An Analysis of Twitter Corpora and the Differences between Formal and Colloquial Tweets*

Análisis de Varios Corpus de Twitter y las Diferencias entre Tweets Formales y Coloquiales

Meritxell Gonzàlez
Oxford University Press, Oxford, United Kingdom
Universitat Politècnica de Catalunya, Barcelona, Spain
meritxell.gonzalezbermudez@oup.com

Abstract: This work reviews recent publications addressing the Twitter translation task, and highlights the lack of appropriate corpora that represents the colloquial language used in Twitter. It also discusses the most well-know issues in the Twitter genre: the use of hashtags and the amount of OOVs, with especial focus in comparing the differences between formal and colloquial texts.

Resumen: Este trabajo resume las publicaciones recientes en el área de la traducción automática de tweets, destacando la falta de un corpus que represente el lenguaje coloquial presente en Twitter. También se tratan los problemas más conocidos del género de Twitter: el uso de hashtags i la gran cantidad de palabras OOV, con especial enfoque en las diferencias entre tweets formales y coloquiales.

Keywords/Palabras clave: corpus, tweets, hashtags, and OOV

1 Introduction

The success and increasing popularity of microblogging has raised the need to analyse and process its content. Traditional methods for natural language processing fail when applied over these texts. The reason is not circumscribed to few nor simple issues. Roughly, microblogs documents do not follow the traditional structure of a formal text or document, they use a number of language variants, styles and registers among other linguistic phenomena, and can even include multimedia content as a way of communication (Jehl, 2010; Gotti, Langlais, and Farzindar, 2014; Kaufmann and Kalita, 2010; Bertoldi, Cettolo, and Federico, 2010).

Machine Translation (MT) is a hard task within the natural language processing field. It has received considerable attention during the last decades, and it is still an active field with many research challenges. As in other natural language processing tasks, it counts among its difficulties the ambiguity of the language, and the need of corpora and a gold standard. The former can be addressed by analysing the context in which a sentence occur, while the second has been typically addressed by combining large amounts of general purpose data and smaller subsets of domain specific datasets. The creation of a gold standard in MT requires the use of parallel data that helps to assess the quality of the output.

When addressing the automatic translation within the microblogging genre, one has to deal with the additional difficulty of having little or no context and the fact that microblogs exhibit fleeting domains. Twitter is not different from other microblogs, and has, in addition, its own particularities. As described in (Jehl, 2010), tweets actually share the spontaneity and expressiveness of the spoken language, but limited to 140 characters. Due this constraint, tweets have usually a very simple syntax. However, they are mined of ungrammaticalities, misspellings and an unlimited number of lexical variants created out of the human imaginary and the common ground of part of the audience.

In this document, Section 2 summarises recent studies in this field and different approaches followed to address these phenomena. Next, Sections 3 to 5 give a numerical analysis of 6 different corpora of tweets written in Basque, Catalan, and Spanish. The goal of this analysis is to

* This work was partially funded by the TACARDI project (TIN2012-38523-C02) of the Spanish Ministerio de Economía y Competitividad.

sketch the content of the Twitter messages (tweets), highlight which are their principal characteristics and discuss the differences between formal and colloquial tweets.

2 Recent Work on Twitter Translation

The automatic translation of tweets, in general, is more difficult than regular MT. Although the MT community has already addressed the translation of tweets, there are still few works in this area, mainly because of the lack of corpora, and especially those showing a fair representation of colloquial texts. The number of authors publishing content in multiple languages is not small, but their messages tend to be correct and well structured, in contrast to those posted by the gross of the users.

2.1 Twitter Corpora

The availability of parallel corpora for Twitter is growing but still scarce. The following four works gathered parallel data following diverse approaches, but them all contain formal texts only. (Gotti, Langlais, and Farzindar, 2013) gathered data from Canadian Government Agencies, written in French and English. This work describes an MT system that uses in-domain parallel data crawled from the links appearing in the tweets. Hence, tuning was conducted with documents from the same domain. The corpus built in (Ling et al., 2014) contains tweets written in Chinese *and* English. This work describes a tool and a methodology to help users to identify parallel excerpts in the messages and to annotate their boundaries. The data obtained with this method was fairly cheap (crowd-sourced) and it resulted to have a high degree of quality. (Jehl, Hieber, and Riezler, 2012) used a corpus of Arabic sentences that were manually translated into English. The data was crawled by filtering the topic (Arabic Spring) and was cleaned and pruned, also by means of crowd-sourcing. Finally, the shared task described in (Alegria et al., 2015) distributed a collection of parallel corpora in the languages spoken in the Iberian peninsula. These corpora have been used in this study and they are detailed in Section 3.

In contrast, the following four works deal with the noisy input from colloquial texts, but either they do not belong to the Twitter genre or they do not contain parallel data. (Kaufmann and Kalita, 2010) describes an MT system able to translate from colloquial English into standard English. The rationale is that traditional NLP techniques can be applied over standardised text. Their methodology includes the use of aligned data from a corpus of SMSs that contains most common acronyms and short forms. (Bertoldi, Cettolo, and Federico, 2010) and (Formiga and Fonollosa, 2012) address the problem of translating noisy input. The former by trying to simulate and generate noisy input automatically; the latter by adding a preprocessing layer to convert the input into clean text. Finally, the corpus described in (Alegria et al., 2014) was distributed to the participants of the TweetNorm shared task. This is a monolingual corpus of Spanish tweets. Since this corpus has been used in this study it is further detailed in Section 3.

2.2 Linguistic Phenomena

Although the previous works addressed different problems, they share a common ground on the principal difficulties of the Twitter genre. First, the translation of hashtags is an open issue that includes its segmentation, identification and analysis of its role in sentences (Gotti, Langlais, and Farzindar, 2014). Second, the correct tokenisation of the text is essential but difficult due the extreme noisiness of the text. Also, making the translation fit in 140 characters can harm the quality of the output, although (Jehl, 2010) addressed this issue in her thesis and reported good results.

The increasing interest in the field has promoted the design of tools to create especialised corpora. However, the human translation of tweets also raises open questions (Šubert and Bojar, 2014). For instance, how to translate idioms and slang, out-of-vocabulary words, onomatopoeias, emphasises (*jajaaaaa*), or irony. But also, how to approach the translation of hashtags and symbols (such as emoticons), how to interpret wrong syntax, find the translated version of a link, and fit the final translation into 140 characters, among others.

All in all, the creation of synthetic corpus to simulate these phenomena seem a feasible approach (Bertoldi, Cettolo, and Federico, 2010), yet out of the scope of this study. Last,

but not least, an appropriate methodology and measures to assess the quality of Twitter translations including its particular characteristics has not been addressed so far.

3 Description of the Used Corpora

The next sections analyse six datasets of tweets from the Tweet-Norm (Alegria et al., 2014), Tweet-MT (Alegria et al., 2015) and Social Media (Saurí, 2013) corpora. The goal is to discuss a few of the phenomena mentioned in the previous section.

A set of four datasets was obtained from the Tweet-MT corpora. It consists of 2 bitexts for Catalan–Spanish and Basque–Spanish language pairs. The four datasets contain both, the development and the test sets for each language: CAES.ca, CAES.es, EUES.eu and EUES.es. The tweets in these datasets were obtained from a sample of manually selected accounts of authors that tend to tweet in various languages, being namely public organisations and personalities. Hence, the content of the messages is mainly formal, i.e., they do not contain misspellings and do not abuse of the use of symbols.

The fifth dataset, TNORM, was obtained from the Tweet-Norm corpus that gathered a random selection of geolocated tweets within the Iberian peninsula, excluding multilingual areas where other languages than Spanish are spoken. The corpus was processed to identify and annotate out-of-vocabulary words. Hence, it contains not only correct messages, but also colloquial ones. The dataset used in this work contains the two development sets and the test provided in the workshop.

The last dataset used in this work is TSM. It is a portion of the Social Media Corpus, and in particular the corpus of tweets in Spanish. It contains a general domain set of tweets randomly selected. So similarly to TNORM, it contains both formal and colloquial tweets. They were manually processed to classify them according to the language of the tweet and annotate different layers such as communication function, polarity, target, and topic. This process included some clean up of the twitter mark-up for privacy reasons. Hence, the author id and user mentions, hashtags and URLs were substituted with the labels @USER, #HASHTAG and [URL], respectively.

	CAES.ca	**CAES.es**
# tweets	4,000	4,000
# tokens	66,559	66,113
avg. tokens/tweet	16.39	16.53
	EUES.eu	**EUES.es**
# tweets	4,000	4,000
# tokens	58,368	51,782
avg. tokens/tweet	14.59	12.94
	TNORM	**TSM**
# tweets	1,132	8,571
# tokens	14,497	123,679
avg. tokens/tweet	12.80	14.43

Table 1: Statistics on number of tweets and tokens in each corpus.

The six datasets were processed to have similar characteristics: the tokens that correspond to the author id and RT (re-tweet) were removed when present, and they were tokenised using an adaptation to Spanish and Catalan languages of the *Twokenize* tool (O'Connor, Krieger, and Ahn, 2010). Table 1 shows the number of tweets, the number of tokens and the average number of tokens per tweet in each corpus. Regardless the differences in nature of the datasets and their size, they show a similar number of tokens per tweet, being CAES.es the dataset with longer ones and TNORM the shortest.

Although tweets are similar in length, a deeper analysis of their content shows remarkable differences between the formal and the colloquial corpora. This section analyses the use of user mentions and URLs whereas Section 4 analyses the use of hashtags. Although dealing with user mentions (@user) and links is not a big issue, they are discussed here to stand out how they are used in Twitter. Table 2 gives the figures for the use of @user and URLs in the body of the messages. @user do not seem to follow any pattern. The number of @user in the two bitexts of the TweetMT datasets is opposite: the EUES datasets contain more than twice @user than the CAES ones, and almost three times the proportion of @user with respect to the number of tokens. Similarly, the TNORM dataset shows a higher use of @user than the TSM one. It is worth to note that not all @user tokens have their counterpart in the translated text, even though this token

	CAES.ca	CAES.es
# @users	743	873
avg. @users/tweet	0.18	0.22
% @users wrt. tokens	1.13%	1.32%
# URLs	3,511	3,525
avg. URLs/tweet	0.88	0.88
% URLs wrt. tokens	5.36%	5.33%
	EUES.eu	**EUES.es**
# @users	1,947	2,070
avg. @users/tweet	0.49	0.52
% @users wrt. tokens	3.76%	3.55%
# URLs	3,461	3,458
avg. URLs/tweet	0.86	0.86
% URLs wrt. tokens	6.68%	5.92%
	TNORM	**TSM**
# @users	665	3,439
avg. @users/tweet	0.59	0.40
% @users wrt. tokens	4.59%	2.78%
# URLs	69	743
avg. URLs/tweet	0.06	0.09
% URLs wrt. tokens	0.47%	0.60%

Table 2: Statistics on user mentions (@users) and URLs use in each corpus.

does not need to be translated.

In contrast, the use of URLs seems to be consistent across the two types of datasets. The four bitexts contain almost the same number of URLs, and we can find almost one URLs in each tweet. In return, TNORM and TSM contain a remarkable small number of URLs, less than 0.1% per tweet. Out of curiosity, the majority of URLs in the bitexts link to documents in the same language as the tweet. Given that the selected authors post multilingual messages, it seems reasonable that they also link to the right URL when available.

4 *On the Importance of the Hashtag Occurrences*

This section analyses the use of hashtags in the datasets. This study and the next one in Section 5 follow the procedure in (Gotti, Langlais, and Farzindar, 2014) that resulted very clear and appropriate to this end. Table 3 shows some statistics on the occurrences of hashtags. The different number of hashtags between formal and colloquial datasets is noticeable. The former contains more than one hashtag per tweet, whereas the latter contains a remarkable

	CAES.ca	CAES.es
# hashtags	3,286	3,821
# hashtag types	198	430
avg. hashtags/tweet	0.82	0.96
% hashtags wrt. tokens	5.01%	5.78%
# tweets > 1 hashtag	1,520	1,750
	EUES.eu	**EUES.es**
# hashtags	4,828	4,608
# hashtag types	584	438
avg. hashtags/tweet	1.21	1.52
% hashtags wrt. tokens	8.27%	8.90%
# tweets > 1 hashtag	2,358	2,364
	TNORM	**TSM**
# hashtags	182	1,046
# hashtag types	157	1
avg. hashtags/tweet	0.16	0.12
% hashtags wrt. tokens	1.26%	0.85%
# tweets > 1 hashtag	103	744

Table 3: Statistics on hashtag use in each dataset.

low number of them.[1] It seems to indicate that formal tweets tend to use hashtags to categorise its topic and, maybe, create a trend. This is also reflected in Figure 1: the most of the formal tweets, in the bitexts, contain one or two hashtag, whereas the most of the colloquial ones have none.

A more interesting issue is the translation of hashtags. In terms of the number of occurrences, each side of the bitexts contain a similar amount. However, the number of hashtag types in CAES.ca is much lower than the ones in CAES.es. A peer review of the hashtag sets reveals that the Spanish versions contain more written variants than their counterparts in Catalan. For instance, the hashtag "#revistapremsa" (Catalan) has four variants in the Spanish text: "#revista", "#revistadeprensa", and "#revistaprensa".

According to (Gotti, Langlais, and Farzindar, 2014), hashtags can be classified by the role they play in the text. They distinguish between hashtags that appear at the beginning of the text (prologue), in the text (inline) and at the end of the text (epilogue). Correctly identifying this role is important since a number of hashtags may have a syntactic function inside the text (inline), or can help to identify the domain of the text (prologue and epilogue). A simple heuristic was used to split the tweets into

[1] The number of hashtag types in TSM is 1 because the corpus contains only the *#HASHTAG* label.

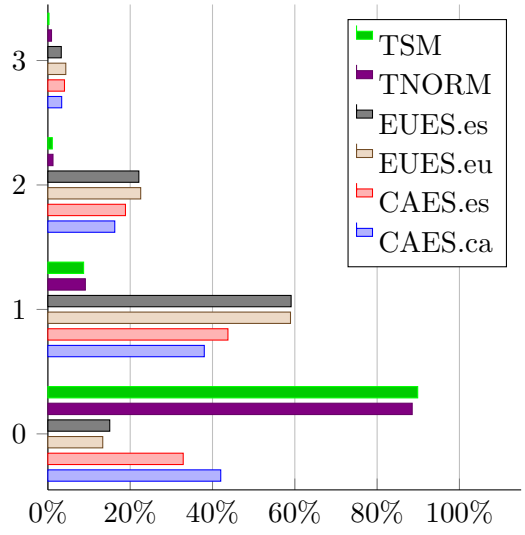

Figure 1: % tweets with exactly n hashtags, for $n \in [0, 1, 2, 3]$.

these three parts, and the results shown are in line with the mentioned study. We can observe, in Table 4, how the hashtag role within the text varies in each corpus. Although in different proportion, the gross of hashtags in the formal datasets appear in the epilogue, which indicates there is a common practice to add any hashtag at the end of the tweet. In contrast, the colloquial datasets have a very few proportion of tweets with either a prologue or an epilogue, but a higher proportion of them appear in the prologues (in comparison to the formal tweets). This behaviour may simply indicate that colloquial tweets do not follow necessarily any common practice. All datasets actually exhibit a low rate of tweets having a prologue, although the EUES bitext show a remarkable higher number in comparison to the rest. Finally, it is worth to note that, although the number of hashtags is lower in the colloquial texts, roughly half of them appear inline, and hence, they play a syntactic role in the message. This is important since they may contain an essential part of the semantics and thus worth to deal with them. Unfortunately, hashtags contains mainly of out-of-vocabulary words, as discussed next in Section 5.

5 *On the OOV words in Twitter*

The use of out-of-vocabulary (OOV) words in Twitter has been claimed to be a hard issue. The reason is not only the high number of misspellings, symbols and orthographic

	CAES.ca	**CAES.es**
% tweets with a prologue	2.85%	3.42%
% tweets with an epilogue	43.6%	49.48%
% of hashtags in a prologue	3.50%	3.61%
% of hashtags in an epilogue	75.72%	73.46%
	EUES.eu	**EUES.es**
% tweets with a prologue	10.28%	10.90%
% tweets with an epilogue	55.23%	55.13%
% of hashtags in a prologue	9.13%	10.63%
% of hashtags in and epilogue	57.27%	60.11%
	TNORM	**TSM**
% tweets with a prologue	2.03%	2.39%
% tweets with an epilogue	5.74%	3.83%
% of hashtags in a prologues	17.03%	20.08%
% of hashtags in a epilogues	40.66%	35.09%

Table 4: Statistics on hashtag (#) use as prologues and epilogues in each dataset.

errors, that could be partially tackled by using spell-checkers, but also the use of specific lexica and lexical variants. For instance, the use of word combinations (e.g., in hashtags), the combination of different languages (especially in multilingual regions, but also English terms) and the unlimited ability of the microblogging sphere to invent new terms.

This section gives a numerical analysis of OOVs that occur in Twitter. In order to conduct this analysis, the datasets were processed to remove the user mentions and URLs, since them all are tokens that do not need to be translated. Some variants of the datasets were built. First, only the CAES bitext was used due the lack of a Language Model (LM) for Basque. Then, since the TNORM annotations provide the corrected forms for some OOV tokens (only spelling variants), they were used to build a new dataset TNORM-S were OOVs were substituted with the correct word when available. In addition, two different versions were created out of each dataset. In the first one (no #symbol), the hashtags' texts were kept (the # symbol was removed) since they play an important role in the text, carry part of the semantics of the message and need to be translated in most of the cases. In the second dataset (no hashtags), all the hashtags were removed. The purpose of this second version is to highlight the impact of hashtags in the perplexity estimation of the texts.

Table 5 shows the results of this analysis. As expected, colloquial datasets contain a

	CAES.ca	CAES.es
% OOV - no #symbol	5.61%	5.14%
% OOV - no hashtags	2.81%	2.20%
ppl - no #symbol	603	644
ppl - no hashtags	520	543
	TNORM	**TNORM-S**
% OOV - no #symbol	14.23%	12.45%
% OOV - no hashtags	13.53%	11.79%
ppl - no #symbol	1,325	1,211
ppl - no hashtags	1,300	1,192
	TSM	
% OOV - no #symbol	9.18%	
% OOV - no hashtags	8.38%	
ppl - no #symbol	1,370	
ppl - no hashtags	1,373	

Table 5: Count of OOV and perplexity (ppl) estimation in each corpus using a LM trained on the "El Periódico" corpus. (This parallel corpora is listed in the ELRA catalog as `http://catalog.elra.info/product_info.php?products_id=1122`)

higher number of OOVs. The TNORM-S contains slightly a lower number of them in comparison to the non-normalised version, which indicates that the use of spell-checkers and the substituion of lexical variants in not enough to deal with OOVs. This is reflected in the figures on the perplexity of the datasets. The perplexity is high across all the datasets, and it slightly decreases after removing the hashtags from the data, indicating that the language used in the text is notable different from the LM. This can be ascribed to the fact that the LM was build using an out-of-domain corpus. In turn, removing the hashtags from the data decreases the amount of OOVs, and seems to have an impact only in the formal dataset, where half of the OOVs occur in the hashtags. However, their proportion is smaller when compared with the colloquial datasets.

For the sake of comparison, the same calculation was carried on using a LM trained on TNORM corpus, the only corpus publicly available out of the two colloquial ones. The new LM was used to obtain the % of OOVs and perplexity estimations on CAES.es and TSM datasets. The results are shown in Table 6. The % of OOVs is higher in both cases, most probably due the small size of the corpus. However, the perplexity of the TSM dataset has decreased. This seems to indicate that the LM was able to capture a high

	TSM	CAES.es
% OOV - no #symbol	11.08%	11.26%
% OOV - no hashtags	10.30%	7.51%
ppl - no #symbol	591	735
ppl - no hashtags	591	669

Table 6: Count of OOVs and the perplexity (ppl) in the TSM and CAES.es corpora using a LM trained on the TNORM corpus.

proportion of the particular characteristics of colloquial tweets, and that these may be recurrent in the colloquial genre and do not appear in formal texts.

6 Conclusions and Further Work

Twitter has its own particularities that makes it a hard genre to deal with. This work reviews recent publications that address the problem of Twitter translation. The number of works in this field is still scarce due the lack of corpora, but also because of the lack of a gold standard and specific evaluation methodologies that can help to assess the quality of a tweet translation. This work also discusses the most well-know issues in the Twitter genre: the use of hashtags and the amount of OOVs, with especial focus on comparing the differences between formal and colloquial texts. The results obtained are preliminary, but they clearly show that these two registers are different not only from a linguistic point of view, but also in terms of tweet structure and content. Further work has to be done to align the hashtags and the OOVs in bitexts corpora and analyse the way their are translated. Also, the annotation layers of the TSM corpus enables the possibility to fine-grain the study, for instance, by focusing in the differences between tweets with different communication functions. To conclude, no major differences were found between languages, but this may be ascribed to the fact that the datasets were obtained from bitexts corpora.

References

Alegria, Iñaki, Nora Aranberri, Pere R. Comas, Víctor Fresno, Pablo Gamallo, Lluís Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2014. TweetNorm_es Corpus: an Annotated Corpus for Spanish Microtext Normalization. In *Proceedings of the*

Ninth International Conference on Language Resources and Evaluation.

Alegria, Iñaki, Nora Aranberri, Cristina España-Bonet, Pablo Gamallo, Hugo G. Oliveira, Eva Martínez, Iñaki San Vicente, Antonio Toral, and Arkaitz Zubiaga. 2015. Overview of TweetMT: A Shared Task on Machine Translation of Tweets at SEPLN 2015. In *Proceedings of the Tweet Translation Workshop co-located with 31th Conference of the Spanish Society for Natural Language Processing, Alacant, Spain, September.*

Bertoldi, Nicola, Mauro Cettolo, and Marcello Federico. 2010. Statistical Machine Translation of Texts with Misspelled Words. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the ACL*, pages 412–419. ACL.

Formiga, Lluís and José A. R. Fonollosa. 2012. Dealing with Input Noise in Statistical Machine Translation. In *Proceedings of COLING 2012: Posters*, pages 319–328, Mumbai, India, December.

Gotti, Fabrizio, Philippe Langlais, and Atefeh Farzindar. 2013. Translating Government Agencies' Tweet Feeds: Specificities, Problems and (a few) Solutions. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 80–89, Atlanta, Georgia, June. ACL.

Gotti, Fabrizio, Phillippe Langlais, and Atefeh Farzindar. 2014. Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. ELRA.

Jehl, Laura. 2010. Machine Translation for Twitter. Master's thesis, University of Edimburgh, United Kingdom.

Jehl, Laura, Felix Hieber, and Stefan Riezler. 2012. Twitter Translation Using Translation-based Cross-lingual Retrieval. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 410–421, Stroudsburg, PA, USA. ACL.

Kaufmann, Max and Jugal Kalita. 2010. Syntactic normalization of Twitter messages. In *Proceedings of the International Conference on Natural Language Processing, Kharagpur, India.*

Ling, Wang, Luis Marujo, Chris Dyer, Alan W. Black, and Isabel Trancoso. 2014. Crowdsourcing High-Quality Parallel Data Extraction from Twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 426–436, Baltimore, Maryland, USA, June. ACL.

O'Connor, Brendan, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In *Proceedings of the International Conference on Web and Social Media (ICWSM). The AAAI Press.*

Saurí, Roser. 2013. Corpus de Dominio Genérico y Específicos (Inglés, Español, Catalán y Portugués). Technical report, Social Media. Métodos y Tecnologías para los medios sociales. Programa CENIT 2010 (CEN-20101037).

Šubert, Eduard and Ondřej Bojar. 2014. Twitter crowd translation – design and objectives. In *Translating and the Computer 36*, pages 217–227, Geneva, Switzerland. AsLing, The International Association for Advancement in Language Technology, Editions Tradulex; AsLing.