Comparison of Automatic Keyphrase Extraction Systems in Scientific Papers

Jesús Ernesto Padilla Camacho, Yulia Ledeneva, René Arnulfo García Hernández

Autonomous University of the State of Mexico,
State of Mexico, Mexico

jernestopc1@gmail.com, yledeneva@yahoo.com, renearnulfo@hotmail.com

Abstract. Nowadays the amount of digital information that found in internet has considerably increased that is why online search is needed to automatically find the corresponding documents. These documents must be verified in order to know whether they contain the required information. A way to simplify the online search is using keywords or keyphrases since they act as filters within a search field. The paper presents the comparison of automatic keyphrases extraction systems based on a collection of scientific papers used in task 5 of SemEval-2010 which calls "Automatic keyphrase extraction from scientific articles". In the experimental section, the results are presented for installable and online systems. We found systems that can match better the author-, reader-, and combined-assigned keyphrases with the keyphrases proposed by an expert. Finally, the obtained results are compared to the results obtained in task 5 of SemEval-2010.

Keywords: Automatic keyphrases extraction, task 5 SemEval-2010, KEA, Alchemy, Wordstat, Extractor.

1 Introduction

At present time the usage of data is a factor of great importance in the public and private sectors. With the constant increase of digital information it needs to be organized for the usage. Nowadays with the technology advances, the searching of information has been facilitated. The keyphrases helps in the information retrieval task because they are very useful for searching information in a big data collection and act as filter to show the most important topics that are described by the author. The keyphrases are the union of words that represent the main ideas of text and provide a brief perception of its content [1-5].

The automatic selection of keyphrases that best describe a text is called Automatic Keyphrase Extraction (AKE). That is to say, the AKE is responsible to perform all the process that is made by a professional indexer, since executing the process automatically reduces factors as the cost of hiring an expert on the subject and the time involved. Witten [3] mentions that the keyphrases usually are elected manually in many academic contexts. The authors assign keywords at the documents they write. The professional indexers usually elect phrases from a "controlled vocabulary" that is relevant for the domain. However, the great majority of documents come without

keyphrases, and manually assigning it is a tedious process that requires a specialized knowledge.

The organization of paper is as follows: in section 2, the related works to AKE are mentioned. In the section 3, the proposed frame work for evaluating AKE systems is described. In the section 4, the dataset, the evaluation tool and the results of experimentation are presented. In the section 5, the conclusions of paper are presented.

2 Related Work

In 2010, Kim et al. [1, 2] perform the shared task "Task 5: Automatic Keyphrase Extraction from Scientific Articles" that was included in the SemEval-2010. The purpose was to develop AKE systems from scientific papers and compare the list of keyphrases proposed by each competitor system with the keyphrases that were assigned by humans to each of the scientific papers. The system that won the best result in the task was the system HUMB [7] with F-score of 27.5% in the combined-assigned configuration.

HUMB is supervised approach that analyzes the structure of document (abstract, conclusions, and references). The selection of candidates that implements is the extraction of n-grams up to 5 words, elimination of candidates that started or terminated with stopwords, filtering of mathematical symbols. The classification of candidates is done by a decision tree. Also the terminology databases GRISP [8] and Wikipedia [9] are used.

Nguyen [10] participates in the task 5 SemEval-2010 with the system WINGNUS. WINGNUS is a supervised approach, one of the main characteristics that employ for the keyphrases extraction is the logic structure of document, to make less text to analyze. The sections are identified where is the most probable is to find the keyphrases. They consider that these sections are abstract, introduction and conclusions. For the classification of candidates employs 19 syntactic functions of which the best result is obtained with the functions such as: *tf* x *idf*, term frequency, substrings frequency, first occurrence and length of the phrase.

El-Beltagy [11] participates in the task 5 SemEva-2010 with the system KP-miner. KP-miner is an unsupervised approach that extracts keyphrases from text in Arab and English. The process consists of three steps: 1.- Selection of candidates where the words are filtered which are not separated by punctuation signs or stopwords, also the frequency of phrase and first occurrence is included. 2.- Weight calculation: term weight, term frequency, IDF weighting, increase factor and term position. 3.- Selected list of final candidate for keyphrases: this is an optional characteristic of the system to refine the candidates.

Bernend [12] participates in the task 5 SemEvla-2010 with the system SZTERGAK. SZTERGAK is a supervised approach. The selection of candidates that employs is the extraction n-grams up to 4 words, the characteristics are grouped in four categories: 1.- Sentence level (length word and POS pattern). 2.- Document level (such characteristics as: acronomity, PMI Sintactic). 3.- Corpora level (*tf-idf* and *keyphraseness*). 4.- External knowledge: use of Wikipedia.

Pianta [13] participates with the system KX. This it is an unsupervised approach. KX employs four steps for the selection of candidates to extract n-grams up to 4 words:

three at corpora level and one extracts the specific document information. For the classification of candidates employs the next characteristics: *idf*, length phrase, position of first occurrence, subsumption and boosting.

The state-of-the-art of AKE systems that not included in task 5 are presented below.

Witten et al. [3] create an algorithm which calls KEA. The proposed algorithm is a supervised approach uses the technique of *Naive Bayes*, which from training data creates a training model that can extract the keyphrases of new documents. KEA employs 2 characteristics: *tf-idf* and first occurrence of phrase.

Turney [5] presents the results of comparison between an extraction model based in a genetic algorithm and an implementation of C4.5 decision trees. Turney informs that genetic algorithm issues better keywords than decision trees.

Mihalcea [6] presents a classification model based on an unsupervised graph that uses the co-occurrence and relation between words that added to the graph to give weights to the vertices. TextRank perform two tasks inside of the information retrieval that are: keyphrase extraction and keyphrase extraction for text summarization.

Medelyan [14] presents Maui, this is a variant of KEA. This is a supervised algorithm for the automatic indexing, uses semantic information extracted from Wikipedia which uses external resources to obtain the best keyphrase extraction based in the titles from Wikipedia.

3 Framework

The dataset of task 5 of SemEval2010 is used. First, the pre-processing is applied. Second, the AKE based on the standard configuration of parameters is performed. Third, Porter stemmer algorithm [15] is applied to obtain the evaluation format. Forth, the evaluation is performed with the tool that evaluates the results, same that it is used in the task 5 of SemEval-2010 (performance.pl). Finally, the systems are compared to the results which are presented in task 5. The evaluated systems are divided in two categories: installable and online systems.

Online systems are those that are run from a web page. The online systems we use for evaluation are mentioned as follows: Alchemy [16] is a commercial system it belongs to the products of IBM family. It offers the AKE as well as entities, text sentiment, classifies the relevance of results, returns the results in different format and it works with a great range of languages. Skyttle [17] is a commercial system for the AKE and text sentiment, it works only in English. Fivefilters [18] is a terms extractor of open source that returns the most important terms. The parameters are: maximum of results, special formats of results, maximum of words per term. It works only in English. Genia Tagger [19] is a commercial terms extractor designed for texts of biomedical area. We use it in this work for learning the performance in other domain. Tree Tagger [19] is a commercial terms extractor that returns the main terms of an analyzed text. It works only in English. Transalated Labs [20] is a terms extractor for identification of the main terms in one text. It works in French, Italian and English.

Installable systems are those that run locally in computers. The installable systems we use for evaluation as follows: KEA [21] is an open source system of supervised approach that from training dataset can perform automatic keyphrase extraction. The parameters are length phrase, minimum occurrence, vocabulary name. It works in

English, Spanish and French. Extractor [22] is commercial system which extracts keyphrases from a text. The parameters are the number of keyphrases for extraction and list of stopwords. It works in English, Spanish, French, German, Japanese and Korean. Wordstat [23] is a commercial system that counts with tools for the text processing. For the automatic keyphrase extraction, the parameters are length phrase, minimum of occurrence, it works with a great range of languages. TexLexAn [24] is an open source system that incorporates different applications as automatic text summarization, plagiarism detection, keyword extraction. It works in English, Spanish, French, German and Italian.

4 Experimental Results

4.1 Dataset

The dataset used is a collection of scientific articles from task 5 of SemEval-2010. The articles come from the digital library ACM. The distribution of the 4 areas that contains the corpora SemEval-2010 [1]. There are three assignments of golden keyphrases: 1.-Author: keyphrases that have been assigned by the authors of scientific articles by defect. 2.-Reader: keyphrases that were assigned by the readers of scientific articles. 3.-Combined: combination between keyphrases of author and reader.

The systems are presented by the top 5, 10, 15 keyphrases and ranked according to F-score by the top 15 keyphrases as originally in SemEval-2010.

In this paper, the evaluation is implemented using a standard configuration with the objective of measuring the performance of the systems under the following parameters:

Number of keyphrases to extract: a list of 15 keyphrases are extracted for each of 100 scientific articles from SemEval2010. **Minimum length**: keyphrase can be considered of the length of one word. **Maximum length**: based on the system HUMB [7], the maximum number of words that can contain a keyphrase is 5. This is done with the purpose of including the major amount of keyphrases with 4 and 5 words, which mostly occurred in the corpora. There are also longer keyphrases, however contains stopwords. **Frequency**: the systems that have this parameter are left by default for the systems that requires it.

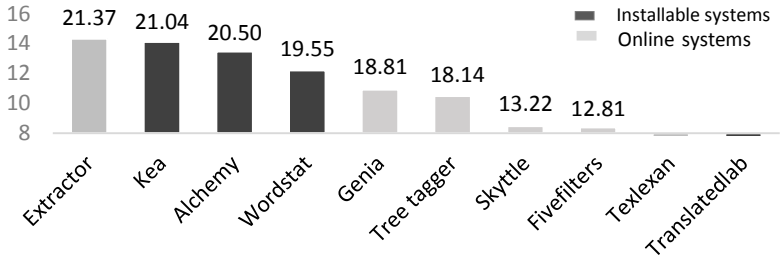

Fig. 1. The performance of author-assigned keyphrases systems (F-score, top 15)

4.2 Results of the Author-assigned Keyphrases

In the author-assigned keyphrases, *KEA* is located in the top 5 as the system with higher result of Precision 15.20%, Recall 19.64% and F-score of 17.14%. For the top 10, KEA again have the highest values of Precision 11.20%, Recall 28.94% and F-score 16.15%. In the top 15, Extractor is positioned in first place for the author-assigned with Precision 9.0%, Recall 34.88% and F-score 14.31%. In table 2, the systems are ranked by the F-score obtained in top 15 keyphrases where the highest values are marked in the top 5, 10 and 15 (see table 1 and figure 1).

Table 1. Results of the systems in the author-assigned (top 15)

		top 5			top10			top15		
System	Rank	P	R	F	P	R	F	P	R	F
Extractor	1	14.80	19.12	16.68	10.40	26.87	15.00	**9.00**	**34.88**	**14.31**
Kea	2	**15.20**	**19.64**	**17.14**	**11.20**	**28.94**	**16.15**	8.87	34.37	14.10
Alchemy	3	14.60	18.86	16.46	10.20	26.36	14.71	8.47	32.82	13.47
Wordstat	4	14.40	18.60	16.23	10.00	25.84	14.42	7.67	29.72	12.19
Genia	5	14.00	18.09	15.78	9.50	24.55	13.70	6.87	26.61	10.92
Tree tagger	6	13.40	17.31	15.11	9.10	23.51	13.12	6.60	25.58	10.49
Skyttle	7	8.20	10.59	9.24	6.40	16.54	9.23	5.33	20.67	8.47
Fivefilters	8	6.00	7.75	6.76	5.50	14.21	7.93	5.27	20.41	8.38
Texlexan	9	5.80	7.49	6.54	4.80	12.40	6.92	4.00	15.50	6.36
Translatedlab	10	5.60	7.24	6.32	4.00	10.34	5.77	3.20	12.40	5.09

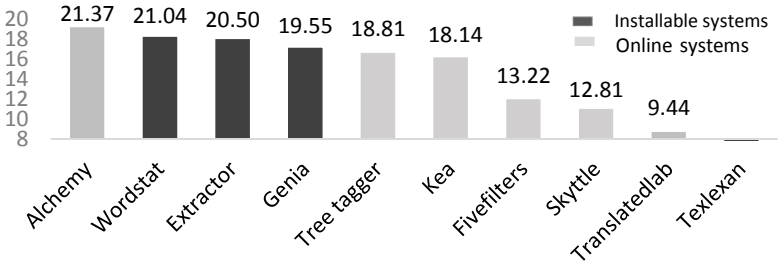

Fig. 2. The performance of reader-assigned keyphrases systems (F-score, top 15)

4.3 Results of Reader-assigned Keyphrases

In the reader-assigned keyphrases, Wordstat is located in the top 5 as system with the highest percentage in Precision 26.20%, Recall 10.88% and F-score 15.38%. For the top 10, Wordstat and Alchemy have the same values in Precision 20.10%, Recall 16.69% and F-score 8.24%. In the top 15, Alchemy with Precision 17.40%, Recall

21.68% and F-score of 19.31%, is positioned in first place of the reader-assigned keyphrases. In table 3, the results are shown where the systems are ranked by the F-score obtained in the top 15 and more higher values are marked in the top 5, 10 and 15 (see table 2 and figure 2).

Table 2. Results of the systems of the reader-assigned keyphrases (top 15)

		Top5			top10			top15		
Systems	Rank	P	R	F	P	R	F	P	R	F
Alchemy	1	25.20	10.47	15.38	**20.10**	**16.69**	**18.24**	**17.40**	**21.68**	**19.31**
Wordstat	2	**26.20**	**10.88**	**15.38**	**20.10**	**16.69**	**18.24**	16.53	20.60	18.34
Extractor	3	19.00	7.89	11.15	17.20	14.29	15.61	16.33	20.35	18.12
Genia	4	24.40	10.13	14.32	18.90	15.70	17.15	15.53	19.35	17.23
Treetrager	5	25.20	10.47	14.79	18.50	15.37	16.79	15.07	18.77	16.72
Kea	6	20.00	8.31	11.74	17.00	14.12	15.43	14.67	18.27	16.27
Fivefilters	7	13.20	5.48	7.74	11.70	9.72	10.62	10.87	13.54	12.06
Skyttle	8	12.40	5.15	7.28	10.80	8.97	9.80	10.00	12.46	11.10
Translatedlab	9	11.20	4.65	6.57	9.10	7.56	8.26	7.93	9.88	8.80
Texlexan	10	10.00	4.15	5.87	7.60	6.31	6.90	6.40	7.97	7.10

4.4 Results of Author- and Reader-, Combined-assigned Keyphrases

In the combined keyphrases, Wordstat is located in the top 5 as the system with the highest result in Precision 32.20%, Recall 10.98% and F-score of 16.38%. The same result, for the top 10, Wordstat is positioned in first place with 24.5% of Precision, Recall 16.71% and F-score of 19.87%. Alchemy in the top 15 is positioned in first place in the combined keyphrases with Precision 21.13%, Recall of 21.62% and F-score of 21.37%. In table 4, the results are shown where the systems are ranked by the F-score obtained in the top 15 keyphrases and more higher values are marked in the top 5, 10 and 15 (see table 3 and figure 3).

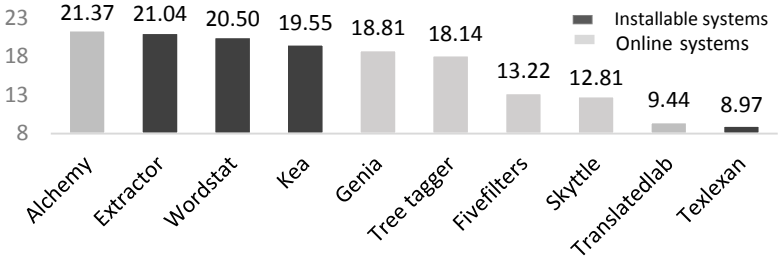

Fig. 3. The performance of combined keyphrases systems (F-score, top 15)

Table 3. Results of the systems in the combined keyphrases (top 15)

		top 5			top10			top15		
Systems	Rank	P	R	F	P	R	F	P	R	F
Alchemy	1	31.20	10.64	15.87	24.40	16.64	19.79	**21.13**	**21.62**	**21.37**
Extractor	2	27.00	9.21	13.73	22.10	15.08	17.93	20.80	21.28	21.04
Wordstat	3	**32.20**	**10.98**	**16.38**	**24.50**	**16.71**	**19.87**	20.27	20.74	20.50
Kea	4	27.80	9.48	14.14	22.80	15.55	18.49	19.33	19.78	19.55
Genia	5	29.60	10.10	15.10	23.00	15.69	18.65	18.60	19.03	18.81
Treetrager	6	30.00	10.20	15.30	22.30	15.21	18.08	17.93	18.35	18.14
Fivefilters	7	16.40	5.59	8.34	14.40	9.82	11.68	13.07	13.37	13.22
Skyttle	8	16.20	5.53	8.25	13.90	9.48	11.27	12.67	12.96	12.81
Translatedlab	9	14.00	4.77	7.12	10.80	7.37	8.76	9.33	9.55	9.44
Texlexan	10	13.40	4.57	6.82	10.70	7.30	8.68	8.87	9.07	8.97

4.5 Comparison of Results with Systems in SemEval-2010

The comparison of the results in this evaluation of the task 5 of SemEval-2010 are presented, with the objective of learning if the systems actually present better performance that already evaluated before.

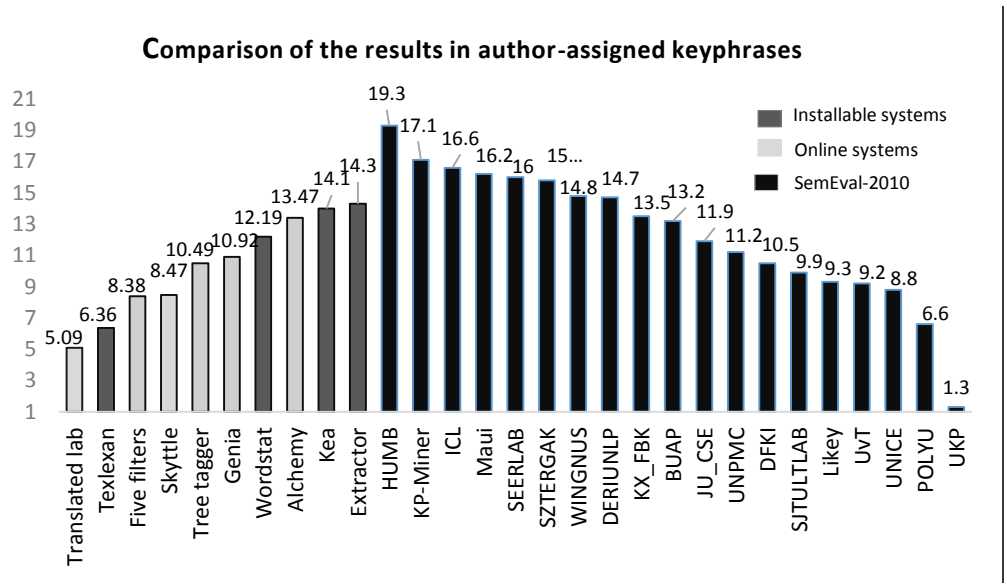

Fig. 4. The performance of the evaluated systems and the systems participated in SemEval-2010 of the author-assigned keyphrases, ranked by F-score of the top 15

Comparison of the results in the author-assigned keyphrases

In this paper, the obtained results of the system *Extractor* of author-assigned keyphrases is 14.31%. DERIUNLP with 14.70% has the similar result in SemEval-2010, while lowest result in this evaluation is *Translatedlab* with 5.09%. In the top 15 in SemEval-2010, UKP obtained 1.3%. In figure 4, the best reached result in this evaluation is *Extractor* 14.31% and in SemEval-2010 is HUMB 19.3%. The bars of color strong gray belong to the installable systems and the bars of color light gray to the online systems that are compared in this paper, while the black bars belong to the systems participated in the task 5 of SemEval-2010 (see figure 4).

Comparison of the results in the reader-assigned keyphrases

In the reader-assigned keyphrases, the system with the highest result is *Alchemy* with 19.31%, its result is similar to DERIUNLP with 19.5% and DFKI with 19.3% in SemEval-2010. The system with the lowest result in this evaluation is *TexLexAn* with 7.1%, while in SemEval-2010 is UKP with 5.2%. In figure 5, the best obtained result in this evaluation is *Alchemy* with 19.31% and in SemEval-2010 is HUMB with 23.5%. The bars of color strong gray belong to the installable systems and the bars of color light gray to the online systems that are compared in this paper, while the black bars belong to the systems participated in the task 5 of SemEval-2010 (see figure 5).

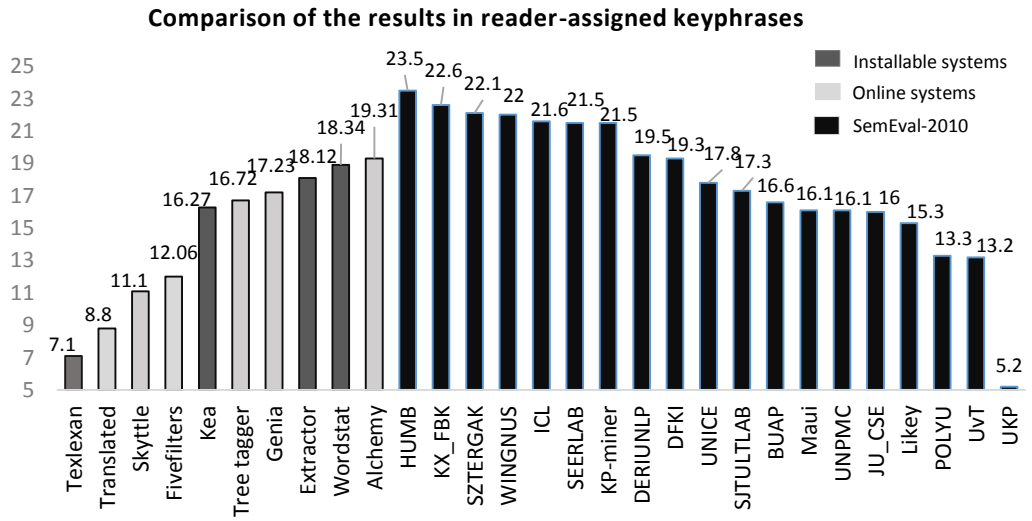

Fig. 5. The performance of the evaluated systems and the systems participated in SemEval-2010 of the reader-assigned keyphrases, ranked by F-score of the top 15

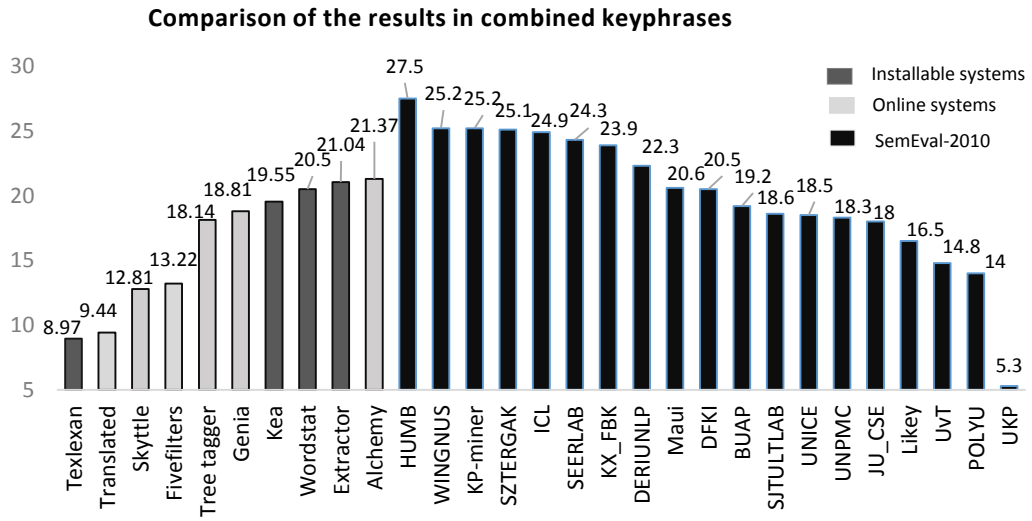

Fig. 6. The performance of the evaluated systems and the systems participated in SemEval-2010 of the combined-assigned keyphrases, ranked by F-score of the top 15

Comparison of the results in the combined keyphrases

In the combined-assigned keyphrases, the system with the highest result in this evaluation is Alchemy with 21.37%. The result can be compared to DERIUNLP with 22.30% and Maui with 20.60%. The system with the lowest result in this evaluation is TexLexAn with 8.97% while in SemEval-2010 is UKP with 5.3%. In figure 6, the best obtained result in this evaluation is Alchemy with 21.37% and in SemEval-2010 is HUMB with 27.5%. The bars of color strong gray belong to the installable systems and the bars of color light gray to the online systems that are compared in this paper, while the black bars belong to the systems participated in the task 5 of SemEval-2010 (see figure 6).

The comparison of the obtained results showed the ranking of the state-of-the-art AKE systems with the already evaluated systems of SemEval-2010.

The results of the systems presented in SemEval-2010 are superior for some systems and are equal for another systems evaluated in this work. We expected that systems evaluated in this paper would present better performance that previously evaluated, considered the time that has passed since 2010.

Also, the presented results show that some terms extraction systems obtain better results than the keyphrase extraction systems.

5 Conclusions

In this paper, the evaluation of systems that automatically extract keyphrases is presented over the commercial free systems that are available in internet for the usage and download. The contribution of the paper is to present the performance of the-state-

of the-art systems and compare the performance with the systems evaluated in SemEval-2010. According to the ranking of systems by the three assignments that contains the gold keyphrases, the system Extractor obtained the first place in the author-assigned while the system Alchemy obtained the first place in the reader- and combined-assigned.

The future work is to test the state-of-the-art systems over other dataset with author- and reader-assigned keyphrases. Other idea is to learn the performance of the systems in different domains. Also, syntactic n-grams [25, 26] and maximal frequent sequences [27-29] will be tested.

Acknowledgment. The work was done with partial support of the Mexico government (CONACyT, SNI, and UAEMex). The authors acknowledge to the Autonomous University of the State of Mexico (UAEMex) for the support.

References

1. Kim, S. N., Medelyan, O., Kan, M. Y., Baldwin, T.: Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics (2010)
2. Kim, S.N., Medelyan, O., Kan, M.Y., Baldwin, T.: SemEval-2010 Task 5: Automatic Keyphrases Extraction from Scientific Articles. Language resources and evaluation, Vol. 47, Issue 3 (2013)
3. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical automatic keyphrase extraction. Proceedings of the fourth ACM conference on Digital libraries, pp. 254–255, ACM (1999)
4. Medelyan, O., Witten, I.H.: Thesaurus based automatic keyphrase indexing. Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, pp. 296–297, ACM (2006)
5. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval, Vol. 2, No. 4, pp. 303–336 (2000)
6. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. Association for Computational Linguistics (2004)
7. Lopez, P., Romary, L.: HUMB: Automatic key term extraction from scientific articles in GROBID. Proceedings of the 5th international workshop on semantic evaluation Association for Computational Linguistics. pp. 248–251 (2010)
8. Lopez, P., Romary, L.: GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains. Seventh international conference on Language Resources and Evaluation (LREC), Valletta, Malta (2010)
9. Wikipedia Database URL: https://en.wikipedia.org/wiki/Wikipedia:Database_download. Consultado 05/3/16.
10. Nguyen, T.D., Luong, M.T.: WINGNUS: Keyphrase extraction utilizing document logical structure. Proceedings of the 5th international workshop on semantic evaluation, pp. 166–169, Association for Computational Linguistics (2010)
11. El-Beltagy, S.R., Rafea, A.: Kp-miner: Participation in Semeval-2. Proceedings of the 5th international workshop on semantic evaluation, pp. 190–193, Association for Computational Linguistics (2010)
12. Berend, G., Farkas, R.: SZTERGAK: Feature engineering for keyphrase extraction. Proceedings of the 5th international workshop on semantic evaluation, pp. 186–189, Association for Computational Linguistics (2010)

13. Pianta, E., Tonelli, S.: KX: A flexible system for keyphrase extraction. Proceedings of the 5th international workshop on semantic evaluation, pp. 170–173, Association for Computational Linguistics (2010)
14. Medelyan, O., Frank, E., Witten, I.H.: Human-competitive tagging using automatic keyphrase extraction. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Vol. 3, pp. 1318–1327, Association for Computational Linguistics (2009)
15. Porter Stemming algorithm, http://tartarus.org/martin/, Consulted 05/3/16
16. AlchemyAPI Keyword Extraction API. Consulted 23/12/15. http://www.alchemyapi.com/products/demo/alchemylanguage.
17. Skyttle. URL: http://www.skyttle.com/demoin. Consulted 23/12/15.
18. Fivefilters Term Extraction. http://fivefilters.org/term-extraction/ Consulted 23/12/15.
19. Termine. Termine web demonstration. http://www.nactem.ac.uk/software/termine, Consultado 23/12/15.
20. Translated Labs. Consulted 23/12/15. http://labs.translated.net/terminology-extraction/.
21. KEA. Keyphrase extraction algorithm. Consulted 23/12/15. http://www.nzdl.org/Kea/.
22. Extractor. Extractor Live Content Demonstration. Consulted 23/12/15. http://www.dbitech.com/trials/dbi_TrialDownloads.aspx.
23. Wordstat 7. Software de análisis de contenido y minería de texto. Consulted 23/12/15. http://provalisresearch.com/es/products/software-de-analisis-de-contenido/.
24. TexLexAn. TexLexAn Analyze, Classify and Summarize any text. Consulted 23/12/15. http://texlexan.sourceforge.net/
25. Sidorov, G.: Syntactic dependency based n-grams in rule based automatic English as second language grammar correction. International Journal of Computational Linguistics and Applications, Vol. 4, No. 2, pp. 169–188 (2013)
26. Sidorov, G.: N-gramas sintácticos no-continuos. Polibits, Vol. 48, pp. 69–78 (2013)
27. García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: A new algorithm for fast discovery of maximal sequential patterns in a document collection. Computational Linguistics and Intelligent Text Processing, pp. 514–523, Springer Berlin Heidelberg (2006)
28. Ledeneva, Y., Gelbukh, A., García-Hernández, R.A.: Terms derived from frequent sequences for extractive text summarization. Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, pp. 593–604 (2008)
29. Ledeneva, Y., García-Hernández, R.A., Gelbukh, A.: Graph ranking on maximal frequent sequences for single extractive text summarization. Computational Linguistics and Intelligent Text Processing, pp. 466–480, Springer Berlin Heidelberg (2014)