Tweet Contextualization:
a Strategy Based on Document Retrieval
Using Query Enrichement and
Automatic Summarization

Jorge Vivaldi and Iria da Cunha

Universitat Pompeu Fabra
Institut Universitari de Lingüística Aplicada
Barcelona
{jorge.vivaldi,iria.dacunha}@upf.edu
http://www.iula.upf.edu

Abstract. The aim of the tweet contextualization INEX (Initiative for the Evaluation of XML retrieval) task at CLEF 2013 (Conference and Labs of the Evaluation Forum) is to build a system that provides automatically information related with different tweets, that is, a summary that explains a specific tweet. In this article, our strategy and results are presented. The methodology for the task in English includes three stages. First, automatic reformulations of the initial queries provided for the task, that is, the tweets, are performed. In this research, we use words sequences that agree with the typical terminological patterns, name entities, hashtags and Twitter users accounts, since we consider that they are representative of tweets' topics. Second, related documents are retrieved from Wikipedia with the search engine Indri, using the reformulated queries. Third, the obtained documents are summarized by using two different automatic summarization systems, in order to provide the final summary associated to each query. Regarding the pilot task for Spanish, our strategy includes a first stage where automatic reformulations of the initial queries provided for the task (similar to English) are carried out. However, it does not include neither the search engine Indri nor the summarization systems REG and Cortex. In this case, we directly extract relevant text passages from Wikipedia pages using the generated queries and we build the summary with the first sentences of these pages.

Key words: Question-Answering, Tweets, Terms, Named Entities, Hashtags, Twitter Users Accounts, Wikipedia, Automatic Summarization, REG, Cortex.

1 Introduction

Twitter is a real-time information network that connects people to the latest stories, ideas, opinions and news about what users find interesting. Nowadays, about 340 millions of tweets are written every day. Each tweet is 140 characters

long. However, 140 characters long messages are rarely self-content. The aim of the tweet contextualization INEX (Initiative for the Evaluation of XML retrieval) task at CLEF 2013 (Conference and Labs of the Evaluation Forum) is to build a system that provides automatically information related with different tweets, that is, a summary that explains a specific tweet. In order to get this aim, it is necessary to combine different types of processing, from information retrieval to multi-document summarization. This INEX task started in 2010, and the results show that best systems combine passage retrieval, sentence segmentation and scoring, named entity recognition and Part of Speech (POS) analysis. The evaluation of the participant systems involves two aspects: informativity and readability. In 2013, the goal of the task and the evaluation metrics has remained unchanged but tweet diversity has been improved. More specially, a significant part of tweets with hashtags have been included in the tweet set. Hashtags are authors' annotation on key terms of their tweets. In the two past years, hashtags have been underused, although they are core components of tweets.

Like in 2012, the use case of this task is the following: given a new tweet, the system must provide some context about the subject of the tweet, in order to help the reader to understand it, i.e. answering questions of the form "what is this tweet about?" using a recent cleaned dump of the Wikipedia. The general process involves: tweet analysis, passage and/or XML elements retrieval and construction of the answer. This context should take the form of a readable summary, not exceeding 500 words, composed of passages from a provided Wikipedia (WP) corpus. The summaries should contain relevant information but also contain as little non-relevant information as possible.

Specifically, this year, 598 tweets in English have been collected by the organizers from Twitter. They were selected among informative accounts (for example, @CNN, @TennisTweets, @PeopleMag, @science...), in order to avoid purely personal tweets that could not be contextualized. Information such as the user name, tags or URLs are provided in JSON format. These tweets are available in a single xml file with three fields: topic, title and txt.

In this edition, an extra set of topics (only tweet texts) has been released in Spanish to try a different language and a slightly different task. Topics in Spanish are opinionated personal tweets about music bands, cars and politics. They were manually selected from CLEF RepLab 2013 test set among those without external url and with at least 15 words. Contextualization should help the reader to also understand the opinion polarity, allusions and humor.

We use two different strategies for the task in English and Spanish. Our strategy for English includes three main stages. First, we perform automatic reformulations of the initial queries provided for the task, that is the title of the tweeets, using terminological patterns, name entities, hashtags and Twitter users accounts. Second, we obtain related documents from WP with the search engine Indri [1], using the reformulated queries. Third, we summarize the obtained documents by using two different automatic summarization systems: REG [2] and Cortex [3], in order to provide the final summary associated to each query. We think that automatic summarization can be useful in this task, taking into

account that a summary can be defined as "a condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source" [4].

This strategy is similar to the one used at QA@INEX track 2010 [5], 2011 [6] and 2012 [7]. In (those three) such INEX editions, we used only the REG summarization system. This year, we use also Cortex, in order to evaluate the impact of the summarization system on the final results. In 2010 and 2011 editions, the system was semi-automatic, while in 2012 and this year the system is totally automatic. The main difference among our participations in all the INEX editions is related with the reformulation and expansion of the queries. This year, the main innovation is the use of the hashtags and the users accounts included in the tweets. As the INEX 2013 organizers suggest, the information given by hashtags can be used to determinate the main topic of the tweet. We combine this information with the information offered by terminological patterns and name entities (that we have previously used in past editions).

Regarding the pilot task for Spanish, our strategy includes a first stage where automatic reformulations of the initial queries provided for the task (similar to English) are carried out. However, it does not include neither the search engine Indri nor the summarization systems REG and CORTEX. In this case, we directly extract relevant text passages from WP pages using the generated queries and we build the summary with the first sentences of these pages.

In Section 2, the state of the art and the resources used are shown. In Section 3, the methodology is explained. In Section 4, the experimental settings and results are presented. Finally, in Section 5, the conclusions and the future work are exposed.

2   State-of-the-art and Resources

2.1   Term and Name Entity Extraction

A term can be defined as a lexical unit that has a specific sense in a thematically restricted domain [8]. Terminology extractors are useful for any NLP task containing a domain specific component such as: ontology and terminological dictionary building, text indexing, automatic translation and summarization systems, among others. There are several methods to obtain the terms from a corpus: methods based on linguistic knowledge or statistical measures, also some authors suggest a combination of both linguistic knowledge and statistically measures. For a review on terminology extraction, see [9]. Most of the tools, in particular those including an important linguistic component, take into consideration the fact that terms usually follow a small number of POS patterns (mostly: noun, noun-adjective and noun-preposition-noun). In this work, we use the terminology extractor YATE [10], to obtain such patterns.

Named Entity extraction may be basically defined as the task to identify names referring to persons, organizations and locations in free text; later this task has been expanded to obtain other entities like dates and numeric expressions. The recent interest in emerging areas like bioinformatics allows to expand

this recognition task to proteins, drugs and chemical names. Name Entity extractors are useful for a number of NLP tasks as question answering, textual entailment and coreference resolution, among others. While early studies were mostly based on handcrafted rules, most recent ones use supervised machine learning as a way to automatically induce rule-based systems or sequence labeling algorithms starting from a collection of training examples. Often, corpus processing tools include some text handling facilities to perform simple Name Entity detection. Some of them are based in language specific peculiarities such as initial upper case letters together with some heuristics for name entities placed at the beginning of the sentence. This is the case of the tool used for this experiment [11].

2.2 Hashtags and Users Accounts

On the one hand, hashtags are sequences of characters and numbers prefixed by a '#'. Social networks largely use this facility to indicate relevant words, groups or topics. From December 2008 to October 2009 hashtags became widely adopted on the site of Twitter. As [12] state:

"A hashtag is the specific name for a tag in Twitter. Hashtags derive their name from the fact that they are preceded by the symbol # also known as a hash mark, e.g., #nowplaying."

Hashtags can be considered as one of the main topics of the tweet. There are several studies on this new type of tagging, which highlight the purpose of these elements. For example, [12] point out that "While tweets without hashtags were also displayed in trending topic lists, the act of tagging a tweet increased the likelihood of a tweet being displayed in a group of tweets on a trending topic."

On the other hand, Twitter is a relevant social network used, as mentioned in the introduction, for INEX competition. In this case, users' accounts are designated by a user name prefixed by '@'. Therefore, this string may be considered as a name entity that could be considered the author of the tweet, or another user related with the tweet.

Both types of information, hashtags and users accounts, will be used in our experiments to improve the queries used to retrieve documents from Wikipedia using Indri and, therefore, to improve the final summary.

2.3 Automatic Summarization

REG [2] is a system for extractive summarization, using a graph approach. The strategy of this system has two main stages: a) to carry out an adequate representation of the document and b) to give a weight to each sentence of the document. In the first stage, the system makes a vectorial representation of the document. In the second stage, the system uses a greedy optimization algorithm. The summary generation is done with the concatenation of the most relevant sentences (previously scored in the optimization stage).

Cortex [3] is a single-document extract summarization system using an optimal decision algorithm that combines several metrics. These metrics result from

processing statistical and informational algorithms on the VSM representation. In order to reduce the complexity, a preprocessing is performed on the topic and the document: words are filtered, lemmatized or/and stemmed. A representation in bag-of-words produces a S[P x N] matrix of frequencies/absences of u = 1, ..., P sentences (rows) and a vocabulary of i = 1, ...,N terms (columns). Cortex system can use several metrics to evaluate the sentences relevance.

3   Methodology

In this section, the two different methodologies used for the task in English and the task in Spanish are explained.

3.1   Methodology for the task in English

The methodology for the task in English includes three stages:

1) Automatic reformulations of the initial queries provided for the task, that is the title of the tweeets, are performed. In this research, we consider that word sequences that agree with the typical terminological patterns, name entities, hashtags and Twitter users accounts are representative elements of tweet topics. Therefore, in order to reformulate and build the query we extract automatically two types of information from the tweets:

a) Terms and name entities. First, we POS tag the tweets file, obtaining proper nouns info, which we consider as name entities. Second, we carry out terminological pattern extraction, by using an already existent module of the YATE term extraction tool [10]. Third, we check if these elements are included in the Wikipedia. If they are included, we maintain them for the query string for Indri. Otherwise, they are erased.

b) Hashtags and users accounts. Usually, hashtags and Twitter accounts are not fully arbitrary strings but very often the actual entity may be discovered just analyzing the string.

First, we extract elements tagged with hashtags (#) and users accounts (@) in each tweet. Second, we check if these elements are included in WP. If they are included, we maintain them for the query string for Indri. If they are not included, the system segments the letters of hashtags and the user accounts, in order to find all the possible combinations of real words (with a maximum of two words). We apply this strategy taking into account that many times hashtags and user accounts include no spaces, so two or more different words can be joined. For our purpose, we need to have the words separated, in order to check if they are included in Wikipedia. For example:

- The user account @BarbraStreisand is not present as such in Wikipedia. Nevertheless, Wikipedia includes "Barbra Streisand", that will be included in our query.
- The hashtag #EUbudget is not present as a single unit in Wikipedia. However, Wikipedia includes both "EU" and "budget", therefore these elements separately will be included in our query.

– The hashtag #NASA_Astronauts is not included as a single unit in Wikipedia. Although, Wikipedia contains "NASA" and "Astronauts" separately. We include both elements in our query.

If the system does not find any real word after applying all possible segmentations in hashtags and users accounts, we do not include these elements in the query string for Indri.

In order to enrich the queries, we use a local installation of a Wikipedia dump[1] to expand the terms with redirection information from such Wikipedia info. Some care has been taken to keep track of multiword sequences as indicated by the Indri query language specification (see examples below). For example, for the tweet:

Behind the scenes on Charlie Brooker's Black Mirror finale

we obtain the following query:

```
scenes,finale,#1(Black Mirror),#1(Charlie Brooker)
```

2) Related documents (specifically 50 texts) are retrieved from WP with the search engine Indri, using the reformulated queries. The resulting set of WP pages has been splitted in several documents. Each document contains the pages relevant to the query.

3) The obtained documents are summarized by using REG and CORTEX summarization systems, in order to provide the final summary associated to each query.

Figure 1 illustrates the methodology for the task in English.

3.2 Methodology for the task in Spanish

Our strategy for Spanish includes neither the search engine Indri nor the summarization systems REG and Cortex. This strategy includes two main stages:

1) To carry out automatic reformulations of the original queries provided for the task. First, name entities, hashtags and Twitter users accounts are extracted from the tweet. Second, the system checks if they have a main page in WP: if they have a page in WP, they are included in the final query; if they do not have a page in WP, they are eliminated. Third, if after this process, the query includes 1 or 0 elements, terminological patterns are used to extract more units from the tweet, in order to include in the query those which have a main entry in WP. For example, for the tweet:

Vaya asco de árbitros hay en la liga BBVA, ya solo por los árbitros que hay, nadie tendría que decir que es la mejor liga del mundo...

we obtain the following query:

---

[1] This resource has been otained using [13].
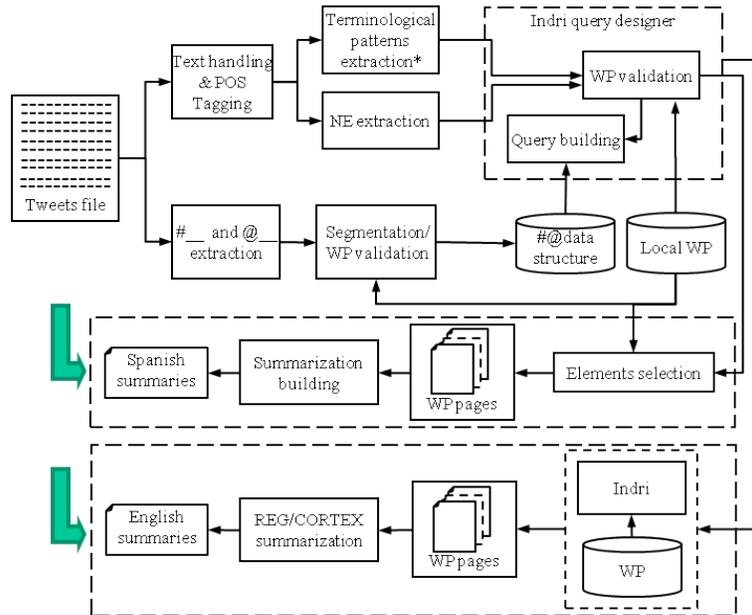
Fig. 1. Full methodology (English/Spanish).

```
mundo,liga,#syn("BBVA" #1(Banco Bilbao Vizcaya Argentaria))
```

2) To extract relevant text passages from WP pages using the generated queries, which are considered as a summary. Our system carries out a summary including about 500 words, giving the same importance to the name entities, the terminological units, the hashtags and the Twitter users accounts included in the query. For example, if a query contains three of these elements, the system extracts the first sentences of the WP page of each of the three, to obtain a summary including a third of sentences for each element.

Some filters are applied before obtaining the final version of the summary: a) ambiguous WP pages are eliminated, b) passages in brackets are erased, c) only passages included at the beginning of the WP page are selected (text appearing after the first subtitle of the WP page is not considered), and d) lists of elements are eliminated.

Figure 1 illustrates the methodology for the task in Spanish.

4   Experiments Settings and Results

In this section, the experiments and results for the task in English and the task in Spanish are explained.

4.1 Experiments for the task in English

For the task in English, we have applied our strategy using the 598 tweets that have been collected by the organizers from Twitter. From these tweets, 529 included different hashtags and/or users accounts. From this set of elements, 391 were found in Wikipedia.

The evaluation of all the participant systems in the INEX 2011 QA Track involves two aspects: informativity and readability. On the one hand, to evaluate informativity, the automatic FRESA [FRamework for Evaluating Summaries Automatically] [14] package is used. This evaluation framework includes document-based summary evaluation measures based on probabilities distribution, specifically, the Kullback-Leibler (KL) divergence and the Jensen-Shannon (JS) divergence. As in the ROUGE package [15], FRESA supports different n-grams and skip n-grams probability distributions. FRESA environment has been used in the evaluation of summaries produced in several European languages (English, French, Spanish and Catalan), and it integrates filtering and lemmatization in the treatment of summaries and documents. FRESA is available in the following link: `http://lia.univ-avignon.fr/fileadmin/axes/TALNE/Ressources.html`.

Informativity has been evaluated based on three overlapping references:

- prior set of relevant pages selected by organizers while building the 2013 topics (40 tweets, 380 passages, 11 523 tokens),
- pool selection of most relevant passages from participant submissions for tweets selected by organizers (45 tweets, 1 760 passages, 58 035 tokens),
- all relevant text merged together with an extra selection of relevant passages from a random pool of ten tweets (70 tweets, 2 378 passages, 77 043 tokens)

Table 1 include the results about informativity. In this table, we have only included the best run of each different team. The complete table can be retrieved in the official website of INEX 2013 (`https://inex.mmci.uni-saarland.de/`).

Informativity results show that the impact of the summarization system in our strategy is not high. The results using Cortex (run 262) are slightly better than the results using REG (run 255).

On the other hand, readability has been evaluated by organizers over the ten tweets having the largest text references (t-rels). For these tweets, summaries are expected to have almost 500 words since the reference is much larger. For each participant summary, organizers have checked the number of words over 500 in passages that are:

- Relevant (T) i.e. clearly related to the tweet.
- Sound (A) i.e. no issues about resolving references to earlier or later items in the discourse.
- Non redundant (R) with previous passages.
- Syntactically (S) correct.

Table 1. Results of informativity in the INEX 2013 Tweet Contextualization Track.

Run	All.skip	All.bi	All.uni	Pool.skip	Pool.bi	Pool.uni	Prior.skip	Prior.bi	Prior.uni
256	0.8861	0.881	0.782	0.8752	0.87	0.7813	0.921	0.9134	0.7814
275	0.8969	0.8924	0.8061	0.8789	0.8745	0.7941	0.9172	0.9106	0.7899
254	0.9242	0.9229	0.8331	0.9162	0.9159	0.8363	0.9473	0.943	0.8223
276	0.9301	0.927	0.8169	0.9333	0.9302	0.8285	0.9718	0.9678	0.8286
270	0.9397	0.9365	0.8481	0.9274	0.9246	0.8418	0.9686	0.9642	0.8529
277	0.9662	0.9649	0.8995	0.9642	0.9626	0.9005	0.9792	0.9773	0.9102
261	0.967	0.9668	0.8639	0.9656	0.9659	0.8666	0.9888	0.9862	0.8687
262	0.9747	0.9734	0.8738	0.9736	0.9727	0.8775	0.9821	0.9788	0.8635
255	0.9783	0.9771	0.8817	0.9759	0.9748	0.8801	0.9938	0.9914	0.8941
265	0.9789	0.9781	0.8793	0.9751	0.9749	0.8821	0.9927	0.9904	0.8845
266	0.9835	0.9824	0.9059	0.9865	0.9859	0.9132	0.9903	0.9877	0.8952
269	0.9999	0.9999	0.9965	0.9999	0.9999	0.9972	1	0.9999	0.9962

Non relevant passages have also been considered non sound, redundant and syntactically incorrect. Runs are ranked according to mean average scores per summary over Soundness, Non redundancy and Syntactically correctness among Relevant passages. Tabla 2 shows the results about readability.

Readability results show that the impact of the summarization system in our strategy is higher than in the case of informativity results: the results using REG (run 255) obtain a higher score than the results using Cortex (run 262). Anyway, this difference is not significant, since none of the systems appears in the best positions.

Since the process is fully automatic, we find some limitations. The main limitation that we have found is that Indri does not retrieve relevant documents. For example, one of the queries includes the name entity "Jennifer Hudson" and only one of the 50 documents retrieved by Indri includes this name entity. This fact limits our results: if the input given to the summarization systems is not relevant, the summary will be not relevant either.

Another limitation of our strategy is the kind of summarization systems that we have used. Both REG and Cortex are mono-document summarization systems. We offer to these systems a single text including the 50 documents retrieved by Indri, and they produce an extractive summary. This fact could produce a lack of correference and coherence, which can affect the final results.

4.2 Experiments for the task in Spanish

For the task in Spanish, we have applied our strategy using the 354 tweets that have been collected by the organizers from Twitter. Nowadays, we do not have the final official results evaluation in the INEX 2013 Tweet Contextualization Track. We will include the results in a more complete version of this paper.

Table 2. Results of readability in the INEX 2013 Tweet Contextualization Track.

Run	Mean Average	Relevancy (T)	Non redundancy (R)	Soundness (A)	Syntax (S)
275	72.44%	76.64%	67.30%	74.52%	75.50%
256	72.13%	74.24%	71.98%	70.78%	73.62%
274	71.71%	74.66%	68.84%	71.78%	74.50%
273	71.35%	75.52%	67.88%	71.20%	74.96%
257	69.54%	72.18%	65.48%	70.96%	72.18%
254	67.46%	73.30%	61.52%	68.94%	71.92%
258	65.97%	68.36%	64.52%	66.04%	67.34%
276	49.72%	52.08%	45.84%	51.24%	52.08%
267	46.72%	50.54%	40.90%	49.56%	49.70%
270	44.17%	46.84%	41.20%	45.30%	46.00%
271	38.76%	41.16%	35.38%	39.74%	41.16%
264	38.56%	41.26%	33.16%	41.26%	41.26%
260	38.21%	38.64%	37.36%	38.64%	38.64%
265	37.92%	39.46%	36.46%	37.84%	39.46%
259	37.70%	38.78%	35.54%	38.78%	38.78%
255	36.59%	38.98%	31.82%	38.98%	38.98%
261	35.99%	36.42%	35.14%	36.42%	36.42%
263	32.75%	34.48%	31.86%	31.92%	34.48%
262	32.35%	33.34%	30.38%	33.34%	33.34%
266	25.64%	25.92%	25.08%	25.92%	25.92%
277	20.00%	20.00%	20.00%	20.00%	20.00%
269	00.04%	00.04%	00.04%	00.04%	00.04%

5 Conclusions and Future Work

In this paper, our strategy for the INEX 2013 Tweet Contextualization Track are presented, both for the task in English and in Spanish. In our work, as explained, a reformulation of the initial queries (tweets) is carried out by using terminological patterns, named entities, hashtags and Twitter users accounts. For the task in English, the automatic summarization systems REG and Cortex are used to summarize the 50 documents obtained by Indri using these queries. We have proved that the use of different summarizers has not a significant impact over the results. Also, we have detected that the main problem is related with the documents retrieved by Indri. Therefore, for the pilot task for Spanish, we have not used Indri. We hope to improve the results in this way.

As usual, ambiguous terms represent a serious problem in a procedure like the proposed in this experiment. In this work, ambiguous terms (that is, terms linked to a WP disambiguation page) have not been considered, in order to prioritize precision. In the future, we would like to experiment with systems such as those proposed by [16] or [17] in order to perform some disambiguation and therefore improve the selection of the relevant elements for the queries. Other improvement could be achieved by clustering the WP pages provided by Indri with the aim to obtain a better WP pages selection and then summarize them.

Acknowledgements

References

1. Strohman, T.; Metzler, D.; Turtle, H.; Croft, B. (2005). *Indri: A language-model based search engine for complex queries*. University of Massachusetts Amherst. CIIR Technical Report IR-407.
2. Torres-Moreno, J-M.; Ramírez, J. (2010). *REG : un algorithme glouton appliqué au résumé automatique de texte*. Proceedings of the 10th Int. Conference on the Statistical Analysis of Textual. Roma, Italia.
3. Torres-Moreno, J. M.; Velázquez-Morales, P.; Meunier, J. G. (2002). *Condensés de textes par des méthodes numériques*. Proceedings of the 6th Int. Conference on the Statistical Analysis of Textual Data (JADT). St. Malo. 723-734.
4. Saggion, H.; Lapalme, G. (2002). *Generating Indicative-Informative Summaries with SumUM*. Computational Linguistics 28(4). 497-526.
5. Vivaldi, J.; da Cunha, I.; Ramírez, J. (2011). *The REG summarization system with question reformulation at QA@INEX track 2010*. Lecture Notes in Computer Science (LNCS) 6932. 295-302. Berlin: Springer.
6. Vivaldi, J.; da Cunha, I. (2012). *QA@INEX Track 2011: Question Expansion and Reformulation Using the REG Summarization System*. Lecture Notes in Computer Science (LNCS) 7424. 257-268. Berlin: Springer.
7. Vivaldi, J.; da Cunha, I. (2012). *INEX Tweet Contextualization Track at CLEF 2012: Query Reformulation using Terminological Patterns and Automatic Summarization*. In Forner, P.; Karlgren, J.; Womser-Hacker, C. (Eds.). CLEF 2012 Evaluation Labs and Workshop, Online Working Notes.
8. Cabré, M. T. (1999). *La terminología: Representación y comunicación*. Barcelona: Institut Universitari de Lingüistica Aplicada, Universitat Pompeu Fabra.
9. Cabré, M. T.; Estopà, R.; Vivaldi, J. (2001). *Automatic term detection. A review of current systems*. Recent Advances in Computational Terminology 2. 53-87.
10. Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. Ph.D. thesis. Universitat Politècnica de Catalunya. Barcelona (Spain).
11. Martínez, H.; Vivaldi, J.; Villegas, M. (2010). *Text handling as a Web Service for the IULA processing pipeline*. Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10). 22-29.
12. Huang, J.; Thornton, K. M.; Efthimiadis, E. N. (2010). *Conversational Tagging in Twitter*. Proceedings of the 21st ACM Conference on Hypertext and Hypermedia.
13. Zesch, T.; Muller, C., Gurevych, I. (2008). *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary*. Proceedings of the 6th Conference on International Language Resources and Evaluation (LREC'08). 1646-1652.
14. Saggion, H.; Torres-Moreno, J-M.; da Cunha, I.; SanJuan, E.; Velázquez-Morales, P.; SanJuan, E. (2010). *Multilingual Summarization Evaluation without Human Models*. Proceedings of the 23rd Int. Conference on Computational Linguistics (COLING 2010). Pekin.

15. Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries.* Proceedings of Text Summarization Branches Out: ACL-04 Workshop. 74-81.

16. Mihalcea, R. (2007). *Wikify! Linking Documents to Encyclopedic Knowledge.* Proceedings of the sixteenth ACM conference on Conference on information and knowledge management: CIKM '07 Pages 233-242. New York.

17. Ferragina, P.; Scaiella U. (2010). *TAGME: On-the-fly Annotation of Short Text Fragments (byWikipedia Entities).* Proceedings of the 19th ACM international conference on Information and knowledge management: CIKM '10 Pages 1625-1628. New York.