TweetMT: A parallel microblog corpus

Iñaki San Vicente[1], Iñaki Alegría[2], Nora Aranberri[2], Cristina España-Bonet[3], Pablo Gamallo[4], Hugo Gonçalo Oliveira[5], Eva Martínez[2], Antonio Toral[6], Arkaitz Zubiaga[7]
[1] Elhuyar, [2] University of the Basque Country, [3] Universitat Politècnica de Catalunya,
[4] Universidade de Santiago de Compostela, [5] University of Coimbra,
[6] Dublin City University, [7] University of Warwick
tweetmt@elhuyar.com

Abstract

We introduce TweetMT, a parallel corpus of tweets in four language pairs that combine five languages (Spanish from/to Basque, Catalan, Galician and Portuguese), all of which have an official status in the Iberian Peninsula. The corpus has been created by combining automatic collection and crowdsourcing approaches, and it is publicly available. It is intended for the development and testing of microtext machine translation systems. In this paper we describe the methodology followed to build the corpus, and present the results of the shared task in which it was tested.

Keywords: Machine Translation, Microblogs, Tweets, Social Media, parallel corpus

1. Introduction

While machine translation (MT) is a mature research field now, the application of MT techniques to tweets is still in its infancy. Tweets are often written from mobile devices, which exacerbates the poor quality of the spelling, and include linguistic inaccuracies, symbols and diacritics. Tweets also vary in terms of structure, including features which are exclusive to the platform, such as hashtags, user mentions, and retweets. These characteristics make the application of MT to tweets a new challenge that requires specific processing techniques to perform effectively.

Despite the paucity of research in the specific task of translating tweets, an increasing interest can be observed in the scientific community (Gotti et al., 2013; Peisenieks and Skadiņš, 2014). Similarly, a related and highly relevant direction of research is the work on MT of SMS texts, such as Munro's study in the context of the 2010 Haiti earthquake (Munro, 2010).

Provided the dearth of benchmark resources and comparison studies bringing to light the potential and shortcomings of today's MT techniques applied to tweets, a corpus was compiled in the framework of TweetMT, a workshop and shared task[1] on MT applied to tweets. Our parallel corpus includes tweets for the following language pairs: Catalan–Spanish (ca-es), Basque–Spanish (eu-es), Galician–Spanish (gl-es), and Portuguese–Spanish (pt-es). Those are the most common pairings between official languages in the Iberian Peninsula.

2. Collecting parallel tweets

To the best of our knowledge, there is no parallel tweet dataset available apart from that produced by (Ling et al., 2013), which differs from our purposes in that they worked on tweets that mix two languages, i.e., providing the translated text within the same tweet. They further improve the quality of the parallel segments by means of crowdsourced annotations (Ling et al., 2014). Since we wanted to work on the translation of entire tweets into new tweets, we generated a corpus for the specific purposes of the TweetMT Workshop.

For corpus generation, we developed a semi-automatic method to retrieve and align parallel tweets. The first step of this method consists in identifying multiple Twitter authors that concurrently tweet in multiple languages and crawl those accounts. The second step involves aligning the collected messages.

The idea behind this methodology is that the languages involved are official in the Iberian Peninsula, being Catalan, Basque and Galician co-official alongside Spanish in the respective regions. This coexistence implies that there is a necessity for multilingualism.

In the end, our methodology was successfully applied for the Catalan–Spanish (*ca-es*) and Basque–Spanish (*eu-es*) language pairs. We manually identified accounts that tend to post messages in multiple languages. This strategy for sampling the authors leads to a prevalence of account types that belong to organizations (government, media, sport clubs, etc.) and famous personalities. Messages from those sources are mostly formal, with well structured sentences and rather infrequent use of slang[2].

Unfortunately, for Portuguese–Spanish (*pt-es*) and Galician–Spanish (*gl-es*) we could not find authors that met these characteristics. It is most surprising in the case of Galician, because being co-official with Spanish it was expected that there would be at least regional government related multilingual accounts. In these latter cases, we had to rely on crowdsourcing to produce actual parallel tweets. Due to time and budget constraints, different to the language pairs that could be automatically aligned, only test sets were generated for *gl-es* and *pt-es* pairs.

Table 1 provides detailed statistics of the datasets used for the tweetMT shared task. A second release of the dataset contains all the correctly aligned tweets.

[1] http://komunitatea.elhuyar.eus/tweetmt/

[2] (González Bermúdez, 2015) analyzes the differences between formal and colloquial tweets

Dataset	Tweets	Authors	Tokens	URL	@user
eu-es$_{\text{dev}}$	4,000	4	181K	2,622	1,569
ca-es$_{\text{dev}}$	4,000	2	161K	3,280	823
eu-es$_{\text{test}}$	2,000	16	37K	1556	673
es-eu$_{\text{test}}$	2,000	16	43K	1535	692
ca-es$_{\text{test}}$	2,000	14	45K	1590	417
es-ca$_{\text{test}}$	2,000	14	46K	1567	502
gl-es$_{\text{test}}$	434	-	7K	274	134
es-gl$_{\text{test}}$	434	-	7K	291	159
pt-es$_{\text{test}}$	1,250	-	19K	674	349
es-pt$_{\text{test}}$	1,250	-	21K	919	583

Table 1: Statistics for the datasets generated.

2.1. Corpus Creation from Multilingual Accounts

2.1.1. Accounts and Collected Data

The initial collection of tweets amounted to 23 Twitter accounts (from 16 authors) for the *eu-es* pair and 19 accounts (from 14 authors) for the *ca-es* pair. In all, 75,000 tweets were collected for *eu-es* and 51,000 tweets for the *ca-es* language pair. The collection includes tweets posted between November 2013 and March 2015. Test sets for the other languages pairs, *gl-es* and *pt-es* were collected through crowdsourcing.

2.1.2. Alignment

Aligning tweets of an author within and across accounts requires both to find matching translations as well as to occasionally get rid of tweets that have no translations. We perform this process semi-automatically, first by automatically aligning tweets that are likely to be each other's translation, and then by manually checking the accuracy of those alignments.

Before we can even align tweets with their likely translations, we needed to identify the language each tweet is written in through language identification (Zubiaga et al., 2014). We used an ngram-based language identifier[3] trained over Twitter data.

We defined a set of heuristics and statistics that would help us find matches quite accurately. Specifically, we looked at the following three characteristics to find likely matches:

- **Publication date.** Translations must be published within a certain period range to be flagged as possible translations of each other. The difference between source and target timestamps must not exceed a certain threshold. The default value of the threshold was 10 hours, although for 9 (5 for *eu-es* and 4 for *ca-es*) accounts the publication date difference was restricted to 1 hour after observing too much noise with the more relaxed standard threshold. The choice between thresholds is ultimately made based on the publication rate of the accounts. We observed that the more tweets they publish the more strict the threshold should be. This was specially so in the case of sport clubs (e.g., @FCBarcelona), with a very high tweeting rate during sport matches, where tweets are either translated within minutes or not translated at all.

- **String length similarity.** We assume translations can not differ greatly in character length, and filter out all candidates that are more 40% shorter or longer than the source message.

- **Overlap of hashtag and user mentions in source and target tweets.** A minimum number of user names and hashtags were required to overlap between source and target parallel tweet candidates. The overlap is computed as the division between the number of entities in the intersection of both tweets and the entities in the union. The threshold is empirically set to 0.76.

- **Longest Common Subsequence ratio (LCSR) between source and target tweets.** LCSR (Cormen et al., 2001) is an orthographic similarity measure, as it tells us how similar two strings are. It is especially reliable when working with closely related languages, as parallel sentences are often very close to each other, because both vocabulary and word order are close.

As for the performance of the heuristics, publication date proximity is effective for filtering out wrong candidates, but it is not enough to find the correct parallel tweet, so it is applied first. The same occurs for the string length heuristic, so it is applied as a second filter. User and hashtag overlap ratio proved successful, up to the point that the contribution of LCSR was minimal.

The output of this alignment can be manually corrected by native speakers of their respective languages. At this point, we split the initial corpus into two datasets: one development set C_{dev} composed of 4,000 parallel tweets for each language pair and one test set C_{test} composed of 2,000 parallel tweets for each language pair.

The development set is limited to accounts with most tweets (2 for *ca-es* and 4 for *eu-es*). Test sets also contain tweets from the authors in the development set, but tweets from new "unseen" authors are also introduced. This way we have the possibility to evaluate systems both on "in-domain" and "out-of domain" scenarios.

Only test sets were manually corrected. Each tweet is reviewed by a single annotator. The overall error rate over the collections manually reviewed to create the test sets was 7% for the *ca-es* language pair (12,500 tweets) and 21% for the *eu-es* language pair (15,045 tweets). The error rate in the development set is estimated as the average error rate of the Twitter accounts that are included in the collection C_{dev} computed from the annotations of those accounts in C_{test}. The error rate in C_{dev} is 3% (*ca-es*) and 33% (*eu-es*). The accounts included in the *eu-es* development set were those providing a larger number of tweets, but also quite noisy in terms of alignment. In contrast, for *ca-es*, accounts with the largest number of tweets were accurately aligned, hence the difference.

Figure 1 summarizes in a boxplot the distribution of the alignment error rates across different Twitter accounts. The boxplot shows a similar error distribution between *ca-es*

[3] http://www.let.rug.nl/vannoord/TextCat/

and *eu-es* language pairs, and in fact, the average error rate per account was 21% for *eu-es* and 18% for *ca-es*. This is explained because for *ca-es* there are two main accounts contributing to the collection and both have low error rates but the rest of the accounts have overall higher error rates. Instead, for *eu-es* source contribution is more evenly distributed.

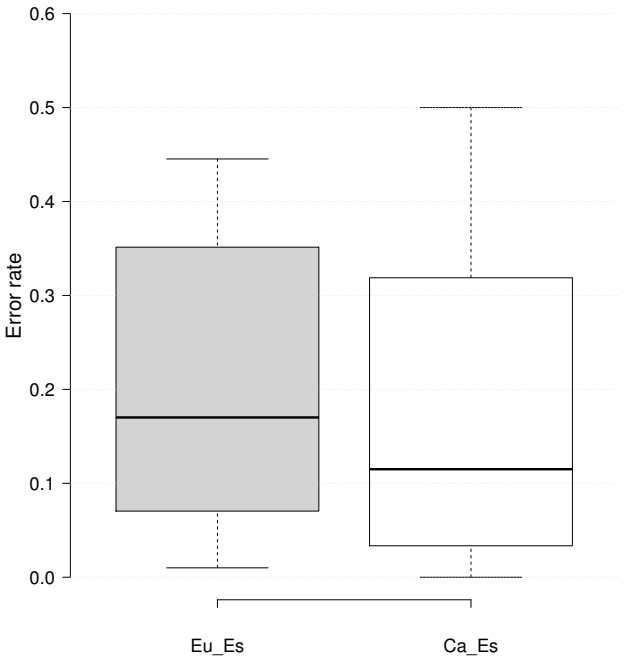

Figure 1: Alignment error rate distribution for *ca-es* and *eu-es* twitter accounts in the test datasets.

2.2. Corpus Creation using Crowdsourcing

As it was said before, bilingual Portuguese–Spanish and Galician–Spanish tweets are not generated sistematically, so we used the crowdsourcing platform CrowdFlower[4] to build the test data for these language pairs. CrowdFlower provides a fast method for collecting annotations from a broad base of paid non-expert contributors over the Web. Besides, in order to assure the quality of the translations, one can impose several constraints on the contributors. In our case we decided to impose restrictions on the countries from which workers were allowed to work (Spain for Galician and Portugal and Brazil for Portuguese); on their language capability by only using Portuguese speakers when needed[5]; and on their performance level by limiting to level 3 contributors (the top level).

In the crowdsourcing tasks for *pt-es* and *gl-es*, the selected contributors had to translate manually, from Spanish to Portuguese and Galician, a dataset with $2,552$ Spanish tweets, taken from both our *ca-es* and *eu-es* parallel corpora, and divided into working tasks of 10 tweets each. Figure 2 shows the translation interface for the *es-pt* task, with the detailed instructions that the participants were given.

[4] `http://www.crowdflower.com/`

[5] This option is not available for Galician.

These instructions were provided to workers in order to make sure that the translations were consistent. For instance, contributors were asked not to translate user mentions (keywords with a leading @) and URLs, while hashtags should only be translated if the contributor considered that it would be natural to use the corresponding Portuguese/Galician hashtag.

As a final result, we obtained a parallel corpus with $2,500$ *pt-es* and 777 *gl-es* tweets which were split into two test datasets with $1,225$ entries for each translation direction for *pt-es* and 388 for *gl-es*. To verify the quality of the translations, samples of 30 tweets were evaluated both for Portuguese and for Galician[6]. In both cases they were considered acceptable by the Portuguese and Galician authors of the current paper, even if some errors were detected. In the case of Galician, we found some mistakes derived from the new spelling rules imposed since 2003. In the case of Portuguese, six errors were found from the 30 tweets evaluated, most of them lexical problems. In example 1 'Buenos Días' was translated to 'Olá' instead of 'Bom dia'. Also 'nos llevan' was translated to the infinitive tense 'levar' instead of being inflected (should be 'levam-nos'):

Example 1
Spanish: 30 pasos lineales nos llevan de 1 a 30. 30 pasos exponenciales *nos llevan* hasta 1000 millones... y el Futuro es exponencial... *Buenos días*!
Portuguese: 30 passos lineares *levar* de 1 a 30. Os 30 passos exponenciais levar até 1 bilhão ... eo futuro é exponencial ... *Olá*!

This mistake can appear because instead of coming up with their own translation, some crowdsourcers have probably resorted to machine translation software. These kind of errors have been solved in the test sets, however, translations among languages do not need to be literal. When compared to the corpus obtained with bilingual accounts, we also see there some examples were the same meaning is expressed in different forms:

Example 2
Spanish: El Mini "A" del IDIDID gana en Calahorra, la V Olimpiada Miguel Jiménez.
Catalan: El Mini "A" del IDIDID guanya la V Olimpiada Miguel Jiménez de Calahorra.

Similarly to the case with *Buenos días* and *Olá*, in Example 2, there is a paraphrasis translation. This is common in parallel corpora, but paraphrases are not very frequent in formal tweets, because public accounts tend to tweet exactly in the same format for both languages. This effect is expected to be further minimised when translations are obtained manually through crowdsourcing instead of automatically, but even in this case some examples appear.

3. The corpus in use: Shared Task Results

The generated dataset has been used in the framework of the TweetMT shared task.

[6] It was not possible to evaluate larger samples due to time constraints of the shared task and the difficulty to acquire translations for those languages.

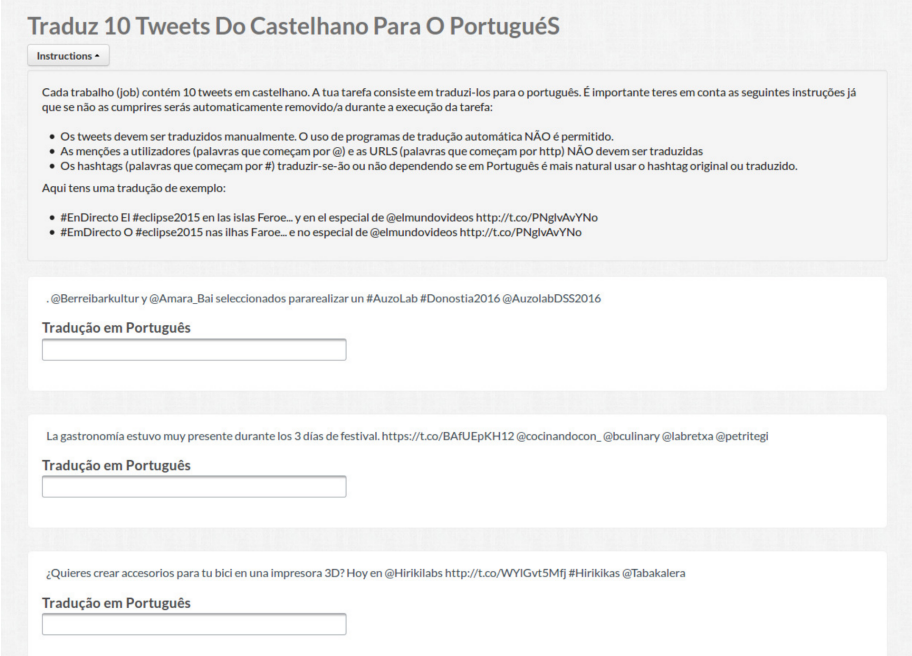

Figure 2: CrowdFlower translation interface.

3.1. Overview of the Task, Resources and Systems

As we have said before, the purpose of the shared task was to translate tweets between all the official languages of the Iberian Peninsula and, at the same time, compare different approaches to tweet MT.

To accomplish this, several resources were made available. On the one hand, the parallel corpora here introduced (cf. Section 2.) were released. Before release, test datasets were preprocessed to replace all user mentions by `IDIDID` and all URLs by `URLURLURL`. Obtaining monolingual corpora of tweets is an easier task that was delegated to the participants. On the other hand, this was an unconstrained task where any kind of data could be used only with the condition that it was reported in the system description. Links to general parallel data, dictionaries and baseline rule-based MT (RBMT) systems were facilitated. Under these conditions, the submitted systems differ not only in the approach but also in the amount of external resources used.

After delivering the test sets, participants had a window of 72 hours to return their translated results. The translated texts would then be extracted and cut down to 140 characters for the official automatic evaluation using several lexical and syntactic metrics.

Out of the 5 registered participants, three teams ended up submitting their results: DCU (Dublin City University) for 3 tracks (*ca-es*, *eu-es*, *pt-es*) (Toral et al., 2015); EHU (University of the Basque Country) for the *eu-es* track (Alegria et al., 2015b); and UPC (Universitat Politècnica de Catalunya) for the *ca-es* track (Martínez-Garcia et al., 2015). So, two teams submitted results for the *eu-es* and *ca-es* tracks, one team participated in the *pt-es* track, and no submissions were received for the *gl-es* pair. The main characteristics of the systems submitted are compiled in Table 2.

3.2. Results

Participants were allowed to submit up to three results per track. Here we outline the results in terms of the BLEU (Papineni et al., 2002) automatic evaluation metric (see Table 3) for all the tracks and systems.

DCU3 system was the best for the *ca-es* direction, a system combining two kinds of statistical MT (SMT) engines plus a RBMT one. For the *es-ca* direction, the two simplest pure phrase-based SMT systems, UPC1 and DCU2, obtained the highest scores. The two teams used very similar corpora in their experiments, so the techniques they used make the difference in this case. The best translator for the *es-eu* language pair is the statistical system EHU1 for the Basque into Spanish direction. When translating from Spanish into Basque, however, DCU2 with the combination of 5 different systems obtains very similar scores, even better than EHU. Finally, DCU3 was the best in the *pt-es* direction. As in the *ca-es* track, their best system is again a combination of two kinds of SMT engines and a RBMT one. On the opposite direction the best system, DCU2, does not include translation options from the RBMT, probably reflecting a lower quality for this engine on tweets.

Notice that the mean length of the tweets gathered for building the corpora is far from reaching the 140-character limit. The longest tweets are the Galician ones with a mean of 97 characters and the shortest ones with 88 characters correspond to Basque. Under these settings, the evaluation of the systems is not affected by our truncation to 140 characters in the submissions. In fact, the result of this truncation only changed the translations of the participants by less than 0.01%, from a mean length of 90.70 characters to 90.62. These changes, of course, are not reflected in BLEU or in any other metric used in the official evaluation (Alegria et al., 2015a).

System	Main Engine	Distinctive features
DCU1	System combination or SMT	Moses and Apertium (ES↔CA), Moses, cdec and Apertium (ES→EU), cdec (EU→ES), Moses (ES↔PT).
DCU2		Moses (ES→CA), Moses, cdec and Apertium (CA→ES, EU→ES), Moses, cdec, ParFDA, Matxin and Morph (ES→EU), Moses and cdec (ES↔PT).
DCU3		Moses, cdec and Apertium (ES→CA, ES↔PT), Moses, ParFDA and Apertium (CA→ES), Moses, cdec, Matxin and Morph (ES→EU), Moses, cdec, Apertium and Morph (EU→ES).
EHU1	SMT	Specific language model and pre- and post-processing for tweets
EHU2	RBMT	Adaptation to Tweets (mainly hashtags)
UPC1	SMT	Moses system
UPC2	SMT	Document-level system (Docent), semantic models

Table 2: Summary of the systems developed by the participants.

System	ca2es	es2ca	eu2es	es2eu	pt2es	es2pt
DCU1	76.73	75.79	25.30	23.22	43.36	36.13
DCU2	76.52	77.75	25.30	**24.44**	43.67	**37.25**
DCU3	**77.70**	75.25	25.44	23.42	**44.28**	36.94
EHU1	–	–	**28.61**	24.34	–	–
EHU2	–	–	–	19.54	–	–
UPC1	68.20	**77.93**	–	–	–	–
UPC2	63.12	–	–	–	–	–

Table 3: BLEU score for all the participant systems and language pairs. Results are obtained only considering the first 140 characters per tweet.

The analysis of the results enables us to draw several conclusions. First, the evaluations for the genre of formal tweets show better results than for other genres such as news in the same language pairs (Alegria et al., 2015b). This may be due to the fact that we aim to translate very short sentences and we have minimised the hardest particularities that would appear in informal tweets such as abbreviations and slang. Second, combining the outputs produced by systems based on different techniques, including RBMT and SMT, can lead to improvements as it generally happens in machine translation (Toral et al., 2015). Finally, the adaptation to this genre can be done in different ways. For example, expanding the context by using a user's tweets within the same day can be of use to boost the performance of the MT system (Martínez-Garcia et al., 2015).

4. Conclusion

The corpus developed as part of the TweetMT shared task has enabled us to come up with a benchmark parallel corpus of tweets for translation applied to four language pairs: *ca-es*, *eu-es*, *gl-es* and *pt-es*. The corpus is publicly available and can be downloaded from the workshop's website[7], which we expect that will enable further research in the field.

[7] `http://komunitatea.elhuyar.eus/tweetmt/resources/`

The *ad hoc* methodology followed to build the corpus has brought to light a few caveats, one of which is the difficulty to find Twitter accounts that tweet simultaneously in multiple languages, which made it unfeasible for two of our language pairs (*gl-es* and *pt-es*). However, it does result in high quality parallel corpora when such Twitter accounts are available, which can be exploited to adapt an MT system to the new domain presented by Twitter. The existence of such resources depends on the social and political balance the target languages maintain.

The results achieved by the participants of the shared task are surprisingly high, especially considering that we are dealing with tweets. Still, it is worthwhile noting that the tweets considered in this shared task can largely be deemed formal and would be difficult to generalize the results to other tweet translation tasks. However, the fact that formal tweets can be accurately translated encourages its use by community managers who tweet in different languages, by making their work easier. One of our main objectives for future work is to further generalize the MT task by including a more representative collection of tweets, to assess the ability of MT systems to translate informal tweets too.

Acknowledgements

This work has been supported by the following projects: Abu-Matran (FP7-PEOPLE-2012-IAPP), PHEME (FP7, grant No. 611233), *Tacardi* (Spanish MINECO TIN2012-38523-C02-01), QTLeap (FP7, grant No. 610516) and TELEPARES (Spanish MINECO FFI2014-51978-C2-1-R).

5. Bibliographical References

Alegria, I., Aranberri, N., España-Bonet, C., Gamallo, P., Gonçalo Oliveira, H., Martínez García, E., San Vicente, I., Toral, A., and Zubiaga, A. (2015a). Overview of tweetmt: A shared task on machine translation of tweets at sepln 2015. In *Proceedings of the Tweet Translation Workshop, at "XXXI Congreso de la Sociedad Española de Procesamiento de lenguaje natural" and CEUR Workshop Proceedings.*, pages 8–19, September.

Alegria, I., Artetxe, M., Labaka, G., and Sarasola, K. (2015b). EHU at TweetMT: Adapting MT engines for formal tweets. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Cormen, T. H., Stein, C., Rivest, R. L., and Leiserson, C. E. (2001). *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition.

González Bermúdez, M. (2015). An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7. CEUR-WS. org.

Gotti, F., Langlais, P., and Farzindar, A. (2013). Translating government agencies' tweet feeds: Specificities, problems and (a few) solutions. *NAACL 2013*, page 80.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

Ling, W., Marujo, L., Dyer, C., Black, A., and Trancoso, I. (2014). Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martínez-Garcia, E., España-Bonet, C., and Màrquez, L. (2015). The UPC TweetMT participation: Translating formal tweets using context information. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Munro, R. (2010). Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, pages 1–4.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Peisenieks, J. and Skadiņš, R. (2014). Uses of machine translation in the sentiment analysis of tweets. In *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014*, volume 268, page 126. IOS Press.

Toral, A., Wu, X., Pirinen, T., Qiu, Z., Bicici, E., and Du, J. (2015). Dublin city university at the tweetmt 2015 shared task. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Zubiaga, A., San Vicente, I. n., Gamallo, P., Pichel, J. R., ia, I. n., Aranberri, N., Ezeiza, A., and Fresno, V. (2014). Overview of tweetlid: Tweet language identification at sepln 2014. *TweetLID@SEPLN*.