones extraídos aplicando lógica proposicional [5], a continuación se muestra un ejemplo:

Oración 1: Leonardo Da Vinci ***painted the*** Mona Lisa.

Oración 2: Mona Lisa ***is the work of*** Leonardo da Vinci.

Equivalencia Lógica 1: ***isPainterOf***(Da Vinci, Mona Lisa).

Equivalencia Lógica 2: ***isWorkOf***(Mona Lisa, Da Vinci).

Regla: ∀x ∀y ***isPainterOf***(x,y) → ***isWorkOf***(y,x).

La complejidad de esta metodología al igual que la de las plantillas en que es necesario construir todos los axiomas que se presentan en el texto que se está estudiando, la generación de reglas de manera automática aún no está totalmente desarrollada.

La detección de la implicación textual utilizando estas dos últimas metodologías, solamente depende de los patrones y las reglas que puedan ser descubiertas en la fase de construcción, sin embargo es conocido que los seres humanos cuando escriben, no emplean reglas ni patrones rígidos, la riqueza de los textos depende mucho del conocimiento propio de la lengua de cada persona, automatizar este proceso es prácticamente imposible, la riqueza del lenguaje y la forma en que se establecen las relaciones en el mismo es infinita, de modo que aplicar plantillas a textos de la vida real, es similar a estudiar una muestra muy pequeña de textos textualmente implicados.

La metodología de patrones de empatamiento, a pesar de su rigidez, puede ser mejorada, con vistas a detectar juicios de implicación textual en textos que no empaten con los patrones construidos, en este sentido está enfocada nuestra investigación.

Este trabajo está organizado de la siguiente forma: en la Sección 2 se presenta el modelo propuesto para abordar la IT. En la Sección 3 se presentan las características de las colecciones de datos y el comportamiento del modelo desarrollado. En la Sección 4 se analizan los resultados obtenidos para tres colecciones de datos. Finalmente las conclusiones y el trabajo a futuro se encuentran en la Sección 5.

2. Modelo basado en plantillas para resolver la IT

Para introducir el modelo propuesto, se considera el siguiente ejemplo:

Oración 1: *La becerra de Manuel está en la calle*.

Oración 2: *Esta es la calle de Manuel Becerra*.

Ambas oraciones comparten la misma cantidad de información, pero no transmiten la misma idea, ante esta evidencia, es necesario proponer un modelo que para descubrir la implicación textual sea capaz de detectar los *tokens* que comparten un par de oraciones, y descubrir si estos *tokens* transmiten la misma idea en ambas oraciones, en este sentido se retoma el concepto de términos anclas de la metodología de empatamiento por plantillas [15].

Consideremos el par de plantillas X ***es obra de*** Y y X ***escribió*** Y, se observa que en ambas platillas se repiten las variables X e Y, estas variables pueden ser sustituidas por cualquier par de *tokens*, como por ejemplo ***Tokio-Blues*** es obra de ***Murakami***, ***Murakami*** escribió ***Tokio-Blues***; o incluso ***Cien-años-de-soledad*** es obra de ***García-Márquez***, ***García-Márquez*** escribió ***Cien-años-de-soledad***. Los tokens ***Cien-años-de-soledad*** y ***García-Márquez*** son denominados términos anclas, porque están presentes en ambas oraciones.

Una vez detectada las posiciones que ocupan en ambas oraciones los términos ancla, se puede extraer por cada oración la relación que sostienen, esta relación corresponde a la información que está contenida entre las anclas, siguiendo el ejemplo, la relación para las oraciones serian: ***es-obra-de*** y ***escribió***, la manera tradicional de validar si un par de relaciones sostiene o no implicación textual, es usar un repositorio de plantillas. Si en el repositorio de plantillas no se encuentran estas relaciones se dice que no existe implicación textual. Es importante destacar que a pesar de que no se encuentren ese par de relaciones en el repositorio, puede existir una implicación textual.

En este trabajo de investigación se propone emplear medidas de similitud semántica, para establecer si dos relaciones significan lo mismo, es decir, se propone un mecanismo que determine si ***es-obra-de*** y ***escribió*** son similares semánticamente.
Se proponen utilizar seis medidas de similitud semántica entre palabra y palabra y dos medidas de similitud estadística, las cuales se explican a continuación:

2.1 Medidas de similitud semántica basadas en la taxonomía de WordNet

Este tipo de medidas utiliza la taxonomía de WordNet [20] para encontrar la relación semántica que existe entre pares de términos, de acuerdo a la jerarquía establecida. La diferencia entre cada una de las medidas radica en la forma en que se recorre esta estructura jerárquica. Las medidas utilizadas son:

- Similitud de P,ath [8],
- Similitud de Leacock-Chodorow [10],
- Similitud de Wu-Palmer [11],
- Similitud de Resnik [12],
- Similitud de Jiang-Conrath [13],
- Similitud de Lin [14].

2.2 Medidas de similitud estadísticas

Para detectar si dos términos se encuentran relacionados se puede aplicar el algoritmo de Análisis Semántico Latente (LSA por sus siglas en inglés) [16] la idea de esta técnica es partir de una colección de documentos y si dentro de los documentos los términos aparecen relacionados es porque se refieren al mismo concepto. La limitante de esta medida está precisamente en la colección de documentos para calcular las relaciones entre los términos. El corpus debe estar conformado por un dominio de información amplio.

En los últimos años se ha utilizado con mucha frecuencia el concepto de Información Mutua (*Pointwise Mutual Information*, PMI [8]) en diversas tareas de procesamiento de lenguaje natural. Esta técnica busca detectar si dos términos son colocaciones, esto quiere decir que la probabilidad de aparición de ellos juntos, en el mismo contexto, supera a la aparición de los términos por separado. Aplicado el PMI a nuestra investigación nos permite detectar la probabilidad de que dos términos estén cerca, partiendo del hecho que si están muy cercanos reflejan el mismo sentido.

La obtención adecuada de los valores PMI dependen totalmente del tamaño del corpus de donde se extraen, de la misma manera que en LSA, se recomienda un corpus de dominio amplio.

2.3 Eliminación de *tokens* para el descubrimiento de la implicación textual

Para detectar el juicio de implicación textual que sostienen dos oraciones, se sigue el siguiente razonamiento: Las oraciones tienen información en común, unas más y otras menos, si se toman estos elementos comunes y se eliminan en ambas oraciones, quedarían términos únicos en cada oración. Los términos sobrantes, al finalizar el proceso de eliminación, se pueden cuantificar en relación a la longitud original de la oración, es decir, que es posible medir el porcentaje de elementos eliminados y así determinar el juicio de implicación textual, por ejemplo si al terminar de eliminar los elementos comunes de un par de oraciones, se obtiene que se ha eliminado más del 80% en ambas oraciones, quiere decir que la cantidad de información compartida es mucha y por lo tanto se dice que existe implicación textual. Como caso particular, se tiene el juicio de implicación textual de contradicción, para detectar éste es necesario introducir elementos de negación propios del idioma con el que se está trabajando, por ejemplo para el idioma inglés se consideran: no, none, nobody, there-*-not, *n't, así como la presencia de antonimia.

Se propone un algoritmo de carácter general que permite obtener el grado de relación entre un par de sentencias, independientemente del tipo de juicio de implicación que se desea detectar, el proceso que aplica es el siguiente: Los términos anclas por simplicidad son los *n-gramas* de palabras más largos presentes en ambas oraciones. Para el procesamiento se tokenizan ambas oraciones y se verifica la distancia a la que se encuentran las anclas, si los términos anclas no exceden un máximo de dos *tokens* entre ellos en ambas oraciones, se verifican las relaciones que sostienen en una y en otra, para ello se emplean las medidas de similitud. Si las relaciones son similares, se eliminan estos términos (criterio de eliminación) y las anclas involucradas en cada oración, ver Algoritmo 1.

El algoritmo propuesto, emplea las siguientes funciones:

- **obtener_anclas:** Esta función genera los *n-gramas* más largos que comparten ambas oraciones.
- **are_near:** Esta función regresa verdadero si en una oración, los términos ancla se encuentran a no más de 2 tokens de distancia, en caso contrario el valor de retorno es falso.
- **obtener_relacion:** Esta función regresa el segmento de texto de una oración, que se encuentra acotado por un par de anclas.
- **are_similar_relations:** Es una función que decide si dos segmentos de texto son similares o están relacionados, internamente esta función hace uso de las 8 medidas de similitud. Dentro de la función se ha establecido un umbral (REL), si la medida de similitud semántica supera dicho umbral entonces se dice que la relación es válida, que en otras palabras significa que es un plantilla válida.

```
Funcion obtener_grado_de_relacion(Sentencia S1,
                                  Sentencia S2,
                        Vocabulario contradiccion){

 anclas = obtener_anclas(S1, S2)
```

```
 tokens_1 = tokenizar(S1)
 tokens_2 = tokenizar(S2)

 hay_contradiccion = false
  for ci in contradiccion
    si ci in s1 or ci in s2 or
       existen_antonimos_entre(s1, s2)
     hay_contradiccion = true

 len_s1_inicial = length(S1)
 len_s2_inicial = length(S2)

 para cada n-grama ai en anclas hacer
  para cada n-grama aj en anclas hacer
   si ai != aj and are_near(ai, aj, tokens_1) and
       are_near(ai, aj, tokens_2) hacer
     r1 = obtener_relacion(ai, aj, tokens_1)
     r2 = obtener_relacion(ai, aj, tokens_2)

   si are_similar_relations(r1, r2) hacer
     eliminar ai, r1 y aj de S1
     eliminar ai, r2 y aj de S2
     eliminar ai y aj de anclas

 len_s1_final = length(S1)
 len_s2_final = length(S2)

 p1 = 1 - (len_s1_final / len_s1_inicial)
 p2 = 1 - (len_s2_final / len_s2_inicial)

 regresar p1, p2, hay_contradiccion
}
```

Algoritmo 1. Algoritmo empleado para obtener el grado de similitud entre un par de sentencias.

3. Caso de estudio

Como la metodología propuesta es de carácter general, se propone utilizar tres colecciones de datos, que buscan detectar juicios de implicación textual, pero de diferente tipo. A continuación se describen las tres colecciones de datos, así como la forma de detectar en cada caso el juicio de implicación.

3.1. Conjunto de datos RTE1

Esta colección está diseñada solamente para detectar los juicios de implicación textual: *entailment* y *no_entailment*. Es una colección que contiene 800 instancias de

las cuales 400 son *entailment* y 400 son *no_entailment* [17]. Para esta colección a partir de los resultados que obtiene el Algoritmo 1, se hace lo siguiente:

Sea *C* el conjunto de elementos comunes entre las oraciones *S1* y *S2*.

Sea *P1* el porcentaje restante de la oración *S1* al eliminar los elementos de *C*.

Sea *P2* el porcentaje restante de la oración *S2* al eliminar los elementos de *C*.

Sea *U* un umbral de porcentaje.

El juicio de implicación se obtiene aplicando las siguientes reglas:

Si *P1* > *U* y *P2* > *U* el juicio de implicación es *entailment*, en caso contrario el juicio de implicación es *no_entailment*.

3.2. Conjunto de datos CLTE

Esta colección maneja que *S1* se encuentre en un idioma y *S2* en otro, sin embargo se puede traducir la oración *S2* para que se encuentre en el mismo idioma de *S1*, que para esta colección es el idioma inglés [18]. La colección está compuesta por 1000 pares de oraciones, divididas en los siguientes 4 juicios:

- *Bidirectional:* Ambas sentencias comparten la misma información, hablan del mismo tópico y tienen el mismo sentido.
- *Backward:* La segunda oración (*S2*) contiene más información que la primera (*S1*), hablan del mismo tópico y poseen el mismo sentido.
- *Forward:* La primera oración (*S1*) contiene más información que la segunda (*S2*), hablan del mismo tópico y poseen el mismo sentido.
- *No_entailment:* Las oraciones *S1* y *S2* no hablan del mismo tópico, ni expresan lo mismo.

Para descubrir el juicio de implicación a partir de los resultados obtenidos del algoritmo 1, se aplican las siguientes reglas.

Sea *C* el conjunto de elementos comunes entre las oraciones *S1* y *S2*.

Sea *P1* el porcentaje restante de la oración S1 al eliminar los elementos de *C*.

Sea *P2* el porcentaje restante de la oración S2 al eliminar los elementos de *C*.

Sea *U* un umbral de porcentaje.

El juicio de implicación se obtiene aplicando las siguientes reglas:

Si *P1* > *U* y *P2* > *U*: El juicio es bidireccional, ya que ambas oraciones comparten mucha información.

Si *P1* < *U* y *P2* > *U*: El juicio es *backward*, porque la oración *S2* está contenida en la oración *S1*, y *S1* contiene mucha más información que *S2*.

Si *P1* > *U* y *P2* < *U*: El juicio es *forward*, porque la oración *S1* está contenida en la oración *S2*, y *S2* contiene mucha más información que *S1*.

Si *P1* < *U* y *P2* < *U*: El juicio de *no-entailment*, ya que ambas oraciones comparen muy poca información.

3.3. Conjunto de Datos SICK

Esta colección fue extraída de los datos ofrecidos en la tarea 1 del SemEval-2014 [9], se propone una colección 4,927 pares de oraciones, donde 2,793 tienen el juicio de implicación neutral, 1,414 tienen juicio de implicación *entailment* y 720 tienen juicio de *contradiction*. El juicio de implicación *neutral* significa que ambas oraciones hablan del mismo tópico. Mientras que el juicio de *contradiction* significa que una es la negación de la otra.

Como el juicio de implicación neutral, puede o no incluir implicación textual, consideramos eliminarlo del conjunto de datos, por lo que finalmente se trabaja con los juicios *entailment* y *contradiction*, haciendo un total de 2,134 pares de oraciones.
Para detectar el juicio de implicación se analiza la salida del Algoritmo 1, si la bandera que indica contradicción esta prendida, se dice que ese par de oraciones tiene un juicio de *contradiction*, en caso contrario se aplican las mismas reglas que para el conjunto de datos RTE1.

4. Análisis de resultados

Se realizaron alrededor de 130,000 diferentes experimentos variando el valor de los umbrales de porcentaje U y REL, que se movieron en un rango de 0.05 a 0.95, con un incremento de 0.05. En la Tabla 1 se muestran los resultados obtenidos utilizando las 8 medidas de similitud. La medida propuesta por Wu-Palmer mostró un desempeño superior al resto.

Las medidas de WordNet, que son las seis primeras en la tabla, muestran un comportamiento similar para los 3 corpus. Para el corpus SICK, se obtiene una exactitud mayor que para el resto de los corpus. Este resultado es coherente con las características propias de este corpus, ya que ha sido construido con más cuidado, por ejemplo, en el corpus CLTE se ha detectado que una misma oración está asociada a dos clases distintas. Es importante destacar que el corpus SICK sólo se encuentra clasificado en tres categorías, mientras que el corpus CLTE esta categorizado en cuatro clases.

Analizando los resultados obtenidos se puede observar que las medidas de similitud PMI y LSA ofrecen un grado de exactitud similar, PMI supera a LSA en una centésima. Se puede concluir que las medidas semánticas que ofrece WordNet, en general, superan a las medidas estadísticas, calculadas a través de PMI y LSA. Se debe destacar que las medidas estadísticas se ven afectadas por las características del corpus de donde se obtienen. Se sabe, que para que PMI ofrezca resultados relevantes, se necesita que el corpus sea muy grande y construido utilizando dominios diferentes. El mismo comportamiento ofrece LSA, ya que los conceptos que va a descubrir tienen que encontrarse en el corpus de referencia.

Tabla 1. Resultados utilizando las 8 medidas.

Medidas	**Corpus**		
	RTE1	CLTE	SICK
Path	0.521	0.574	0.652
Leacock-Chodorow	0.521	0.568	0.661
Wu-Palmer	0.525	0.576	0.679
Resnik	0.520	0.554	0.677
Jiang-Conrath	0.520	0.554	0.677
Lin	0.520	0.554	0.677
LSA	0.521	0.566	0.664
PMI	0.521	0.574	0.672

Por otro lado, la buena inferencia de relaciones, está en función de la calidad de los términos anclas, si estos son consistentes se asegura que la relación entre ellos es válida, mientras que si los términos anclas son ruidosos, las relaciones que se logran inferir son invalidas.

4.1. Patrones válidos

A continuación se muestran algunos patrones válidos detectados por la metodología:

- X **take-captive** Y = X **kidnap** Y
- X **peculiar-anthropomorphic-creature** Y = X **stranger** Y
- X **nickname** Y = X **know-as** Y
- X **arrest** Y = X **take-imprison** Y
- X **together-know** Y = X **collective-refer** Y
- X **call** Y = X **well-know** Y
- X **down-payment** Y = X **discount** Y
- X **initial** Y = X **preliminary** Y
- X **associate** Y = X **link** Y
- X **topple** Y = X **overthrow** Y

La metodología propuesta, también permite determinar si las relaciones extraídas no son válidas, esto se hace posible gracias al umbral REL, a continuación se muestran algunos patrones no válidos.

4.2. Patrones no válidos

- X **place** Y != X **consist** Y
- X **shell** Y != X **bomb** Y

- X **europe** Y != X **ban** Y
- X **jail** Y != X **arrest** Y
- X **justice** Y != X **begin** Y
- X **hindi** Y != X **use** Y
- X **get** Y != X **produce** Y
- X **happen** Y != X **center** Y

Se ha detectado que los patrones no válidos tienen una relación directa con la calidad de las anclas, que éstas están en función de los procesos de tokenización y consolidación de términos, que consiste en la generación de *n-gramas* comunes. Es importante destacar que los patrones son extraídos en tiempo de ejecución, directamente del par de texto de los que se desea descubrir el juicio de implicación.

Los resultados obtenidos no pueden ser comparados de manera directa con la metodología propuesta, y que estas no detectan el juicio de implicación a través del uso de plantillas. Las plantillas se han propuesto de manera general, pero no han sido aplicadas sobre ningún corpus. Para visualizar el comportamiento se ha realizado una comparación de los resultados obtenidos por esta metodología con propuestas de los mejores modelos a nivel internacional y de modelos desarrollados por nuestro grupo de investigación.

Todos los algoritmos del estado del arte emplean aprendizaje supervisado, mientras que esta metodología es no supervisada; los algoritmos del estado del arte emplean módulos de normalización, análisis estadístico, análisis semántico y hasta teoría de grafos, mientras que la propuesta introducida se basa en la eliminación de plantillas. Todo esto podría conducirnos a una mala interpretación de los resultados obtenidos, sin embargo en la tabla 2 se ofrece una comparativa de los resultados obtenidos por esta metodología, empleando como medida de similitud a Wu-Palmer, así mismo se muestran los resultados de otras propuestas de solución empleadas por nuestro grupo de investigación.

Tabla 2. Comparación de resultados.

Corpus	Ejecución	Exactitud
RTE1	Manning	0.631
	BUAP-Plantillas	0.525
CLTE	HDU-Run2	0.632
	BUAP-Run2	0.366
	BUAP-Plantillas	0.576
SICK	Illinois-LH_run1	0.845
	BUAP-Run1	0.796
	BUAP-Plantillas	0.679

Analizando la tabla 2, con respecto al corpus CLTE, puede apreciarse que la aproximación utilizando plantillas supera considerablemente a la aproximación enviada al SemEval-2012 (BUAP-Run2), esta aproximación está basada en la selección de características que comparten ambas sentencias [19] y este modelo fue supervisado, es decir se aplicó un modelo de clasificación para detectar el juicio de implicación textual.

A pesar de que esta metodología ha ofrecido mejores resultados no supera a la aproximación HDU-Run2 que ofreció un grado de exactitud del 63%.

La aproximación enviada por nuestro grupo de investigación al SemEval-2014 (BUAP-Run1) aprovecha el tipo de juicio que desea descubrir, ya que construye un conjunto de reglas que le permite prácticamente detectar el juicio CONTRADICTION, y es por ello que el grado de exactitud es prácticamente del 80%. La aproximación utilizando plantillas no incluyó el juicio de implicación neutral, por lo que el resultado obtenido no es comparable.

La metodología propuesta para el Corpus del RTE, no logró detectar más de la mitad de los juicios de implicación *entailment* y *no_entailment*.

5. Conclusiones

Hemos presentado un método para resolver el problema de implicación textual, independientemente del corpus que se está estudiando. La metodología propone inicialmente encontrar anclas entre las dos oraciones, con el objetivo de comparar la similitud entre las relaciones que quedan entre los términos anclas, una vez aplicado el algoritmo de eliminación. Para la búsqueda de la similitud entre las relaciones se propuso utilizar las 6 medidas de WordNet, el Algoritmo LSA y el Algoritmo PMI. Esta propuesta viene dada porque si quedan términos en el medio de las anclas, después de la eliminación, quiere decir que no se pudieron aplicar criterios de sinonimia, pero partimos de la hipótesis que si son términos similares o hablan del mismo contexto, en la taxonomía de WordNet son más cercanos.

Las medidas estadísticas LSA y PMI, no ofrecieron resultados satisfactorios, se considera que esto es producto del corpus que se utilizó para obtenerlas. Se construyó un corpus utilizando EuroParl, Documentación de ayuda de PHP, OpenOffice, novelas extraídas del Proyecto Gutemberg y por último un fragmento filtrado de la Wikipedia. A pesar de que el corpus obtenido es de dominio amplio hay pares de palabras de las que no se pudo detectar el grado de similitud. Es importante destacar que esta metodología no se puede utilizar para el idioma español, pues no se dispone de WordNet en español.

La ventaja de esta propuesta es que propone resolver el problema de detectar el juicio de implicación textual de manera no supervisada y nos permite descubrir patrones de relación, a partir de los términos anclas. Con estos patrones como trabajo a futuro, se puede detectar fácilmente el juicio de implicación.

Referencias

1. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognizing textual entailment challenge. Machine learning challenges: evaluating predictive uncertainty, visual object classification, and recognising textual entailment, pp. 177–190 (2006)
2. Zhao, J., Lan, M., Niu, Z.: ECNUCS: Recognizing Cross lingual Textual Entailment Using Multiple Text Similarity and Text Difference Measures. Atlanta, Georgia, USA, pp. 118–124 (2013)

3. Zhao, S., Lan, X., Liu, T., Li, S.: Application driven statistical paraphrase generation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL, Vol. 2, pp. 834–842 (2009)
4. Kouylekov, M., Magnini, B.: Building a large-scale repository of textual entailment rules. In: Proceedings of LREC (2006)
5. Bos, J.: Is there a place for logic in recognizing textual entailment? Linguistic Issues in Language Technology, Vol. 9 (2013)
6. Harabagiu, S., Hickl, A.: Methods for using textual entailment in open-domain question answering. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 905–912 (2006)
7. Carrillo, M., Vilariño, D., Pinto, D., Tovar, M., León, S., Castillo, E.: FCC: Three Approaches for Semantic Textual Similarity. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics, Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 631–634 (2012)
8. Mihalcea, R., Corley, C. Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. AAAI, Vol. 6, pp. 775–780 (2006)
9. Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R. Zamparelli, R.: A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (2014)
10. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In: Fellbaum, Christiane (ed.), WordNet, pp. 265–283 (1998)
11. Zhibiao, W., Stone, M.: Verb Semantics and Lexical Selection. In: ACL, pp. 133–138 (1994)
12. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453 (1995)
13. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc of 10th International Conference on Research in Computational Linguistics (ROCLING'97), pp. 19–33 (1997)
14. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304 (1998)
15. Zhao, S., Xiang, L. Ting, L. Shen, L.: Application-driven statistical paraphrase generation. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol. 2, pp. 834–842 (2009)
16. Deerwester, S. Dumais, S., Furnas, G., Landauer, T.: LSA: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science (1986-1996), pp. 391 (1990)
17. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognizing textual entailment challenge. Springer Berlin Heidelberg, pp. 177–190 (2006)
18. Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., Giampiccolo, D.: Semeval-2013 Task8: Cross-lingual Textual Entailment for Content Synchronization. In: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013) (2013)
19. Vilariño, D., Pinto, D., León, S., Alemán, Y., Gómez-Adorno, H.: N-gram based Feature Evaluation for the Cross-Lingual Textual Entailment Task. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Vol. 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pp. 124–127, (2013)
20. http://wordnetweb.princeton.edu/perl/webwn

Análisis de la diversidad morfosintáctica en las locuciones verbales

Belém Priego Sánchez

Laboratoire Lexiques, Dictionnaires, Informatique, CNRS (UMR 7187),
Université Paris 13, Sorbonne Paris Cité,
Francia

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

belemps@gmail.com

Resumen. La lengua permanece en constante cambio, disponiendo de una gran cantidad de palabras, combinación y unión de éstas, que permiten expresar un concepto determinado y que son utilizadas en diversos contextos. La fusión de las palabras permite construir frases; la inmensa mayoría de frases están formadas por un verbo y una o varias variables. El verbo exige y realiza una rigurosa selección de los sujetos y de los componentes que pueden acompañarle. A partir del análisis realizado a un tipo particular de frases, en este artículo se presentan los patrones morfosintácticos de las locuciones verbales, cuyo elemento principal es un sintagma verbal y presentan una gran diversidad morfosintáctica. El análisis realizado a las locuciones verbales fue desarrollado para el idioma español.

Palabras clave: Locución verbal, morfosintaxis, patrones morfosintácticos.

1. Introducción

La comunicación es un sistema complejo dado que para llevarla a cabo se necesita de un lenguaje (oral, escrito, señas, braille, etc.) que permita expresar una situación concreta y de manera diferente en cada circunstancia determinada tomando como referente el horizonte cultural compartido. A lo largo de la vida se adquiere, capta y memoriza información que complementa la base de conocimiento léxica personal y que permite desarrollar la habilidad de comunicación. Si se toma como base el lenguaje oral, esta información es representada en forma de frases, expresiones, dichos, la cual permite formular una idea o un concepto y visualizarlas como un todo capaz de expresar variadamente una situación en contextos diferentes. Dicha información, en ocasiones, se llega a plasmar de forma escrita en lenguaje natural (como español, francés, inglés o cualquier otro idioma) en libros, periódicos, revistas u otros textos y en formato digital. Siendo

ésta un recurso importante para la humanidad, la cual al ser almacenada es posible analizarla y obtener resultados benéficos como herramientas que ayuden y faciliten la explotación de esta información y que permitan obtener conocimiento de una lengua a través del acervo lingüístico y cultural que poseen.

La lengua permanece en constante cambio, disponiendo de una gran cantidad de palabras y unión de éstas que a veces no se considera lógica pero que permite expresar un concepto determinado y que no son unidas libremente, sino que son utilizadas y repetidas durante años como bloque de combinaciones. Ya desde la antigüedad, el hombre dedicado a la ciencia sintió un enorme interés por determinadas expresiones cotidianas, las cuales quiso agrupar y estudiar. Este tipo de expresiones que regularmente son fácilmente dominadas por los hablantes nativos de una lengua, plantean un gran desafío para los sistemas computacionales en términos de su interpretación debido a su naturaleza flexible y heterogénea. Estas expresiones no son tan frecuentes en los recursos léxicos como en los textos del mundo real y por lo tanto presentan un gran reto de estudio para diversas tareas de la lingüística computacional.

Este trabajo se enfoca principalmente en un tipo de expresiones denominadas unidades fraseológicas, las cuales no constituyen enunciados completos, tienen la principal característica de ser una combinación estable de dos o más términos y pueden constituir el núcleo de sintagmas verbales. Se analizan expresiones como por ejemplo: *chuparse el dedo, costar un ojo de la cara, dar de sí, mover cielo y tierra, ser el vivo retrato de alguien*, entre otras; es decir, expresiones denominadas locuciones verbales. A partir del análisis realizado se muestran los patrones morfosintácticos de estas construcciones lingüisticas y se identifican en un corpus periodístico.

El presente artículo está organizado de la siguiente manera. La Sección 2 expone de manera general a la unidades fraseológicas, se centra en la definición y características de las locuciones, particularmente de las locuciones verbales. Así como, algunos trabajos sobresalientes encontrados en la literatura. La Sección 3 describe la metodología llevada a cabo para la identificación de los patrones morfosintácticos. La Sección 4 muestra los resultados obtenidos al buscar los patrones en un corpus. Finalmente en la Sección 5 se presentan las conclusiones y perspectivas de este trabajo.

2. Unidades fraseológicas

La fraseología, disciplina lingüística encargada del estudio de la combinación de palabras características de una lengua, ha incrementado su estudio en los últimos años, dado que ha sido un espacio de interés para investigadores atraídos hacia estas construcciones lingüísticas por el hecho de ser un modo particular de habla de los pueblos como material de estudio científico e implica diferentes dimensiones del lenguaje: lingüísticas, pragmáticas, culturales y muchas más. Este hecho refuerza el interés de estudio de este trabajo.

En la lingüística española se emplean diferentes denominaciones a la combinación de estas palabras; a manera de ilustración se citan en este trabajo los

nombres que servirán para dar una idea de la abundante nomenclatura que existe alrededor de estas denominaciones comúnmente identificadas como: unidad fraseológica, fraseologismo, modismos, locuciones, expresiones fijas, dichos, frases hechas, expresiones idiomáticas, expresiones, frases, refranes, proverbios, modos de decir, aforismos, entre otras (en [16] aparecen hasta 64 denominaciones diferentes). En este trabajo se emplea el término *unidad fraseológica* para referirse a las combinaciones de palabras que tienen un significado como un todo, debido a que esta denominación goza de gran aceptación en la filología española, es conocido en la fraseología internacional. Se hace hincapié de la existencia de diferentes tipos de unidades fraseológicas y se remarca que en este trabajo de investigación se analiza solamente un tipo particular de ellas.

Las clasificaciones de las unidades fraseológicas, de manera cronológica, han sido establecida por lingüistas como [6,8,25,27,13,5], entre otros. A partir de ellas, en [7] se propone realizar una clasificación, que parte de [12] y que toma en cuenta algunas características de los trabajos mencionados. Por tanto, en este trabajo, se toma como base esta clasificación que divide en dos grupos a las unidades fraseológicas. El primer grupo, corresponde a las unidades que no constituyen un enunciado completo, las cuales incluyen a las *colocaciones* y a las *locuciones*. El segundo grupo, lo forman las unidades que constituyen un enunciado completo, denominadas como *enunciados fraseológicos*.

Las colocaciones, desde el punto de vista del sistema de la lengua son sintagmas completamente libres que presentan cierto grado de restricción combinatoria determinada por el uso. Éstas son generadas a partir de reglas y generalmente son de base semántica. Algunos ejemplos son: *asumir una responsabilidad, conciliar el sueño, banco de peces, correr un rumor, declararse una epidemia, desempeñar un cargo, estallar una guerra, fuente fidedigna, relación estrecha, negar rotundamente, rematadamente loco, tomar una decisión, una rebanada de pan, viaje relámpago, zarpar un barco*, entre otras.

Las locuciones, son definidas por [6] como una combinación estable de dos o más términos, que funciona como elemento oracional y cuyo sentido unitario consabido no se justifica, sin más, como una suma del significado normal de los componentes. Las diferentes definiciones de locución en español han seguido esta caracterización. Las locuciones se han divido según la función oracional que desempeñen, independientemente de que sean conmutables por palabras simples o por sintagmas. En [7] se distinguen los siguientes tipos:

- Locuciones nominales: *el qué dirán, mosca muerta, patas de gallo, santo y seña,* entre otras.
- Locuciones adjetivas: *corto de medios, de armas tomar, más suave que el algodón, sano y salvo,* entre otras.
- Locuciones adverbiales: *boca con boca, con el corazón en la mano, de par en par, más de la cuenta, por lo pronto,* entre otras.
- Locuciones verbales: *cargársela, dar de sí, meter a alguien en cintura, ir y venir, saber de qué pie cojea,* entre otras.,
- Locuciones prepositivas: *a pesar de, delante de, en lugar de, gracias a,* entre otras.

- Locuciones conjuntivas: *así que, como si, dado que, mientas tanto, puesto que, tan pronto como,* entre otras.
- Locuciones clausales: *como quien dice, como Dios manda, hacérsele a alguien agua la boca, subírsele a alguien, revolvérsele a alguien las tripas, ...).*

Los enunciados fraseológicos, de acuerdo a [14], constituyen cadenas autónomas del habla y, en cuanto a tales, se formulan con entonación independiente, como corresponde a su carácter de unidades mínimas de comunicación. En los enunciados fraseológicos se distingue entre las *paremias* y las *fórmulas rutinarias.* En la primera diferenciación tenemos ejemplos como *errar es humano perdonar es divino, las paredes oyen, poderoso caballero es don dinero, por la boca muere el pez,* entre otras; y en la segunda por ejemplo *a eso voy, ¿qué hay?, ¿qué te digo yo?, ¿puedo ayudarle?, hasta luego, ya lo creo,* entre otras.

En este artículo se decidió trabajar con las locuciones verbales que serán descritas en la subsección 2.1.

2.1. Locuciones verbales

Las locuciones verbales constituyen el núcleo de sintagmas verbales, es decir, están formadas por un núcleo verbal, acompañado por sus complementos. Desde el punto de vista sintáctico expresan procesos y actúan como los predicados, con o sin complementos. Estas unidades fraseológicas, igual que los verbos, se combinan con el sujeto y los complementos para formar una oración.

Las locuciones verbales presentan las mismas características como el resto de las locuciones, las cuales son: fijación interna, unidad de significado y fijación externa parsamática. La fijación es la característica más importante de las locuciones. Así, en las locuciones verbales, la forma fija significa que excepto el verbo, el resto de los elementos que forman parte de la locución no pueden modificarse, tampoco pueden sustituirse por otras palabras ni es posible introducir algo nuevo entre ellos. El verbo que constituye la locución cambia según la persona, el número, el tiempo y el modo de la oración. Sin embargo, la fijación de las locuciones no es absoluta, más bien es relativa y es posible encontrar una locución que tiene dos o más formas en las que es posible sustituir una de las palabras que la constituyen o donde es posible introducir entre las palabras de la locución otra palabra.

En general, las locuciones no constituyen actos de habla ni enunciados, es decir, necesitan combinarse con otros signos lingüísticos y que equivalen a sigtagmas. Estas unidades fraseológicas, no constituyen enunciados completos y, generalmente, funcionan como elementos oracionales. Las locuciones son diferenciadas de las combinaciones libres de palabras de la lengua por su institucionalización, su estabilidad sintáctico-semántica y su función denominativa. La estabilidad es la característica esencial para delimitar unos fenómenos lingüísticos de otros; se incluyen tanto aspectos léxicos-semánticos como los morfosintácticos.

Por lo tanto, existen variadas pruebas para comprobar la cohesión semántica y morfosintáctica. El primer caso, se refleja en el carácter de unidad de significación en la lengua que presentan dichas unidades, ya tengan significado

compositivos (el significado se deduce de acuerdo a sus componentes de la locución) o traslaticio (el significado no se puede deducir de los componentes de la locución). Éste es complementado, con el segundo caso, cohesión morfosintáctica, traducido en la determinadas pruebas y operaciones formales, que comprueban no sólo la estabilidad formal de las locuciones, sino también su integridad semántica. Las principales pruebas aplicadas a los elementos de las locuciones son: a) Sustitución, b) Eliminación, c) Deficiencias transformativas.

Al hacer uso de las locuciones verbales, se debe de tener en cuenta que son propias de determinadas situaciones comunicativas. Esto significa que no todas las locuciones se pueden usar en cualquier situación comunicativa. Existen las que son propias de situaciones en las que existe mucha confianza entre los interlocutores (en los diccionarios se marcan como coloquiales, informales o familiares), por ejemplo: *aguar la fiesta, romper el hielo, ser harina de otro costal, traer cola*, y muchas más. Existen locuciones verbales que no tienen ninguna marca y por eso su uso es más amplio, por ejemplo *abrir paso, dar carpetazo, pasar la factura*, entre otras. Finalmente las locuciones verbales que tienen un significado vulgar, muchas veces contienen vulgarismos y su uso puede expresar la falta de respeto.

Cabe destacar que el uso principal de las locuciones verbales, y de cualquier locución, es su distribución geográfica, distinguiéndose así, locuciones de ámbito general y locuciones locales que se usan sólo en determinadas zonas.

2.2. Trabajo relacionado

Entre los autores interesados en el estudio de las unidades fraseológicas, podemos encontrar por un lado grupos vinculados a corrientes de lingúística teórica [23,24,4] y, por otro lado, corrientes vinculadas a la práctica terminográfica y la estandarización de unidades fraseológicas [3,1]. En las últimas décadas, sin embargo, ambas corrientes comparten el interés por las tecnologías de extracción automática de unidades fraseológicas. A partir del interés por la extracción de estas unidades, algunos autores se han centrado en identificar patrones sintácticos, morfológicos o la mezcla de ambos que ayuden a determinar la estructura interna de esta combinación de palabras.

En [10] se propone una herramienta basada en el etiquetado de las partes de la oración y el alineamiento de palabras para extraer candidatos a unidades fraseológicas y sus traducciones. La lista de unidades candidatas contiene frases de varias palabras y de una sola palabra. Las unidades fraseológicas de varias palabras coinciden con un conjunto de patrones sintácticos definidos por expresiones regulares y se identifican mediante una búsqueda en el documento etiquetado con las partes del discurso.

En [15], la indexación y recuperación de unidades fraseológicas se realiza mediante la combinación de la sintaxis y la morfosintaxis. Esencialmente lo que el sistema hace es tomar como entrada una lista de términos (unidades fraseológicas) y un corpus. La lista es precompilada manual o automáticamente, ésta es extendida detectando todas las variantes del término y marcando sus ocurrencias en el corpus.

En [18] se tiene como objetivo reconocer estas unidades lingüísticas en inglés, asignarles su significado y traducirlas en francés. Este proyecto incluye la identificación de las unidades fraseológicas, la construcción de un recurso léxico y la aplicación en la traducción. El sistema trata de mapear el término de una base de datos léxica que incluye la traducción, la información lingüística y metalingüística de las palabras para que de acuerdo a las partes de la oración, el estilo y dominio de la unidad fraseológica, las restricciones y demás características lingüísticas se identifique en el corpus la unidad y se proporcione su mejor traducción encontrada.

Existen en la literatura más trabajos relacionados con la temática de este artículo, sin embargo, en este trabajo de investigación no se pretende hacer un análisis exhaustivo sino presentar solamente aquellos considerados como los más relevantes partiendo de trabajos seminales. Para el caso del español, algunos trabajos dedicados al análisis morfosintáctico de las locuciones y que determinan diferentes tipos de ellas, se presentan en [6,8,27,13,5,7]. Adicionalmente, en [20,17,2,26,9,11,19] se analizan otro tipo de patrones (semántico, composicional, léxico), con el fin de extraer a estas unidades y determinar las características que podrían generalizarse en estas unidades lingüísticas.

En la siguiente sección se describe la metodología llevada a cabo para el análisis de la diversidad morfosintáctica de las locuciones verbales.

3. Metodología

Con el fin de identificar los patrones morfosintácticos en las locuciones verbales, se parte de la taxonomía, de las locuciones realizada por [7], anteriormente descrita, en su clasificación de las unidades fraseológicas en español. A partir de dicha taxonomía, se decide centrarse en las locuciones verbales, debido a que la mayoría de frases está formada por un verbo y una o varias variables. El verbo exige y realiza una rigurosa selección de los sujetos y de los componentes que pueden acompañarle. Estas frases se encuentran fusionadas en la oración para enunciar algo de manera más amplia, pero al separarse de la oración tienen sentido completo, es decir, tienen información semántica por ellas mismas y constituyen el núcleo de sintagmas verbales.

En dicha taxonomía se clasifican los tipos de locuciones verbales de acuerdo a su variedad morfosintáctica, los cuales comprenden: a) Locuciones formadas por dos núcleos verbales unidos por conjunción, b) Locuciones compuestas de verbo y pronombre, c) Locuciones compuestas de verbo, pronombre y partícula, d) Locuciones de verbo más partícula asociada a éste, con complemento opcional, e) Locuciones formadas por verbo copulativo más atributo, f) Locuciones formadas por verbo más complemento circunstancial, g) Locuciones formadas por verbo más suplemento h) Locuciones formadas por verbo más objeto directo y i) Locuciones negativas. En este trabajo son denominadas como Tipo 1, Tipo 2, ..., Tipo 9; respectivamente. Con base en esta taxonomía, se prosigue a inquirir ejemplos de locuciones verbales que cumplan con la variedad morfosintáctica y que ayuden a determinar los patrones morfosintácticos.

En cuanto a los ejemplos utilizadas, se emplearon las locuciones verbales presentadas en [22], debido a que éstas fueron recuperadas manualmente. Posteriormente, estas locuciones verbales se clasifican de acuerdo a los tipos de la taxonomía empleada según sus componentes. Una vez clasificadas se necesita saber su estructura morfosintáctica para de esta manera obtener los patrones, así que, las locuciones fueron etiquetas con FreeLing[1]. En la tabla 1 se presenta una muestra de las locuciones verbales identificadas de acuerdo a su tipo y sus respectivas etiquetas morfosintácticas[2].

Para la búsqueda de los patrones morfosintácticos identificados, se seleccionó un fragmento del corpus periodístico presentado en [21], el cual contiene aproximadamente 1,960,373 palabras. La identificación de los patrones morfosintácticos en el corpus se ha realizado de dos diferentes maneras, una tomándo en cuenta el contexto y la otra sin tomarlo en cuenta. En la primera aproximación, se ha utilizado una ventana de cinco palabras a la izquierda de la locución verbal y cinco palabras a la derecha, denominándolas *contexto izquierdo* y *contexto derecho*, respectivamente.

Básicamente la metodología propuesta de este trabajo considera tener dos elementos escenciales: 1) Una lista de locuciones verbales, y 2) Un conjunto de textos, ambos etiquetados morfosintácticamente. Del primer recurso léxico se obtienen los patrones morfosintácticos, y éstos son buscados en el corpus de textos con la finalidad de obtener una lista de posibles locuciones verbales, las cuales concuerdan con los patrones morfosintácticos obtenidos de las locuciones semilla (ver figura 1).

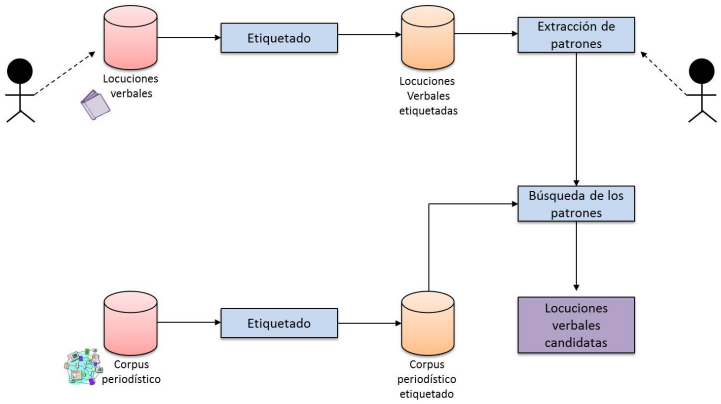

Fig. 1. Metodología empleada para la identificación de patrones morfosintácticos en las locuciones verbales.

[1] Para más información de la herramienta, consultar http://nlp.lsi.upc.edu/freeling/

[2] Para una referencia del significado del etiquetado morfológico de Freeling referirse a http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Tabla 1. Ejemplo de locuciones verbales identificadas de acuerdo a sus etiquetas morfosintácticas.

Tipo de locución verbal	Ejemplos	Etiquetas morfosintácticas (resultados de FreeLing)
Tipo 1	dar y tomar	VMN0000 CC VMN0000
	ir y venir	VMN0000 CC VMN0000
	llevar y traer	VMN0000 CC VMN0000
Tipo 2	apañársela	VMN0000 PP3CN000 PP3FSA00
	arreglársela	VMN0000 PP3CN000 PP3FSA00
	cargársela	VMN0000 PP3CN000 PP3FSA00
Tipo 3	brincarse la barda	VMN0000 PP3CN000 DA0FS0 NCFS000
	darse su taco	VMN0000 PP3CN000 DP3CS0 NCMS000
	tomarla con (alguien/algo)	VMN0000 PP3FSA00 SPS00 (PI0CS000/PI0CS000)
Tipo 4	dar de sí	VMN0000 SPS00 CS
	ir con (uno)	VMN0000 SPS00 PI0MS000
	tomar (algo/a alguien) por	VMN0000 (PI0CS000/SPS00 PI0CS000) SPS00
Tipo 5	ser ajonjolí de todos los moles	VSN0000 AQ0CS0 SPS00 DI0MP0 NCMP000
	ser el vivo retrato de alguien	VSN0000 DA0MS0 AQ0MS0 NCMS000 SPS00 PI0CS000
	ser gacho	VSN0000 AQ0CS0
Tipo 6	decir hasta la despedida	VMN0000 SPS00 DA0FS0 NCFS000
	dormir como un tronco	VMN0000 CS DI0MS0 NCMS000
	meter a alguien en cintura	VMN0000 SPS00 PI0CS000 SPS00 NCFS000
Tipo 7	meter las cuatro	VMN0000 DA0FP0 Z
	oler a cuero quemado	VMN0000 SPS00 NCMS000 VMP00SM
	pagar el pato	VMN0000 DA0MS0 NCMS000
Tipo 8	chuparse el dedo	VMN0000 PP3CN000 DA0MS0 NCMS000
	mover cielo y tierra	VMN0000 NCMS000 CC NCFS000
	saber de qué pie cojea alguien	VMN0000 SPS00 DT0CN0 NCMS000 VMIP3S0 PI0CS000
Tipo 9	no haber vuelta de hoja	RN VMN0000 NCFS000 SPS00 NCFS000
	no poder ver ni en pintura a alguien	RN VMN0000 VMN0000 CC SPS00 NCFS000 SPS00 PI0CS000
	no tener un pelo de tonto	RN VMN0000 DI0MS0 NCMS000 SPS00 NCMS000

4. Resultados

En este trabajo de investigación se han identificado 34 patrones morfosintácticos que sirven como semilla para encontrar posibles locuciones verbales dentro de un corpus de textos. Dichos patrones han sido obtenidos mediante el etiquetado

morfosintáctico de una lista semilla de 43 locuciones verbales. En la tabla 2 se presenta una muestra de los patrones morfosintácticos identificados como más frecuentes en el corpus de textos periodísticos.

Tabla 2. Muestra de patrones morfosintácticos de las locuciones verbales.

Estructura sintáctica	**Patrón morfosintáctico**
V + Prep	VMN0000 SPS00
V + Det + Nom + Adj	VMN0000 DA0FS0 NCFS000 AQ0CS0
V + Pron + Prep	VMN0000 PP3FSA00 SPS00
V + Prep + Conj	VMN0000 SPS00 CS
V + Pron + Det + Nom	VMN0000 PP3CN000 DA0MS0 NCMS000
V + Det + N + Prep + Det + Nom	VMN0000 DI0MS0 NCMS000 SPS00 DA0FS0 NCFS000
V + Det + Nom	VSN0000 DA0FS0 NCFS000
V + Prep + Pron	VMN0000 SPS00 PI0CS000
V + Prep + Pron	VMN0000 SPS00 PI0MS000
V + Pron + Prep	VMN0000 PI0CS000 SPS00

Tabla 3. Ejemplo de las locuciones verbales encontradas en el corpus periodístico.

Frecuencia de aparición	**Locución verbal candidata**
357	llegar/llegar/VMN0000 a/a/SPS00
201	contar/contar/VMN0000 con/con/SPS00
152	participar/participar/VMN0000 en/en/SPS00
117	tratar/tratar/VMN0000 de/de/SPS00
113	apoyar/apoyar/VMN0000 a/a/SPS00
110	cumplir/cumplir/VMN0000 con/con/SPS00
106	salir/salir/VMN0000 de/de/SPS00
99	ir/ir/VMN0000 a/a/SPS00
93	ver/ver/VMN0000 con/con/SPS00
90	acudir/acudir/VMN0000 a/a/SPS00

En la Tabla 3 se presenta un ejemplo de las 10 locuciones verbales encontradas como más frecuentes en el corpus y que empatan con el patrón morfosintáctico indicado en la misma Tabla. En total, se extrajeron 3,083 resultados coincidentes con los patrones registrados.

En la figura 2 se puede observar que de los 10 patrones morfosintácticos más frecuentes (ver Tabla 2), el primero obtiene un 80 % de cobertura con respecto a los demás. Esto se encuentra derivado del hecho de ser un patrón demasiado general que parte de locuciones verbales semilla tales como: “ir con”.

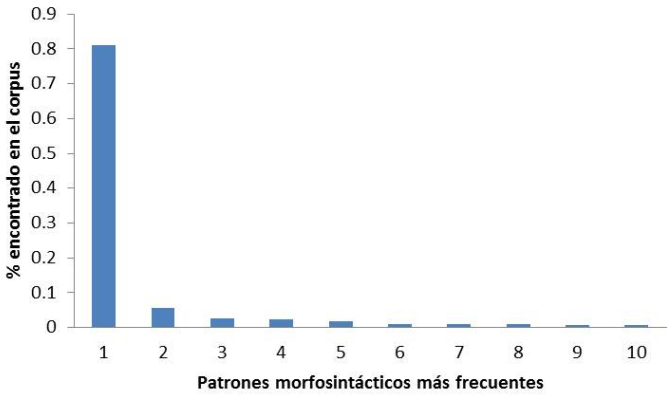

Fig. 2. Porcentaje de las locuciones verbales más frecuentes encontradas en el corpus.

Cabe mencionar que de los 34 patrones morfosintácticos detectados a partir de las locuciones semilla, solamente se encontraron coincidencias sobre 18. Esto significa, que 16 patrones no han arrojado posibles locuciones verbales. En la Tabla 4 se muestran tales patrones; una discusión sobre los mismos sigue a continuación.

Observando los patrones que no encontraron coincidencias en el corpus de textos podemos ver que en general se trata de una secuencia no habitual de etiquetas morfológicas, cuya frecuencia es muy baja en los corpus textuales. El corpus utilizado tiene únicamente 5,000 noticias (361 palabras en promedio por noticia), y por tanto, la probabilidad de encontrar una de estas secuencias es muy baja.

5. Conclusiones y perspectivas

En este trabajo de investigación se presentan experimentos tendientes a la identificación automática de locuciones verbales a partir de textos planos. La metodología propuesta indica tomar como entrada un conjunto de locuciones verbales semilla que sirvan para encontrar un conjunto de patrones morfosintácticos, los cuales son posteriormente utilizados para encontrar coincidencias de los mismos sobre un corpus de textos (en nuestro caso, fue del género periodístico).

El experimento realizado sobre un conjunto inicial de 43 locuciones verbales, permitió encontrar 34 patrones morfosintácticos. De éstos, únicamente 18 encontraron coincidencias en el corpus de textos.

Como trabajo a futuro se desea ampliar el corpus de textos para determinar si es posible encontrar coincidencias de todos y cada uno de los patrones morfosintácticos detectados a partir de las locuciones verbales semilla. Adicionalmente, sería importante evaluar el filtrado de locuciones verbales candidatas para incrementar la precisión en la identificación de las mismas.

Tabla 4. Conjunto de patrones morfosintácticos de los cuales no se encontraron coincidencias en el corpus de textos.

Patrón morfosintáctico
VMN0000 PP3FSA00 SPS00 PI0CS000
VSN0000 DA0MS0 AQ0MS0 NCMS000 SPS00 PI0CS000
VMN0000 Z NCFP000 SPS00 PI0CS000
VMN0000 SPS00 PI0CS000 SPS00 NCFS000
VMN0000 SPS00 PI0CS000 CS VMIP3S0 SPS00 NCMS000
VMN0000 SPS00 DT0CN0 NCMS000 VMIP3S0 PI0CS000
VMN0000 PP3CN000 DA0MP0 NCMP000 SPS00 DA0FS0 NCFS000
VMN0000 DA0FS0 NCFS000 SPS00 NCFS000 SPS00 DA0MS0 VMP00SM
VMN0000 DA0FP0 SPS00 NP00000
RN VMN0000 Z NCMP000 SPS00 RG
RN VMN0000 VMN0000 CC SPS00 NCFS000 SPS00 PI0CS000
RN VMN0000 SPS00 DT0CN0 NCFS000 VMN0000 PP3CN000
RN VMN0000 SPS00 DA0FS0 NCFS000 DA0FS0 AQ0FS0
RN VMN0000 PP3CSD00 CC VMN0000 PP3CSD00 PI0CS000 SPS00 PI0CS000 SPS00 PI0CS000
RN VMN0000 NCMS000 CC VMIP1S0
RN VMN0000 NCFS000 SPS00 NCFS000

Referencias

1. Arntz, R., Picht, H.: Introducción a la terminología. In: Fundación Germán Sánchez Ruipérez. Barcelona (1988)
2. Baldwin, T.: Deep lexical acquisition of verb-particle constructions. Comput. Speech Lang. 19(4), 398–414 (Oct 2005), http://dx.doi.org/10.1016/j.csl.2005.02.004
3. Cabré, T., Estopá, R.: Introducción a la teoría general de la terminología y a la lexicografía. In: Institut Universitari de Lingüística Aplicada. Barcelona (1979)
4. Cabré, T., Estopá, R.: Unidades de conocimiento especializado, caracterización y tipología. In: Cabré, M. T.; Bach, C. (eds.) Coneixement, llenguatge i discurs especialitzat. Barcelona (2005)
5. Carneado Moré, Z., Tristá Pérez, A.M.: Estudios de la fraseología. La Habana: Academia de Ciencias de Cuba. Instituto de literatura y lingüística (1983)
6. Casares, J.: Introducción a la lexicología moderna. In: C.S.I.C. Madrid (1950)
7. Corpas Pastor, G.: Manual de fraseología española. Gredos, Madrid (1996)
8. Coseriu, E.: Structure lexicale et enseignement du vocabulaire. In: Actes du premier colloque international de linguistique apliquée. pp. 175–217 (1966)
9. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions. pp. 25–32. MWE '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1613704.1613708
10. Dagan, I., Church, K.W.: Termight: Identifying and translating technical terminology. In: ANLP. pp. 34–40 (1994), http://dblp.uni-trier.de/db/conf/anlp/anlp1994.html#DaganC94

11. Davis, A.R., Barrett, L.: Lexical semantic factors in the acceptability of english support-verb-nominalization constructions. ACM Trans. Speech Lang. Process. 10(2), 5:1–5:15 (Jun 2013), http://doi.acm.org/10.1145/2483691.2483694
12. Gramley, S., Pätzold, K.M.: A survey of modern English. Londres-Nueva York, Routledge (1992)
13. Haensch, Wolg, G., Ettinger, L., Werner, S.: La lexicografía. De la lingüística teórica a la lexicografía práctica. Gredos, Madrid (1982)
14. Hernando Cuadrado, L.: Sobre las unidades fraseológicas en español. In: Actas de la Sociedad Española de Lingüística. XX Aniversario. vol. 1, pp. 538–546 (1990)
15. Jacquemin, C., Klavans, J.L., Tzoukermann, E.: Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In: Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics. pp. 24–31. EACL '97, Association for Computational Linguistics, Stroudsburg, PA, USA (1997), http://dx.doi.org/10.3115/979617.979621
16. Martins Baltar, M.: La locution entre langue et usages. In: ENS Editions, Fontenay-St. Cloud (1997)
17. McCarthy, D., Keller, B., Carroll, J.: Detecting a continuum of compositionality in phrasal verbs. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. pp. 73–80. MWE '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1119282.1119292
18. Michiels, A., Dufour, N.: Defi, a tool for automatic multi-word unit recognition, meaning assignment and translation selection. In: Proceedings of the first international conference on language resources and evaluation,. pp. 1179–1186 (1998)
19. Nissim, M., Zaninello, A.: Modeling the internal variability of multiword expressions through a pattern-based method. ACM Trans. Speech Lang. Process. 10(2), 7:1–7:26 (Jun 2013), http://doi.acm.org/10.1145/2483691.2483696
20. Piao, S.S.L., Rayson, P., Archer, D., Wilson, A., McEnery, T.: Extracting multiword expressions with a semantic tagger. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18. pp. 49–56. MWE '03, Association for Computational Linguistics, Stroudsburg, PA, USA (2003), http://dx.doi.org/10.3115/1119282.1119289
21. Priego Sánchez, B., Pinto, D., Mejri, S.: Metodología para la identificación de secuencias verbales fijas. Research in Computing Science 85, 45–56 (2014), http://rcs.cic.ipn.mx/2014_85/Metodologia%20para%20la%20identificacion%20de%20secuencias%20verbales%20fijas.pdf
22. Priego Sánchez, B., Pinto, D., Mejri, S.: Towards the automatic identification of spanish verbal phraseological units. Research in Computing Science 96, 65–73 (2015), http://rcs.cic.ipn.mx/2015_96/Towards%20the%20Automatic%20Identi_cation%20of%20Spanish%20Verbal%20Phraseological%20Units.pdf
23. Sager, J.: A practical course in terminology processing. In: Amsterdam/Philadelphia: John Benjamins (1990)
24. Sager, J.: La terminología: representación y comunicación. In: Institut Universitari de Lingüística Aplicada. Barcelona (1999)
25. Thun, H.: Probleme der phraseologie. In: Beihefte zur zeitschrift für romanische philologie 168. Tubinga, Max Niemeyer (1978)
26. Zhang, Y., Kordoni, V., Villavicencio, A., Idiart, M.: Automated multiword expression prediction for grammar engineering. In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. pp. 36–44. MWE '06, Association for Computational Linguistics, Stroudsburg, PA, USA (2006), http://dl.acm.org/citation.cfm?id=1613692.1613700

27. Zuluaga, A.: La función del diminutivo en español. In: Thesaurus XXV. pp. 23–48 (1980)

Metodología basada en grafos para la identificación de perfiles de usuario

Patricia Espinoza, Darnes Vilariño, David Pinto,
Josefa Somodevilla, Mireya Tovar

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla,
México

{patricia.efong,mariajsomodevilla}@gmail.mx, {darnes,dpinto,mtovar}@cs.buap.mx

Resumen. En la presente investigación se propone un modelo para la identificación de perfiles de usuario, a través de la creación y análisis de un grafo de co-ocurrencia. Se utilizan 4 corpus en Inglés: de Blogs, de Redes sociales, de Críticas y de Twitter y 2 corpus en Español: de Blogs y de Críticas para el desarrollo de los grafos. Para la creación y extracción de la información del grafo se han utilizado las herramientas NetworkX[1] (creación del grafo) y Gephi[2] (extracción de características del grafo). En general el corpus de Blogs en el idioma Español fue el que presentó los mejores resultados.

Palabras clave: Perfil de usuario, grafos de co-ocurrencia, medidas de centralidad.

1. Introducción

En el mundo actual, se generan contenidos electrónicos de todo tipo, todos los días. Blogs, Twitter, Facebook, son algunas de las plataformas mas comunes para compartir textos de algún tema en particular. Suponiendo que alguien quisiera analizar esos textos para determinar alguna característica en particular o común entre ellos, sería casi imposible, debido al volumen de información que existe actualmente. Por esta necesidad de automatización surgen tareas enfocadas al análisis de los textos, la que se aborda en esta investigación es la tarea de identificar el perfil de un autor de manera automática.

Dicha tarea, se basa en encontrar patrones de escritura entre diferentes grupos, los cuales pueden incluir el género, la edad, el lenguaje nativo y la nacionalidad, entre otras cosas. Esta tarea ha ganado gran relevancia debido a las aplicaciones que se le pueden dar, por ejemplo en análisis forenses, en seguridad y hasta en mercadotecnia.

El enfoque principal de esta investigación, es determinar correctamente el género (female, male) y el rango de edad (18-24, 25-34, 35-49, 50-64, 65+) del

[1] https://networkx.github.io/
[2] http://gephi.github.io/

autor de un documento dado. Para cumplir este objetivo se desarrolló un modelo de aprendizaje automático a partir del análisis de grafos de co-ocurrencia que permite encontrar aspectos relevantes de cada documento.

Los documentos son extraídos de 4 corpus en Inglés: de Blogs, de Redes sociales, de Críticas y de Twitter y 2 corpus en Español: de Blogs y de Críticas. Dichos corpus fueron obtenidos de la conferencia internacional PAN 2014[3].

La estructura del artículo es la siguiente. En la sección 2 se presentan los trabajos desarrollados en la literatura con respecto al uso de grafos para diferentes problemas de clasificación. La sección 3 presenta la descripción del el modelo de clasificación. La discusión acerca de los resultados obtenidos se presenta en la sección 5. Finalmente la conclusión del presente trabajo de investigación se realiza en la sección 6.

2. Estado del arte

Se realizó un estudio sobre los trabajos desarrollados en esta área, enfatizando sus avances y el tipo de diseño que se utiliza al momento de crear los grafos en diferentes tareas, así como sus aportaciones científicas, encontrando el siguiente panorama general:

En el trabajo desarrollado por [7], lo que se busca es realizar consultas sobre una base de datos de grafos indexados, para esto, la representación de los grafos se hace por medio de un código o *canonical label* al que llaman *DFS Code*, si dos grafos son iguales entonces comparten el mismo código. Dicho código es generado al realizar una búsqueda en profundidad en el grafo.

De igual forma en [4] proponen un método para representar una imagen de manera formal, la cual consiste en un conjunto de objetos con propiedades y relaciones. Se busca hacer la representación a través de un grafo etiquetado dirigido, el problema que se aborda es el de cuales propiedades seleccionar para la construcción del grafo. En esta aproximación los objetos son representados por los nodos, y las relaciones y propiedades son las aristas.

En [1] el objetivo de los autores es diseñar un motor de búsqueda que haga uso de la estructura de los hiperenlaces de la Web para encontrar sitios web de interés. Este motor de búsqueda es capaz de encontrar no solo palabras clave o de algún tema en particular, si no que puede buscar un hiperenlace con una estructura deseada. En ese grafo cada URL representa un vértice etiquetado como '_page_', las aristas están etiquetadas como '_hyperlink_' y apuntan de una URL padre a un URL hijo. También se hace un análisis del texto de cada página, se eliminan signos de puntuación, palabras cerradas, etiquetas HTML y todas las palabras restantes se agregan al grafo como un nodo nuevo etiquetado con la palabra y se relacionan con la página correspondiente (nodo '_page_') por medio de una arista etiquetada como '_word_'.

Otro trabajo que utiliza grafos para representar la información es presentado en [2], donde el problema a resolver es la correferencia de entidades. Una entidad

[3] http://pan.webis.de/

es un objeto o un conjunto de objetos del mundo real y una mención es una referencia textual a una entidad. El objetivo de este trabajo es identificar a que entidad hace referencia una mención, para esto utilizan una representación del espacio de correferencia mediante un grafo no dirigido, en donde los nodos representan todas las menciones en el texto y las aristas relacionan a los nodos que se refieren a la misma entidad. Cada arista tiene un peso asignado, el cual representa el grado de confianza de correferencia entre esos nodos.

En los trabajos [3] y [6] también se busca resolver el problema de correferencia. Ambos de igual forma que en el trabajo anterior, crean un grafo donde los nodos son las menciones y las aristas modelan una relación entre esas menciones. Cada arista tiene un peso asignado y en cada trabajo se utiliza un método específico para calcular ese peso.

Por último en [5] se busca hacer un anális del significado de un texto mediante una representación de ese texto en un grafo dirigido, en el cual las palabras del texto se representan por los nodos y las relaciones entre las palabras se representan por las aristas. Un punto interesante de este trabajo es que se crean aristas entre las palabras que están directamente conectadas (una detrás de otra), pero también se conectan palabras que están separadas por un número de palabras definido, para que las palabras que son usadas dentro de un mismo contexto estén conectadas.

Las investigaciones revisadas demuestran que la representación de los textos mediante grafos se está utilizando en la actualidad, ya que existen diversas herramientas que permiten crear grafos con un gran número de nodos y aristas, además de que los grafos logran representar de manera correcta diferentes niveles del lenguaje. Todos estos trabajos nos sirven como referencia y línea base para crear un modelo efectivo, pero es importante destacar que no importa el modelo que se esté evaluando, siempre va a ser más simple detectar el género, que la edad, pues los hombres y las mujeres escriben o se interesan por temas diferentes independientemente de la edad que tienen. Un aspecto importante a estudiar es la técnica de clasificación que se debe usar y su comportamiento frente a los modelos en los que se aplique.

3. Metodología

A continuación se presenta la metodología desarrollada.

3.1. Preprocesamiento del corpus

Debido a que el corpus con el que se trabaja es descargado directamente de la página del PAN, es necesario varias operaciones antes de trabajar con él, algunas de ellas son:

1. Separar el corpus por autor.
2. Separar el corpus por género.
3. Sustituir los símbolos HTML que pueda contener el texto, por su equivalente en utf8.

Para el último punto se desarrolló un diccionario de símbolos HTML, el proceso se puede observar en la figura 1.

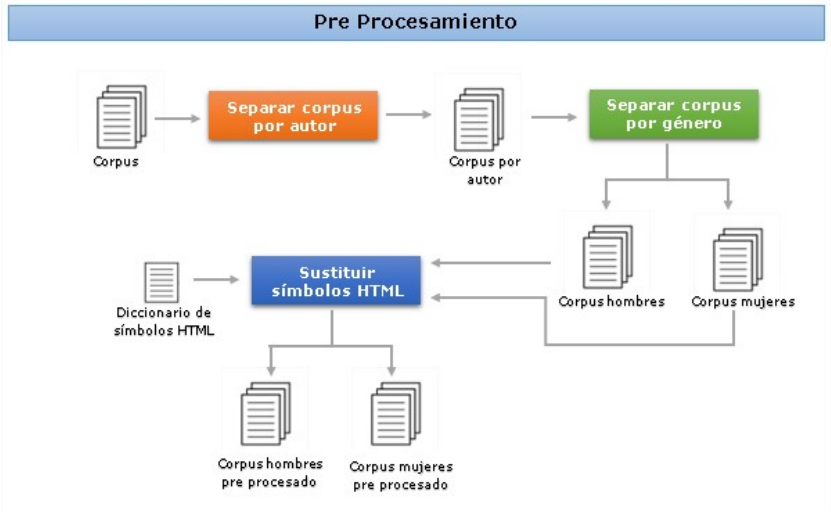

Fig. 1. Preprocesamiento estándar del corpus.

El preprocesamiento estándar se hace para limpiar el corpus de manera general, posteriormente se remueven del corpus las **palabras cerradas**(artículos, conjunciones, verbos auxiliares, etc) , ya que son las que se utilizan con más frecuencia, pero en realidad no aportan significado o no cambian el contexto del texto. La detección de las palabras cerradas se hace a partir de un diccionario en Inglés y otro en Español.

En el tercer paso se sustituyen las palabras restantes en el texto por su correspondiente lema, esto se hace con el objetivo de simplificar y hacer más eficientes los procesos posteriores, ya que se reduce la complejidad de la red resultante, disminuyendo el tamaño del vocabulario. Para realizar este proceso se utilizó la función parse[4] que viene dentro de la librería de Clips utilizada en aproximaciones anteriores.

Como último paso se eliminan los signos de puntuación, los números y se lleva todo el texto a minúsculas (lo que evita que una misma palabra sea considerada como dos palabras diferentes). Todo este proceso se refleja en la figura 2.

Un fragmento del texto resultante se puede observar a continuación:

```
currently see wave idea datum center throw traditional model
datum center management air accelerate demand process datum
storage capacity globally come together environmental demand
create area.
```

[4] http://www.clips.ua.ac.be/pages/pattern-en

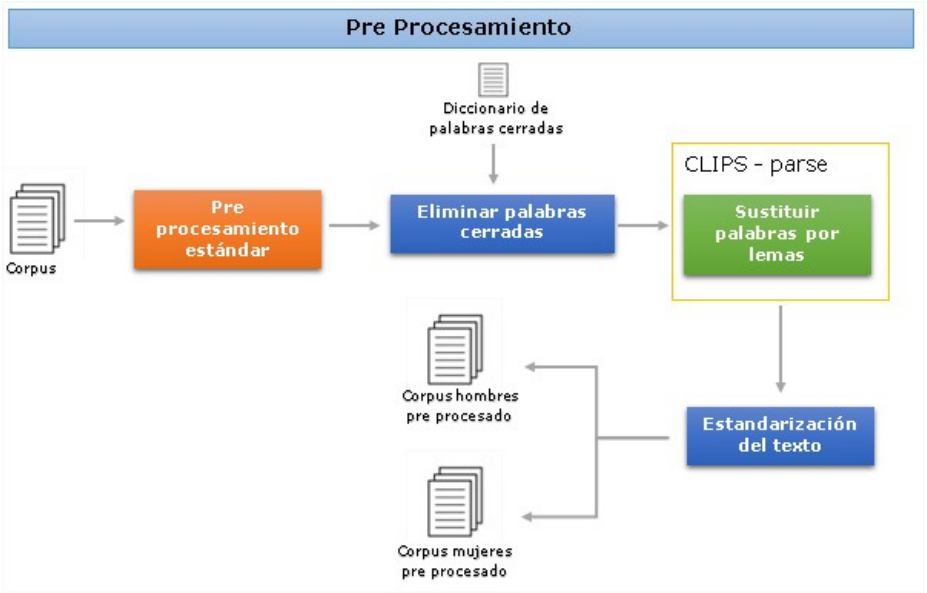

Fig. 2. Preprocesamiento del texto para la creación del grafo.

3.2. Creación del grafo

Despúes de realizar el preprocesamiento de los corpus, el siguiente paso es usar el texto resultante para crear un grafo de co-ocurrencia. Este tipo de grafos se ha convertido en una de las formas más simples y efectivas de representar las relaciones entre las palabras, ya que su implementación es muy fácil de realizar.

Se dice que dos palabras co-ocurren si entre ellas se encuentra un número fijo de palabras, a esto se le llama ventana. En este caso se utilizaron dos tipos de ventanas: una para relacionar los términos que están uno junto al otro (ventana de 0), y otra para relacionar palabras dentro de una ventana igual a 3. El objetivo de la segunda ventana es el de reforzar la relación entre palabras que ocurren en contextos similares.

Formalmente, un grafo de co-ocurrencia dirigido G es un par ordenado G=(V,E), donde:

- V : Conjunto de vértices o nodos los cuales representan las palabras del texto.

- E : Conjunto de pares ordenados de elementos de V que representan la relación entre estos nodos:

En la figura 3 se puede observar el grafo para la siguiente oración: *"currently see wave idea datum center throw traditional model datum center management*

air accelerate demand process datum storage capacity globally come together environmental demand create area"; Se muestran las relaciones que se crean entre las palabras no secuenciales.

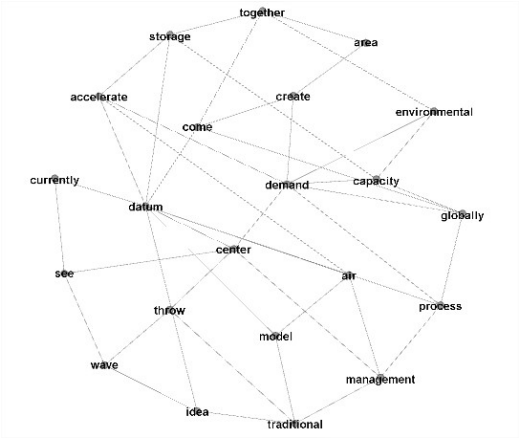

Fig. 3. Grafo de co-ocurrencia.

El proceso para la creación del grafo se puede observar en la figura 4. Se desarrolló un grafo por género {female, male}, se separó el corpus por grupos de edad y se creó un grafo por cada grupo de edad, este proceso se realizó por cada corpus en Inglés y en Español. Al final se obtuvo un total de **72** grafos, los cuales se guardan en un formato xml, para posteriormente crear una representación visual del mismo por medio de Gephi y calcular las medidas de centralidad deseadas.

3.3. Extracción de las carácteristicas del grafo

El desarrollo de grafos de co-ocurrencia permite extraer las palabras relevantes dentro del texto, por medio de las medidas de centralidad y de modularidad, disponibles dentro de la herramienta de análisis de grafos Gephi. Estas medidas se explican a continuación:

- **Interconectividad (Betweenness centrality)**: es un indicador de la centralidad de un nodo dentro de la red. Es igual al número de veces que se pasa por ese nodo para llegar a otros nodos en el grafo o dicho de otra forma, es el número de veces que un nodo aparece al calcular el camino más corto de los otros nodos en la red. Los nodos con una interconectividad alta se pueden decir que son los que tienen mayor influencia dentro de la red, ya que son capaces de representar el contexto en el que se encuentra una cierta palabra.

En la figura 5 se tiene un ejemplo de esta medida, utilizando el mismo grafo de la sección anterior, pero ahora el tamaño de los nodos está dado por el grado

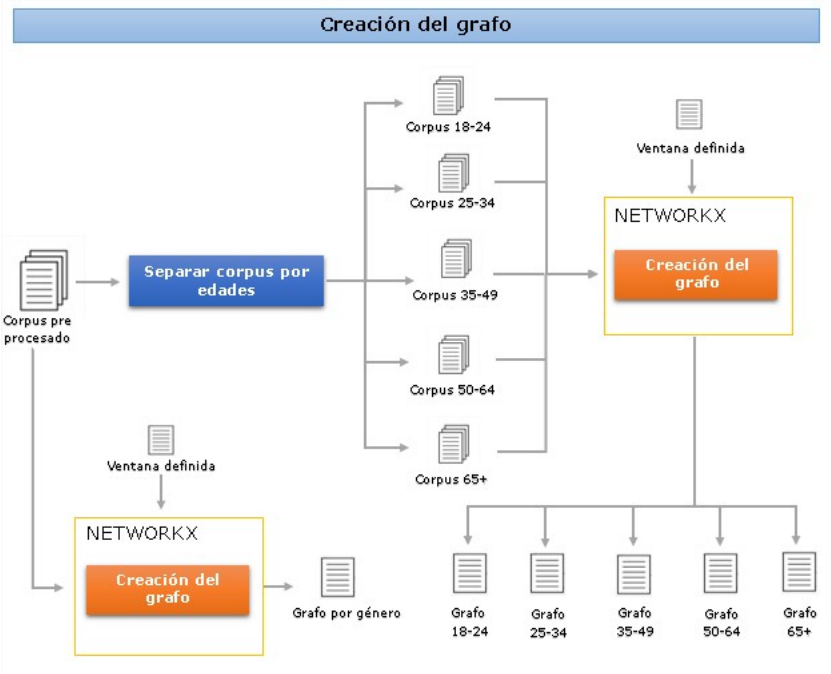

Fig. 4. Creación del grafo.

de interconectividad, fácilmente se puede observar que los más grandes son los más interconectados ya que conectan los dos extremos del grafo.

- **Modularidad (Modularity)**: Mide la fuerza con la que se divide una red en módulos (grupos, clusters o comunidades). Los grafos con gran modularidad tienen conexiones densas entre los nodos que se encuentran en el mismo módulo y conexiones escasas entre nodos de otros módulos. Para esta tarea, nos ayuda a encontrar palabras que se relacionan en torno a un tema dado.

Siguiendo el mismo ejemplo, en la figura 6 se muestran por color los clusters en los que se agrupan los nodos del grafo, el tamaño de los nodos está dado por la medida anterior.

El proceso para el análisis del grafo se puede observar en la figura 7. Se recibe el grafo en formato XML y se calcula el grado de interconectividad entre los nodos, esta herramienta permite observar visualmente los nodos con mayor interconectividad del grafo, ya que pueden ser filtrarlos por tamaño y color.

La segunda medida que se calcula es la modularidad, para que agrupe los nodos por comunidades y se puedan distinguir cada comunidad con un color. Al final lo que interesa es obtener una lista de palabras, en donde cada palabra tenga 2 medidas, el grado de interconectividad y la comunidad a la que pertenece.

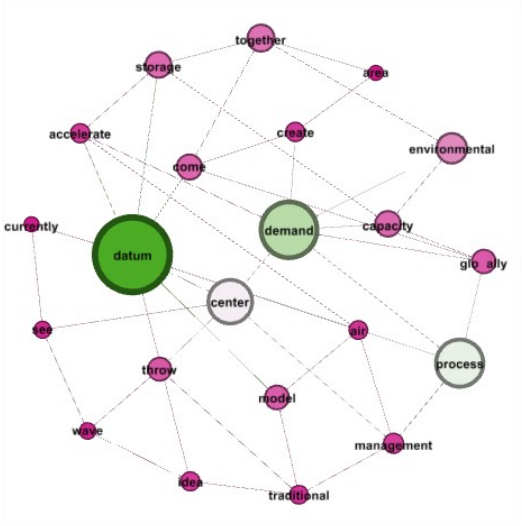

Fig. 5. Ejemplo de interconectividad.

3.4. Proceso de clasificación

Se desarrolló un modelo supervisado el cual se puede observar en la figura 8. Como primer paso se tiene el preprocesamiento, que se realiza para preparar los corpus, posteriormente la creación y análisis del grafo. Después se seleccionan las características o palabras relevantes para ese corpus y esa clase, y se realiza un conteo de las veces que aparece cada palabra en cada documento. También se utilizan todas las comunidades resultantes del análisis y cada vez que se cuenta una palabra, se incrementa el valor de la comunidad o comunidades a las que pertenece.

Se genera un vector por cada documento, donde la longitud de éste es igual al número de palabras elegidas más el número de comunidades . Cada posición del vector corresponde al número de veces que aparece esa palabra en el documento y en el caso de las comunidades, corresponde al número de palabras que pertenecen a esa comunidad en el documento. El atributo clasificador corresponderá al género del autor. Una vez que se han construido los vectores se utiliza como clasificador a la máquina de soporte vectorial para crear el *Modelo de clasificación por género*.

Posteriormente se separan por género los vectores y se les asigna el atributo clasificador correspondiente al rango de edad del autor. Aquí se crean dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre*. Para que a cada modelo solo entren vectores que correspondan a ese género.

En la fase de pruebas se realiza el mismo proceso para crear los vectores con los documentos de prueba que son evaluados con los modelos construidos.

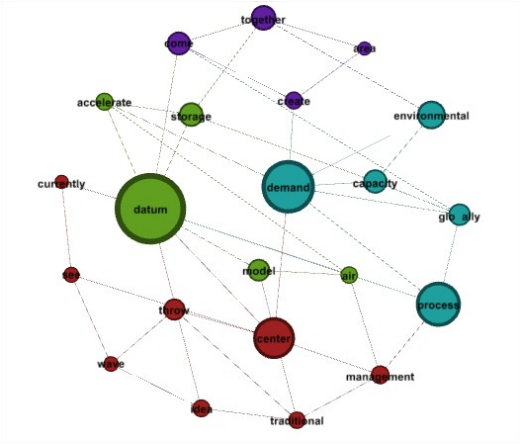

Fig. 6. Ejemplo de modularidad.

4. Descripción de los experimentos

Para estos experimentos se tomaron varios conjuntos de palabras para analizar el comportamiento del clasificador, para cada clase {female, male} de cada corpus y cada experimento se probó por documento y por autor. Con estos conjuntos de palabras de crearon los modelos para clasificar los documentos por género y por edad, los experimentos se explican en detalle a continuación:

- **Experimento 1**: Se escogieron las 1000 palabras con mayor interconectividad (véase el Capítulo 3.3) de cada clase {female, male}.
- **Experimento 2**: Se tomaron todas las palabras del vocabulario de cada clase, excluyendo las que tienen una interconectividad igual a cero.
- **Experimento 3**: Se excluyeron las que tienen una interconectividad igual a cero. Se dividió el total de palabras entre 2 y se tomó mil palabras arriba de la mitad y mil palabras abajo de la mitad, un total de 2000 palabras por clase.
- **Experimento 4**: Se excluyeron las que tienen una interconectividad igual a cero. Se calculó el promedio de la interconectividad de cada palabra y se tomó mil palabras arriba del promedio y mil palabras abajo del promedio, un total de 2000 palabras por clase.

Los experimentos 3 y 4 se realizaron con la hipótesis de que las palabras con mediana interconectividad serían más representativas de su clase, ya que se repitieron menos las palabras entre las clases, a comparación de los experimentos anteriores.

Por último se realizaron 2 experimentos más, pero ahora específicamente para crear un modelo para calcular la edad de los autores de los documentos. Para esto se crearon 10 grafos adicionales por cada corpus, se tienen dos clases para el género {female, male} y 5 clases para la edad {18-24, 25-34, 35-49,

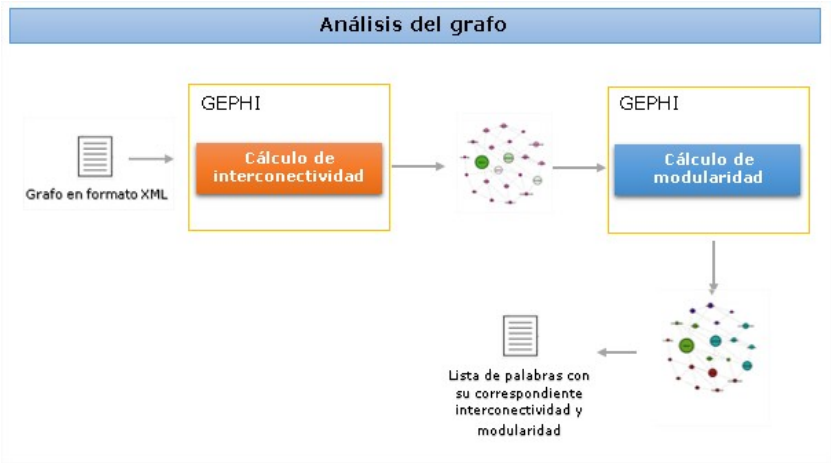

Fig. 7. Análisis del grafo.

50-64, 65+}, posteriormente se creó un grafo por cada clase género-edad (female-18-24, female-25-34, etc). Obteniendo como resultado 5 conjuntos de palabras con su respectiva interconectividad por cada género, para entrenar cada modelo edadHombre y edadMujer (véase 3.4) se utilizaron instancias correspondientes al género del modelo que se entrenó.

- Experimento 5: Se escogieron las 1000 palabras con mayor interconectividad de cada clase (female-18-24, female-25-34, etc), con un total de 5000 palabras para cada modelo.
- Experimento 6: Se escogieron las 1000 palabras con mayor interconectividad de cada clase como en el experimento anterior, pero se observó que las clases que más se confunden entre ellas son: 25-34, 35-49 y 50-64. Debido a esto se decidió tomar las siguientes mil palabras con mayor interconectividad de estas clases en particular, 1000 palabras para las clases 18-24 y 65 y 2000 palabras para las clases mencionadas anteriormente dio un total 8000 palabras para cada modelo.

A continuación se muestran los resultados de los experimentos para cada corpus, en **negritas** los mejores resultados de cada tipo de experimento (por autor o por documento). Debido a que los Experimentos 5 y 6 se diseñaron para calcular la edad, no aplican las pruebas sobre el corpus por género, esto se indica con N/A.

5. Resultados

En el siguiente cuadro se muestra un resumen con los mejores resultados de los experimentos de cada corpus, se muestra en **negritas** los mejores resultados

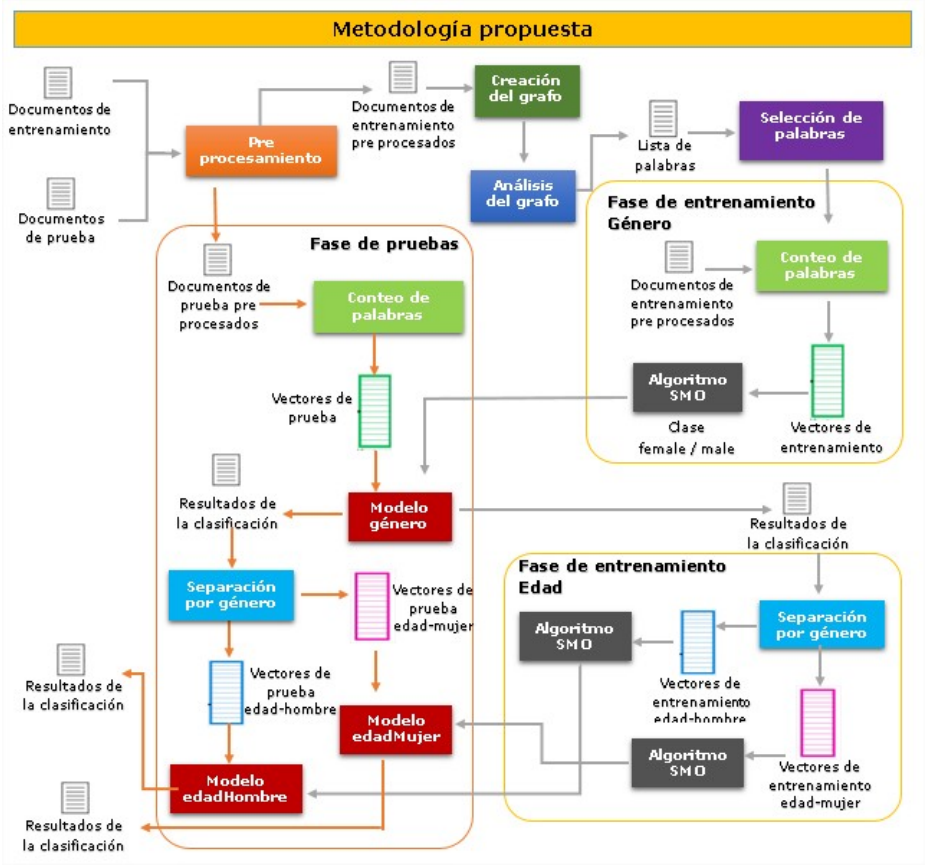

Fig. 8. Metodología para el modelo creado a partir de Gephi.

por idioma. La clasificación se realizó con el algoritmo máquinas de soporte vectorial (SMO) implementado en weka.

Se puede observar en el cuadro 1 que los mejores resultados los obtuvieron los corpus de blogs, pero en general el corpus de blogs en español fue el que presentó los mejores resultados, tanto para el género como para la edad. Otro detalle importante a resaltar es que para la edad, el experimento con mejor desempeño para casi todos los corpus fue el número 6 y para el caso del género fueron el 2 y el 1.

6. Conclusiones

Se desarrolló un modelo para la detección del perfil de un autor (género y edad) mediante grafos de co-ocurrencia. Se pudo observar que el comportamiento

Tabla 1. Resumen de la primera aproximación para ambos idiomas.

Número de características	Tipo de clasificación	Tipo de organización	Presición
INGLÉS			
Blogs			
Experimento 2	Por género	Por documento	**80.76**
Experimento 6	Por edad (mujeres)	Por documento	**67.58**
Experimento 6	Por edad (hombres)	Por documento	**73.36**
Reviews			
Experimento 1	Por género	Por autor	66.82
Experimento 5	Por edad (mujeres)	Por autor	33.89
Experimento 6	Por edad (hombres)	Por documento	31.63
Socialmedia			
Experimento 1	Por género	Por documento	62.39
Experimento 6	Por edad (mujeres)	Por documento	57.67
Experimento 6	Por edad (hombres)	Por documento	56.85
Twitter			
Experimento 4	Por género	Por autor	72.1
Experimento 1	Por edad (mujeres)	Por documento	61.47
Experimento 5	Por edad (hombres)	Por documento	70.61
ESPAÑOL			
Blogs			
Experimento 2	Por género	Por documento	**84.79**
Experimento 6	Por edad (mujeres)	Por Documento	**74.92**
Experimento 2	Por edad (hombres)	Por Documento	**84.24**
Socialmedia			
Experimento 1	Por género	Por autor	63.67
Experimento 5	Por edad (mujeres)	Por autor	47.64
Experimento 6	Por edad (hombres)	Por documento	41.36

del modelo fue un poco diferente para ambos idiomas, superando los resultados de blogs en Español al de blogs en Inglés, siendo lo opuesto en el caso de socialmedia.

En el caso de los experimentos se puede concluir que la idea de realizar grafos por edad (Experimento 5 y 6) fue buena, ya que estos fueron los que mejor resultados brindaron para casi todos los casos. Y para el caso del género, las mil palabras con mayor interconectividad (Experimento 1) fueron las que mejor comportamiento tuvieron. Con los resultados obtenidos se puede afirmar que la confeccíon de grafos de co-ocurrencia para seleccionar caraterísticas para construir modelos de clasificación permite el desarrollo de modelos efectivos, ya que en la mayoría de los experimentos superan las precisiones reportadas en la literatura.

Referencias

1. Cook, D., Manocha, N., Holder, L.B.: Using a graph-based data mining system to perform web search|. International Journal of Pattern Recognition and Artificial Intelligence 17(705) (2003)
2. Cristina, N., Gabriel, N.: Bestcut: a graph algorithm for coreference resolution. In: EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 275–283 (2006)
3. Jie, C., Michael, S.: End-to-end coreference resolution via hypergraph partitioning. In: COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics. pp. 143–151 (2010)
4. Krahmer, E., Verleg, A., Erk, S.: Graph-based generation of referring. In: Computational Linguistics archive. pp. 53–72 (2003)
5. Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. Nodus Labs (2011)
6. Vincent, N.: Graph-cut-based anaphoricity determination for coreference resolution. In: NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 575–583 (2009)
7. Yan, X., Yu, P.S., Han, J.: Graph indexing: A frequent structure-based approach. In: SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data. pp. 335–346 (2004)

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27, Centro Histórico, México, D.F.
octubre de 2015
Printing 500 / Edición 500 ejemplares