Expert Systems with Applications 40 (2013) 6624–6630

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre

Elena Lloret *, Manuel Palomar

NLP research group, Department of Software and Computing Systems, University of Alicante, E-03080 Alicante, Spain

ARTICLE INFO

Keywords:
Generation
Twitter
Automatic tweet generation
Text summarization
Informativeness
Indicativeness
Interest

ABSTRACT

In recent years, Twitter has become one of the most important microblogging services of the Web 2.0. Among the possible uses it allows, it can be employed for communicating and broadcasting information in real time. The goal of this research is to analyze the task of automatic tweet generation from a text summarization perspective in the context of the journalism genre. To achieve this, different state-of-the-art summarizers are selected and employed for producing multi-lingual tweets in two languages (English and Spanish). A wide experimental framework is proposed, comprising the creation of a new corpus, the generation of the automatic tweets, and their assessment through a quantitative and a qualitative evaluation, where informativeness, indicativeness and interest are key criteria that should be ensured in the proposed context.

From the results obtained, it was observed that although the original tweets were considered as model tweets with respect to their informativeness, they were not among the most interesting ones from a human viewpoint. Therefore, relying only on these tweets may not be the ideal way to communicate news through Twitter, especially if a more personalized and catchy way of reporting news wants to be performed. In contrast, we showed that recent text summarization techniques may be more appropriate, reflecting a balance between indicativeness and interest, even if their content was different from the tweets delivered by the news providers.

1. Introduction

The Web 2.0, and in particular, social media have revolutionized the way we communicate and interact with people, businesses and organizations (Hanna, Rohm, & Crittenden, 2011). This has also affected how information is consumed on the Internet, leading to new reading, sharing, or writting habits when it concerns online texts. Focusing on the journalism genre, the study conducted in Hahn (2013) showed that people do not actively look for news, but find them on social media instead. In particular, it was reported that 43% of young people, aged 16–24, found their news on social media rather than through search engines. Moreover, online news readers tend to scan rather than read them in depth (Holmqvist, Holsanova, Barthelson, & Lundqvist, 2003), as well as they gather news from several sources rather than a single destination site (Aitamurto & Lewis, 2013). Statistics show that, although 8 out of 10 people read headlines, only 2 go onto read its content.[1]

Twitter[2] is a social media service that is increasing in size at a significant rate. It has more than 200 million active users and 400 million tweets each day (March 2013[3]). In the context of newswire, Twitter is by far more widespread for gathering, broadcasting and reporting news than other social networks, such as Facebook (Hahn, 2013). This may be due because it has a huge potential for disseminating information in real-time, which can range from broadcasting breaking news to reporting a piece of news to generate debate and opinion. In Kwak, Lee, Park, and Moon (2010), it was shown that over 85% of topics addressed in Twitter were headlines or persistent news. Furthermore, it is common for almost all major news sources to have active accounts in social media services like Twitter to take advantage of the enormous reach these services provide, existing at the same time, an intense and fast paced competition for attention among news items published online (Bandari, Asur, & Huberman (2012)).

On the other hand, headlines should satisfy two relevant principles (Ifantidou (2009)): (i) summarize the content of a piece of news, and (ii) attract the attention to the full-text newspaper article. Additionally, if the headline has to be broadcasted through Twitter, it needs to be very concise, having at maximum 140 characters. However, the automatic tweets derived from press agencies consist of the headline of the news, and in some of the cases, they also contain a link to the whole article. Although headlines are

* Corresponding author. Tel.: +34 965902961.
E-mail addresses: elloret@dlsi.ua.es (E. Lloret), mpalomar@dlsi.ua.es (M. Palomar).

[1] http://www.copyblogger.com/writing-headlines-that-get-results/.

[2] https://twitter.com/.

[3] http://blog.twitter.com/2013/03/celebrating-twitter7.html.

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.021

normally short, the limited number of characters allowed in a tweet (only 140 characters) make that headlines with more than 140 characters have to be shortened. This may lead to incomplete tweets, where the user needs to access the Website, if he or she wants to read the complete headline or news. For instance, the headline:

> *"Age of consent should be lowered to 13 to stop persecution of old men and sex assault victims should not get anonymity, says leading barrister"*

has to be shortened to fit in a tweet:

> *"Age of consent should be lowered to 13 to stop persecution of old men and sex assault victims...* http://bit.ly/10qjMT9via @MailOnline"

However, as it can be seen, the shortened version does not summarize the information, but also just cuts the sentence, providing the link to the full-text news, that would be necessary to access if a user wants to read at least the full headline.

This issue together with the fact that the research conducted in Petrovic et al. (2013) showed that Twitter could report the same events as newswire providers, may be detrimental, decreasing the role of news providers in social media.

One way to add value to such tweets is by generating them differently. Instead of reporting the headline associated to the news, we could take advantage of Natural Language Processing (NLP) tools, and specifically Text Summarization (TS), for producing an *ultra-concised* summary containing the *essential* information of the news.

Therefore, the goal of this research is to analyze the task of automatic tweet generation from a TS point of view in the context of the journalism genre. This means that our study focuses on the generation of tweets for reporting and broadcasting information. This research work has three main contributions: (i) we investigate to what extent state-of-the-art TS techniques are appropriate for generating a summary from a large piece of news in the form of a tweet; (ii) we propose a novel experimental and evaluation framework, comprising the creation of a new corpus, the testing of TS approaches, and the quantitative and qualitative assessment of the generated tweets; and (iii) we analyze the task in two languages (English and Spanish), comparing the potentials and limitations of current multi-lingual summarization systems when applied to this task.

The results show that even though current tweets delivered by news providers contain relevant information, they may be not considered interesting from a human point of view. However, specific TS techniques were shown appropriate to be used to generate automatic tweets, keeping a balance between information indicativeness and interest.

The remainder of this article is organized as follows. Section 2 covers the related work and puts our work into perspective. Then, Section 3 describes our data, and Section 4 explains the TS systems analysed within the scope of this research. Section 5 reports the experiments and evaluation conducted, and finally, Section 6 concludes the article and outlines future work.

2. Related work

The task of automatic tweet generation could be related at least with two NLP tasks: (i) headline generation, and (ii) keyword extraction. On the one hand, it is similar to automatic headline generation in the sense that the aim is to produce a very brief summary containing the most relevant information from a piece of news. Headline generation is a well-studied task within single-document summarization. During different editions of DUC competitions[4] there was a specific task aiming at producing headlines no longer than 50 words. The techniques employed for producing these very short summaries included the use of lexical and named entities chains (Fuentes, Massot, Rodríguez, & Alonso, 2003), information about the topic (Wan, Dras, Paris, & Dale, 2003), parsing and trimming (Dorr, Zajic, & Schwartz, 2003) or language models (Soricut & Marcu, 2007). Recent research has focused on applying the headline generation task to produce titles (Lopez, Prince, & Roche, 2012; Tseng, 2010), image captions (Woodsend & Lapata, 2010), or even story highlights (Woodsend, Feng, & Lapata, 2010). On the other hand, the techniques employed in automatic keyword extraction (Romero, Moreo, Castro, & Zurita, 2012, 2013) could be useful for tweet generation, specifically for identifying the set of relevant keywords that could be transformed into hashtags for producing the tweet, or could be combined for generating a new sentence. This would be another manner of presenting a tweet, which is out of the scope of this research.

Moreover, TS has been applied to Twitter from different perspectives, but not with the aim of generating automatic tweets. For instance, TS has been used to: (a) summarize a set of related tweets about the same topic (Connor, Krieger, & Ahn, 2010; Sharifi, Hutton, & Kalita, 2010; Weng, Yang, Chen, Wang, & Lin, 2011; Liu, Liu, & Weng, 2011); (b) generate event summaries (Chakrabarti & Punera, 2011), or (c) produce opinion summaries from tweets (Ganesan, Zhai, & Viegas, 2012). An interesting novel task of tweet contextualization was proposed within CLEF 2012 forum.[5] This task consisted of generating a comprehensive summary of 500 words that provided additional information and could be used for better understanding a tweet.

In this respect, TS has a great potential for Twitter. It could be useful for improving the broadcasting of relevant information, as well as to help better comprehend the information expressed in one or several tweets.

Concerning the automatic tweet generation task on its own, we are only aware of a previous research work. In Lofi and Krestel (2012), a potential application of automatic tweet generation is described, where the tweets are used to increase the transparency of government actions. The proposed approach combines several NLP techniques, such as topic classification, or TS; however, it is just a task proposal, and none of the techniques and the system as a hole are analyzed or evaluated.

To the best of our knowledge, there is no previous research aiming to analyze the effectiveness of state-of-the-art TS systems for generating a tweet automatically, and although it could be a similar task as the one of headline extraction, tweet generation is even more challenging, due to the 140 character size limit, thus having to select the appropriate information. Otherwise, the tweet may be useless and could be considered as spam.

3. Corpus development

Since there was not any previous corpus for addressing the task of automatic tweet generation, we created our own corpus. Such corpus consisted of a collection of newswire documents from online newspaper sites, but taking into account that they should have a Twitter account, as well as they should allow readers to directly share the news articles through this social media channel.

For collecting the documents in our corpus, we specifically selected 10 different English and Spanish newspaper online Websites (5 from the UK, and 5 from Spain). Table 1 shows the information sources that have been employed for the corpus collection.

[4] http://www-nlpir.nist.gov/projects/duc/.

[5] https://inex.mmci.uni-saarland.de/tracks/qa/.

Table 1
Online newspapers for corpus collection.

Language	Media	URL
English	BBC	www.bbc.co.uk
English	The Guardian	www.guardian.co.uk
English	The Independent	www.independent.co.uk
English	The Scotsman	www.scotsman.com
English	The Telegraph	www.telegraph.co.uk
Spanish	El País	www.elpais.com
Spanish	El Mundo	www.elmundo.es
Spanish	Público	www.publico.es
Spanish	ABC	www.abc.es
Spanish	La Razón	www.larazon.es

Instead of focusing on a particular domain or type of news, we gathered the news which were among the top most read, commented, or shared. This resulted in a corpus that not only news from different domains were contained, such as sports, science, technology, or culture, but also the corpus would be useful for carrying out other type of user studies in the future (e.g., comparison across languages, interest in the types of news, evolution of events, etc.).

In total, we collected 200 news, i.e., 100 news for each language within a 10-day period. Such news were pre-processed, removing all unnecessary extra information, such as images, external links, advertisements, and even headlines. Only the main body of the news was kept. In addition, we also recorded the tweet associated to each of the news, which normally coincided with its headline (further details concerning the analysis of the tweets derived from online news is provided in Section 3.1).

Table 2 shows some statistics the corpus. As it can be seen, the news articles are quite long in both languages. Taking into account that a tweet must not exceed 140 characters, the required compression rate for the summaries (i.e., the length of the summary with respect to the length of the full-text news, with respect to the number of characters) should be around 3% for both English and Spanish. This is an extremely small compression rate that have not been previously tested by any system, since it has been shown that summaries with a compression rate of 20% and 30% of the source document are effective surrogates (Morris, Kasper, & Adams, 1992), so these are the most common compression rates analyzed in the literature.

Although the corpus is not very big (200 documents in total), we believe that it is big enough to conduct this preliminary study for analyzing the capabilities of current TS systems for generating automatic tweets, and determine whether they are good synthesizers or not. Nevertheless, it would be possible to easily extent the corpus with a higher number of news, as well as to increase the number of sources.

3.1. Tweet analysis

As it was aforementioned, apart from the full-text news, the corpus also contains the corresponding tweet associated to each of them. Previous to the generation of our own tweets in an automatic way, we wanted to understand how they were delivered by online newspaper Twitter accounts. Therefore, we examined in detail the tweets collected within our corpus.

Table 2
Corpus statistics for news articles.

Number of documents		English	Spanish
		200	200
Number of words	Min.	121	125
	Max.	2869	4007
	Avg.	853.32	741.11
Number of characters	Min.	1039	687
	Max.	16451	23084
	Avg.	4915.19	4363.85

Regardless the language (English or Spanish) and the information source, the tweets consist of the headline of the news, followed by a link to the full text. This happens for all the tweets.

However, depending on the newspaper, the manner in which the link is provided varies in length, and normally at the time of posting it is shortened. Moreover, a common characteristic of the tweets concerns with specifying where the information comes from. The media analyzed have their own accounts in Twitter, and therefore, most of their tweets (although not all of them) include the "@" symbol, in order to mention the source of information (e.g.,*"via @Telegraph"*). Specifically, 55% of the English tweets in our corpus contain this information, whereas this percentage raises to 79%, for the Spanish ones. We observed that in the cases this information is not included, either the whole URL where the name of the newspaper is mentioned (e.g., *http://www.publico.es/deportes/425370/los-leones-asaltan-el-teatro-de-los-suenos*), or the source information itself (e.g., *BBC News*) is provided. With respect to the links, since Twitter only allows entries of 140 characters (including white spaces), a common phenomenon for posting a tweet is to shorten the original URL, so more space is devoted for the information itself.

Table 3 provides some properties about the tweets generated from news items. We computed the number of words and characters of the whole tweet, as well as the same information for its content (i.e., without considering URL links or information sources). We refer to the latter as "info tweet". For our experiments and evaluation, only the "info tweet" will be considered.

As it can be seen from the figures reported in Table 3, the number of words is very low. This is explained by the fact that, for the full tweet the URLs have been considered as a unique word. This is the reason why the figures showin for the full and info tweet are quite similar.

Furthermore, it is worth mentioning the fact that whereas for English, the average tweet length is 164.02 characters, which surpasses the Twitter 140 character-limit restriction. In light of this fact, we wanted to analyze whether this was a common tendency. After checking the length of the individual tweets for both languages, we found out that this occured for some Spanish tweets as well. 66% of English tweets, and 44% of Spanish tweets had length problems. This meant that the deadline of the news was longer than 140 characters, and could affect the readability of the tweet when published, since at the time of publishing the tweet, this was incomplete. As we previously mentioned in Section 1, the strategy adopted by the Twitter sharing service is to truncated the headline until it fitted within 140 characters. For achieving this, the exceeding characters or words are replaced by "...".

An example of two tweets (one for English -EN- and one for Spanish -ES) are shown in Table 4. The one at the top is taken from a news appearing in *"The Guardian"*, whilst the latter belongs to a Spanish newspaper *"El Mundo"*.[6]

4. Text summarization approaches

The aim of this section is to explain the TS systems employed as well as justifying this selection. For this initial study, six state-of-the-art summarizers were selected for extracting relevant content from documents (i.e., extractive summarizers). Our goal here is to assess whether the techniques employed are appropriate for extracting relevant content from documents in the form of a tweet.

[6] English translation for the Spanish tweet: Shakira stops the traffic in Barcelona.

Table 3
Tweet properties for each language.

		English	Spanish
Number of words (full tweet)	Min.	1	4
	Max.	27	19
	Avg.	14.05	11.58
Number of words (info tweet)	Min.	1	16
	Max.	26	16
	Avg.	12.01	9
Number of characters (full tweet)	Min.	52	55
	Max.	103	268
	Avg.	164.02	135.17
Number of characters (full tweet)	Min.	20	14
	Max.	129	92
	Avg.	73.90	61.14

Table 4
Examples of tweets derived from a news.

EN	Gorilla genome analysis reveals new human links http://gu.com/p/362kt/twvia@guardian
ES	Shakira para el tráfico en Barcelona http://mun.do/wHx5uzvía@elmundoes

Although there is a wide number of summarizers developed by the research community, not all of them were suitable for our purposes. In the selection of the summarizers, we took into consideration that the TS system or approach: (i) should be multi-lingual, and it should work, at least, for English and Spanish, the languages dealt with in our study; (ii) should have been previously tested on newspaper articles; and (iii) should be currently available or easily reproducible. Next, each of the selected TS systems is described:

- *SweSum*[7]: This TS system (Hassel, 2007) employs different features for determining the importance of a sentence, mainly: (i) sentences in the beginning of the text are given higher scores than the ones at the end; (ii) sentences containing numerical data are given a higher score than the ones without numerical values; and (iii) sentences which contain keywords (frequent terms) are scored higher. All these features are normalized and combined to obtain the total score of each sentence. Then, the highest scored sentences are extracted until the desired summary length, which can be configured by the user.
- *Open text summarizer*[8]: This is a multi-lingual summarizer able to generate summaries in more than 25 languages, including English and Spanish. The idea behind this system is that the important facts in an article are described with many of the same words while redundant information uses less technical terms and is not related to the main subject of the article. In this approach, keywords are identified by means of word occurrence, and sentences are given a score based on the keywords they contain. Some language-specific resources, such as stemmers and stop word lists are employed. It has been shown that this system is very competitive, since it obtains better performance than other multi-lingual TS systems (Yatsko & Vishnyakov, 2007).
- *AutoSummarize*[9]: This summarizer is integrated into Microsoft Word and it also generates summaries in several languages. Since it is a commercial system, the implementation details are not revealed. However, from our experience in using it, and the experiments carried out, we deduce that it mainly uses positional and statistical features.
- *Extractor*[10]: For summarizing documents, this system uses a genetic programming approach which itself provides an automatic learning process, which allows the summarizer to work on different domains without re-training it. This approach is the result of the research carried out in Turney (2000), where several learning algorithms were analyzed and evaluated for determining the best for the keyphrase extraction task. Currently, it is also a commercial system which has an online demo for testing it.
- *COMPENDIUM*[11]: This summarizer (Lloret, 2011) relies on two main issues for generating summaries: (i) redundant information identification and (ii) relevance detection. The former uses textual entailment as a method for removing repeated information, whereas the later takes into consideration statistical and cognitive-based features for selecting and extracting relevance in documents. Originally, COMPENDIUM was a mono-lingual summarizer specifically implemented for English; however, we adapted it to work for Spanish in order to carry out the experiments proposed in this paper.
- *Language-specific Summarizer*: This can be considered as a TS approach rather than a end-user TS system. It was proposed in Lloret et al. (2011), and due to the good results it obtained for multi-lingual newswire summarization, we decided to evaluate it in the context of this research. Basically, this TS approach takes into account language-specific resources, i.e., Named Entity Recognizers and specific Knowledge Bases, such as WordNet (Fellbaum, 1998) or EuroWordNet (Ellman, 2003) for concept identification. The relevance of a sentence is assigned based on the occurrence of relevant named entities and concepts. Finally, as in the previous summarizers, the highest scored sentences are extracted until a specific summary length.

5. Evaluation and discussion

To quantify to what extent current TS techniques would be appropriate to generate an automatic tweet, a very brief summary (i.e., with only 140 characters) has to be produced for each of the news articles in the corpus. Since all the evaluated TS systems follow an extractive approach, such summary will take the form of a single short sentence. In light of this, the main limitation is that the most important sentence detected by the summarizers may not fit within the length restriction. In this case, a possible decision is either to shorten the sentence or to select the most relevant sentence among those which satisfy the 140 character length. In order to avoid generating incoherent or incomplete tweets, we opt for the latter strategy, and we left for future work the task of regenerating language.

Therefore, a pool of potential tweet sentences is first built, keeping only those ones whose length do not surpass 140 characters.

After filtering out the sentences above 140 characters, the TS systems described in Section 4 were used to determine the relevance of the remaining sentences, and the top score one was extracted to be the final tweet. This means that we select as a tweet the most relevant sentence of the pool of potential tweet sentences. In addition to the systems described in Section 4, a TS method based on the term frequency, which selects as important sentences those ones that contains high frequent words, was considered as a baseline. It is important to mention that the headline of the full-text news was not considered as an additional baseline,

[7] http://swesum.nada.kth.se/index-eng.html.
[8] http://libots.sourceforge.net/.
[9] http://www.microsoft.com/education/autosummarize.aspx
[10] http://www.extractorlive.com/upload_demo.html.
[11] http://intime.dlsi.ua.es:8080/compendium/.

because this headline was indeed the tweet produced by the newspaper Websites, and therefore we also evaluate them.

Regarding the evaluation, our aim was to determine the performance of the different TS approaches in a quantitative and qualitative manner. On the one hand, for the quantitative assessment, we focused on testing the informativeness of the automatic tweets by comparing them to the original tweets already available for the news. On the other hand, we also performed a qualitative evaluation through a human assessment of the tweets. For this, a user survey was designed with the purpose of knowing users' opinions with respect to different criteria.

5.1. Quantitative evaluation

For the quantitative evaluation, we considered the original tweet for each news as a gold-standard, and we employed the ROUGE tool (Lin, 2004). ROUGE allows us to compare the content of an automatic summary with respect to an ideal one, according to several measures based on different n-grams lengths. The most common ones are: unigrams (ROUGE-1), bigrams (ROUGE-2), bigrams with a distance of 4 words between them (ROUGE-SU4), and the longest common subsequence (ROUGE-L). In our evaluation, we computed the recall value of ROUGE-1, ROUGE-2, ROUGE-SU4 and ROUGE-L metrics. This manner, we could ensure that the most similar tweets to the original ones may be informative enough for providing the gist of the news article.

Table 5 shows the performance (recall) of the TS systems analyzed for English and Spanish.

As it can be seen, the TS system that generates the best tweets for English is the "Open Text Summarizer", being statistically significant with respect to the baseline at a 95% confidence level (t-test). It is important to stress that the difference in the results between "Open Text Summarizer", "COMPENDIUM" and "SweSum" are not statistically significant. For Spanish, the baseline overperforms all the remaining TS systems, although the results obtained are not statistically significant compared to these systems, except for "Extractor". Although the general ROUGE results for both languages are quite similar, it is worth observing the little difference (not statistically significant) between the results achieved by "Open Text Summarizer" system (the best system results for English) with respect to the baseline in Spanish.

For both languages, English and Spanish, the recall results range from 0.12000 to 0.16000, approximately for ROUGE-1. However, we observed that the similarity between automatic tweets and the original ones was lower for Spanish. This is explained by the differences in the writing style of each language. Sentences in the Spanish news documents are usually longer than in English, and therefore, when we selected the ones not surpassing 140 characters, less candidate sentences were obtained, and therefore, the loss of information may have a greater impact.

The quantitative evaluation provided us an idea of which TS systems could generate more informative tweets, assuming that informativeness was determined by the information contained in the original tweets (i.e., headlines). However, this does not imply that such tweets are the best ones from a human perspective or they attract the user's attention to read the whole news. Therefore, in order to assess other criteria apart from informativeness, we also performed a qualitative evaluation.

5.2. Qualitative evaluation

For this type of evaluation, a user survey was designed, where each tweet was rated according to the two questions shown in Table 6. These questions were about the topic of the tweet (i.e., its indicativeness), and the interest it could be arisen in reading the full-text of the news after having a look at the tweet (i.e., its interest). We would like to note that concerning our second criteria, we do not want to include in this group those tweets that are not understandable, and therefore one has to go to the full-text of such news to have an idea of what it is about.

A group of 16 external users participated in the evaluation. For each question, the tweets were rated by two users with respect to a binary classification (*"yes"* or *"no"*). The users were shown the tweet, the link to the full-text of the news, and the two questions. They were required to access to the full-text of the news through the URL where the tweet came from, so that manner they could better evaluate the tweets. Moreover, they were given no clues about the method employed for generating the tweets, so they did not know whether they were the original or the automatically-generated tweets.

Table 7 reports the results of our survey, showing the percentage of tweets that were rated as *"yes"*. Specifically, these figures represent the percentage of tweets that were good with respect to the evaluated criteria. For computing these figures, a tweet was considered to be good only if the two assessors rated it with a *"yes"* value.

As far as English is concerned, we observed that, more than 60% of the tweets generated by "COMPENDIUM" and the "Language-specific summarizer" are indicative of the topics of the full news. This contrasts to the results obtained in the quantitative evaluation, where these summarizers were not the best ones. This is explained by the fact that the vocabulary contained in these tweets is not identical to the vocabulary stated in the original tweets, and therefore, ROUGE results decrease. Moreover, it is worth noting that the percentage of orginal tweets that were rated as good concerning their indicativeness is slightly lower, but still is among the three best approaches. This is not the case for Spanish, where the original tweets were rated the best with respect to their indicativeness. Nevertheless, they are followed by the tweets generated by "COMPENDIUM" and the "Language-specific summarizer", which also obtained higher results than 45%. As it can be seen, it seems that these summarizers are the most suitable ones for producing indicative tweets.

Regarding the interest generated by the tweets, we generally obtained low results. This means that most of the tweets did not

Table 5
ROUGE results (recall) of the automatic tweets (R-1 = ROUGE-1; R-2 = ROUGE-2; R-L = ROUGE-L) (the best results are highlighted in boldface).

TS system	English			Spanish		
	R-1	R-2	R-L	R-1	R-2	R-L
Original (upper bound)	1.000	1.000	1.000	1.000	1.000	1.000
Baseline (TF)	0.132	0.0244	0.119	**0.158**	**0.037**	**0.138**
SweSum	0.147	0.027	0.121	0.134	0.034	0.118
Open Text Summarizer	**0.162***	**0.036***	**0.142***	0.149	0.029	0.131
AutoSummarize	0.124	0.024	0.110	0.135	0.025	0.115
Extractor	0.133	0.019	0.116	0.126	0.019	0.108
COMPENDIUM	0.140	0.024	0.125	0.142	0.017	0.127
Lang-specific summarizer	0.132	0.024	0.117	0.144	0.023	0.126

Table 6
Questions for the manual assessment of the tweets.

Indicativeness	When reading the tweet, is it easy to identify the topics of the news?
Interest	Is the tweet interesting? that is, after reading it, are you curious and would you like to know and read more about the news?

Table 7
Survey results: percentage of tweets that were rated as "yes" (the best results are highlighted in boldface).

Tweet	English		Spanish	
	Indicativ.	Interest	Indicativ.	Interest
Original	55.56	27.42	**66.23**	22.38
Baseline (TF)	43.50	**38.50**	35.42	30.00
SweSum	23.00	14.00	34.92	29.03
Open Text Summarizer	46.67	7.14	26.51	21.95
AutoSummarize	36.23	28.00	39.62	16.28
Extractor	30.43	31.75	40.68	19.64
COMPENDIUM	**61.02**	36.07	46.48	**41.94**
Language-specific summarizer	**66.15**	35.62	58.97	**46.03**

Table 8
Examples of automatic tweets

Indicative and interesting	Robert Peston, financial expert and BBC business editor, said none of the bank's top executives would be receiving their bonuses this year.
Indicative but not interesting	The Barcelona coach, Pep Guardiola, said: "He's the best [ever]". There is no other like him. The numbers speak for themselves.
Not indicative but interesting	Having got that off my chest, I feel a little better now.
Not indicative and not interesting	By Oliver Smith, Lonely Planet Magazine.

attract the interest of the reader, thus going unnoticed. For English, the highest result was obtained by the baseline, for which around 38% of the tweets were interesting, followed by the tweets generated by "COMPENDIUM" and the "Language-specific summarizer". For Spanish, "COMPENDIUM" and the "Language-specific summarizer" were again the TS systems which generated the best tweets, achieving around 41% and 46%, respectively. This means that both systems could be the most appropriate ones to be used for producing automatic tweets.

From this evaluation, we would like to note two additional issues: (i) the best TS systems in the quantitative evaluation are not the best ones in the qualitative evaluation. This is due to the fact that the quantitative assessment is carried out assuming that original tweets are the best ones, whereas from the qualitative evaluation it has been shown that we can produce tweets completely different from the original but equally good, or even better. (ii) Original tweets, even though they are directly extracted from the headlines of the news and are very informative, might not be the most appropriate way of communicating a news through the new social media channels, such as Twitter. From a human perspective, external factors could be influencing the qualitative evaluation, such as background knowledge, interest, familiarity with Twitter, etc. This makes that this type of evaluation is subjective rather than objective. However, it allows us to have an idea of how humans perceive the information in the social media, and also whether the information that automatic TS systems could be used with the purpose of adding value to an original headline that is spread via Twitter. Having a look at the "interest" criteria, it can be observed that the original tweets were not among the best performing tweets. It was obtained that only the 27.42% and 22.38% of the tweets were interesting, for English and Spanish, respectively. This would correspond to the 5th and 4th ranking position out of the 8 tested approaches. Although they are good in capturing the main gist of a news, sometimes this is not sufficient for capturing the attention of the reader.

Analyzing in detail the results of the user survey, we observed that we could distinguished between four cases: (i) there were tweets that were rated both as indicative regarding the topic and also interesting; (ii) tweets neither indicative nor interesting; (iii) tweets where the topics were not directly clear, but the users found them catchy and interesting to read more about; and (iv) tweets, which were very indicative, but on the contrary, there was no interest in them. These cases affect both the automatic tweets and the orginal tweets directly derived from the news headline. Table 8 shows several examples of automatic tweets showing the identified types of tweets.

The classification of the tweets in one of these groups will strongly depend on the user preferences or his/her knowledge about a topic, as well as the purpose of the tweet. For instance, in a more formal context, a more informative and structured tweet may be more appropriate, whereas in other environments, informality might not be a problem or even capturing the interest of a user might have the highest priority. These issues make us to believe that it would be more useful the automatic generation of personalized tweets according to users' preferences, which was out of the scope of this study. However, this aspect would be an interesting issue to be analyzed in the future.

6. Conclusion and future work

This article presented an initial analysis and comparison of several extractive summarizers that were employed for generating automatic tweets. We focused on the journalism genre, and therefore we produced tweets from news documents for two languages, English and Spanish. Our purpose was to analyze to what extent TS is appropriate for performing this task.

Due to the novelty of the task, we created a corpus of news documents and their corresponding tweets for being tested in our experimental framework. Through the analysis and experiments conducted, we provided some insights concerning the characteristics of the generated tweets from a quantitative and qualitative perspective. On the one hand, we assumed that the original tweet suggested by the newspaper online Website (i.e., the headline of the news article) contained all the necessary information to be considered very informative. Therefore, in our quantitative evaluation, we assessed if similar information was also covered by the automatic tweets, employing the ROUGE evaluation tool. On the other hand, we also wanted to check what real users thought about the tweets (both the original and the automatic ones) in terms of their indicativeness and interest. For this, we carried out a user survey, finding out that some summarizers (i.e, COMPENDIUM or the "Language-specific summarizer") were suitable for producing indicative as well as interesting tweets, overperforming the original ones. Although the suitability of a tweet would be tightly related to the context where it will be disseminated and the purpose of it, we showed that different tweets from the headlines could be also suitable, and even more interesting. It was observed that the more similar a tweet is to the original one does not necessarily mean that it is better. As it was shown, the original tweets, which were considered as model tweets from the point of view of their informativeness, were not among the most interesting ones regarding the human evaluation.

Moreover, thanks to this evaluation, we could identify four types of tweets, that could be used differently depending on the specific needs, opening up the opportunity to address the task of personalized tweet generation. For instance, an indicative and interesting tweet could be appropriate for attracting the attention of the reader, since s/he would probably be curious about the content of the whole news article.

Despite having shown that current TS techniques could be used as a starting point to the automatic generation of tweets, there is still a lot of room for improvement. The key challenging issue to be tackled in the future is how to address abstractive summarization together with natural language generation, in order to improve the automatic generation of a tweet, so the 140 character length allowance for tweets can be filled with the suitable information, or biased it according to the user preferences (generation of personalized tweets). Furthermore, apart from investigating these issues, we would also like to analyze if there exists any correlation between the informativeness and interest. This will be done analyzing a higher number of tweets and extending the evaluation framework.

Acknowledgements

This research work has been partially funded by the Spanish Government (“Ministerio de Economía y competitividad) through the project “Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano” (TIN2012–31224), and by the Valencian Government through projects PROMETEO (PROMETEO/2009/199) and ACOMP/2011/001. We would also like to thank the users who participated in the manual evaluation.

References

Aitamurto, T., & Lewis, S. C. (2013). Open innovation in digital journalism: Examining the impact of Open APIs at four news organizations. *New Media & Society, 15*(2), 314–331.

Bandari, R., Asur, S. & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. In *Proceedings of the 6th international conference on weblogs and social media, Dublin, Ireland*.

Chakrabarti, D. & Punera, K. (2011). Event summarization using tweets. In *Proceedings of the 5th international conference on weblogs and social media*.

Connor, B. O., Krieger, M., & Ahn, D. (2010). TweetMotif: Exploratory search and topic summarization for twitter. *Artificial Intelligence (May)*, 384–385.

Dorr, B., Zajic, D., & Schwartz, R. (2003). Hedge trimmer: A parse-and-trim approach to headline generation. *Proceedings of the HLT-NAACL 03 on Text summarization workshop* (Vol. 5, pp. 1–8). Association for Computational Linguistics.

Ellman, J. (2003). Eurowordnet: A multilingual database with lexical semantic networks. In Piek vossen (Ed.) (Vol. 9, pp. 427–430). kluwer academic publishers. 1998. Natural Language Engineering.

Fellbaum, C. (1998). *WordNet: An electronical lexical database*. Cambridge, MA: The MIT Press.

Fuentes, M., Massot, M., Rodríguez, H. & Alonso, L. (2003). Mixed approach to headline extraction for duc 2003. In *Proceedings of document understanding conference (DUC 2003), Edmonton, Alberta, Canada*.

Ganesan, K., Zhai, C., & Viegas, E. (2012). Micropinion generation: An unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 869–878). New York, NY, USA: ACM.

Hahn, N. (2013). What good is Twitter? The value of social media to public service journalism. Tech. rep., European Broadcasting Union.

Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business Horizons, 54*(3), 265–273.

Hassel, M. (2007). Resource lean and portable automatic text summarization. Ph.D. thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden.

Holmqvist, K., Holsanova, J., Barthelson, M., & Lundqvist, D. (2003). *The mind's eye: Cognitive and applied aspects of eye movement research*. Amsterdam: Elsevier Science. Ch. Reading or Scanning? A study of newspaper and net paper reading, pp. 657–670..

Ifantidou, E. (2009). Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics, 41*(4), 699–720.

Ittoo, A., & Bouma, G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Systems with Applications, 40*(7), 2530–2540.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of association of computational linguistics text summarization workshop* (pp. 74–81).

Liu, F., Liu, Y. & Weng, F. (2011). Why is “SXSW” trending? Exploring multiple text sources for twitter topic summarization. In *Proceedings of the workshop on language in social media* (pp. 66–75).

Lloret, E. (2011). Text summarisation based on human language technologies and its applications. Ph.D. thesis, University of Alicante.

Lloret, E. & Palomar, M. (2011). Finding the best approach for multi-lingual text summarisation: A comparative analysis. In: *Proceedings of the international conference recent advances in natural language processing, Hissar, Bulgaria* (pp. 194–201).

Lofi, C., & Krestel, R. (2012). iParticipate: Automatic tweet generation from local government data. In *Database Systems for Advanced Applications. Lecture Notes in Computer Science* (Vol. 7239, pp. 295–298). Berlin Heidelberg: Springer.

Lopez, C., Prince, V., & Roche, M. (2012). Just title it! (by an online application). In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 31–34). Stroudsburg, PA, USA: Association for Computational Linguistics.

Morris, A. H., Kasper, G. M., & Adams, D. A. (1992). The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research, 3*(1), 17–35.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I. & Shrimpton, L. (2013). Can twitter replace newswire for breaking news? In *Proceedings of the 7th international AAAI conference on weblogs and social media*.

Romero, M., Moreo, A., Castro, J., & Zurita, J. (2012). Using wikipedia concepts and frequency in language to extract key terms from support documents. *Expert Systems with Applications, 39*(18), 13480–13491.

Sharifi, B., Hutton, M.-A. & Kalita, J. (2010). Summarizing microblogs automatically. In *Proceedings of the annual conference of the North American chapter of the association for computational linguistics* (pp. 685–688).

Soricut, R., & Marcu, D. (2007). Abstractive headline generation using widl-expressions. *Information Processing & Management, 43*(6), 1536–1548.

Tseng, Y.-H. (2010). Generic title labeling for clustered documents. *Expert Systems with Applications, 37*(3), 2247–2254.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval, 2*(4), 303–336.

Wan, S., Dras, M., Paris, C. & Dale, R. (2003). Using thematic information in statistical headline generation. In *Proceedings of the ACL 2003 workshop on multilingual summarization and question answering* (Vol. 12, pp. 11–20).

Weng, J. -Y., Yang, C. -L., Chen, B. -N., Wang, Y. -K. & Lin, S. -D. (2011). IMASS: An intelligent microblog analysis and summarization system. In *Proceedings of the ACL-HLT 2011 system demonstrations* (pp. 133–138).

Woodsend, K., Feng, Y. & Lapata, M. (2010). Title generation with quasi-synchronous grammar. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 513–523).

Woodsend, K. & Lapata, M. (2010). Automatic generation of story highlights. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 565–574).

Yatsko, V., & Vishnyakov, T. (2007). A method for evaluating modern systems of automatic text summarization. *Automatic Documentation and Mathematical Linguistics, 41*, 93–103.