Overview of TweetMT: A Shared Task on Machine Translation of Tweets at SEPLN 2015

Introducción a TweetMT: Tarea Compartida sobre Traducción Automática de Tuits en la SEPLN 2015

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

Resumen: Este artículo presenta un resumen de la tarea conjunta que tuvo lugar en el marco del taller TweetMT celebrado junto con SEPLN 2015, que consiste en traducir diversas colecciones de tweets en varios lenguajes. El artículo describe el proceso de recolección y anotación de datos, el desarrollo y evaluación de la tarea y los resultados obtenidos por los participantes.
Palabras clave: Traducción Automática, Microblogs, Tuits, Social Media

Abstract: This article presents an overview of the shared task that took place as part of the TweetMT workshop held at SEPLN 2015. The task consisted in translating collections of tweets from and to several languages. The article outlines the data collection and annotation process, the development and evaluation of the shared task, as well as the results achieved by the participants.
Keywords: Machine Translation, Microblogs, Tweets, Social Media

1 Introduction

While research in machine translation has been studied for a while now, the application of machine translation techniques to tweets is still in its infancy. The machine translation of tweets is a challenging task which, to a great extent, depends on the spelling and grammatical quality of the tweets that one has to provide the translations for. In fact, the difficulty of a tweet translation process varies dramatically for different types of tweets ranging from informal posts to formal announcements and news headlines posted by social media editors or community managers. The former 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 exclusively used in the platform, such as hashtags, user mentions, and retweets, among others. These characteristics make the application of machine translation tools to tweets a new problem that requires specific processing techniques to perform effectively.

The machine translation of tweets is usually tackled in two different ways: (1) as a direct translation task (tweet-to-tweet), or (2) as an indirect translation task (tweet normalization to standard text (Kaufmann and Kalita, 2010), text translation and, if needed, tweet generation). Despite the fact that the direct translation approach would look like the natural approach in an ideal scenario, the lack of parallel or comparable corpora of tweets for the working languages (Petrovic, Osborne, and Lavrenko, 2010) makes the indirect approach a more viable solution in most of the cases. Alternatively, researchers have also tried to gather similar tweets in other languages, leveraging Cross-Lingual Information Retrieval techniques (Jehl, Hieber, and Riezler, 2012).

Despite the paucity of research in the specific task of translating tweets, an increasing interest can be observed in the scientific community (Gotti, Langlais, and Farzindar, 2013; Peisenieks and Skadiņš, 2014). Similarly, a related and highly relevant direction of research is the work on

machine translation 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 machine translation techniques applied to tweets, we organized TweetMT, a workshop and shared task[1] on machine translation applied to tweets. This workshop is a follow-up to two other related workshops organized in previous years also at SEPLN: TweetNorm 2013 (Alegria et al., 2013) and TweetLID 2014 (Zubiaga et al., 2014). The workshop intended to be a forum where researchers had a chance to compare their methods, systems and results and the task focuses on MT of tweets between languages of the Iberian Peninsula (Basque, Catalan, Galician, Portuguese, and Spanish).

As a starting point, and especially given the little work performed so far in the field, the corpora we compiled for the shared task includes tweets that are mostly formal and correctly written, while keeping the brevity inherent to tweets. While the corpora might not be fully representative of the texts that one can find on Twitter, it is instead intended to boost the work performed within the field, encouraging researchers to submit preliminary contributions that will then help better understand the state of the art so that future work can be set forth. As this research matures, subsequent corpora will include a wide variety of informal and misspelled tweets to keep making progress.

2 Creation of a Benchmark Dataset

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, providing the translated text within the same tweet. 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.

In order to facilitate corpus generation, we developed a semi-automatic method to retrieve and align parallel tweets. The semi-automatic method consists in identifying multiple Twitter authors that tweet identical content, albeit in different languages, either from a single account or from two different accounts. Hence, whenever possible, the parallel corpora have been generated from multilingual Twitter accounts; this methodology was applied for the Catalan–Spanish (*ca-es*) and Basque–Spanish (*eu-es*) language pairs, as we found authors that concurrently tweet in these languages. However, we did not find authors that meet these characteristics for the other two language pairs, i.e., Portuguese–Spanish and Galician–Spanish (*pt-es* and *gl-es*); in these cases, the parallel tweets were manually produced through crowdsourcing. Different to the language pairs that could be automatically aligned, in the latter cases only test sets were generated due to time and budget constraints.

The following sections give details about the creation of the datasets. Table 1 shows some statistics of those datasets.

2.1 Corpus Creation from Multilingual Accounts

The corpus creation process out of multilingual Twitter accounts can be divided into two steps: (i) identifying the accounts and collecting the messages, and (ii) semi-automatic alignment of translated tweets.

2.1.1 Accounts and Collected Data

Different to (Ling et al., 2013), we do not aim for mixed language tweets, where the source and target segments are included in the same tweet, but rather we manually select a number of authors that tend to post messages in various languages. It is worth noting that this strategy for sampling the authors leads to a prevalence of account types that belong to organizations and famous personalities.

We identified two kinds of "authors" following this strategy: (i) authors that use a single account to post messages in different languages, and (ii) authors that have parallel accounts to post in different languages using separate accounts. 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.

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

The initial corpus was then split into two datasets: one development-set composed of 4,000 parallel tweets for each language pair and one test-set composed of 2,000 parallel tweets for each language pair.

Author distribution in the development set was limited to account 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.

As we said before, one of the limitations of our strategy is that it is only applicable to certain language pairs. The linguistic realities of Basque and Catalan (both are considered to be co-official together with Spanish in certain regions that support bilingualism) make the application of such methods viable for our purposes. Unfortunately, it was not the case for *pt-es* and *gl-es* pairs. It is understandable that few or no users have the need to tweet both in Spanish and Portuguese, which have little or no geographical overlap; it was however a surprise not to find any such example for Galician and Spanish, which has the same status as Catalan and Basque of being co-official. In consequence, we only could provide development corpora for the *eu-es* and *ca-es* language pairs. For the Galician–Spanish and Portuguese–Spanish language pairs, test sets were manually generated through crowdsourcing. Specifically, we used the CrowdFlower platform to translate tweets into the other language. Section 2.2 further discusses this process.

2.1.2 Alignment

The large volume of tweets collected in the previous step needs to be properly aligned in order to create the parallel corpus. 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). While Twitter does provide the language ID along with tweet's metadata, Basque and Catalan are never tagged as such by Twitter, so that we implemented our own language identification module to identify these languages. Language identification is done by using TextCat[2] trained over Twitter specific data.

Once we have an author's tweets separated by language, and hence with source language tweets and target language tweets separated, we need to align them with likely translations for each tweet. For the automated process, 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 flapped as possible translations of each other. The difference between source and target timestamps must not exceed a certain threshold. The value of the treshold was set overall to 10 hours, although for a few accounts the publication date difference was restricted to 1 hour after empirically detecting too much noise with the more relaxed standard threshold.
- **Overlap of hashtag and user mentions in source and target tweets.** It is very rare to change the user (@) mentions across language, only in a few cases was observed that phenomenon (e.g., using @FCBarcelona_ca in a tweet in catalan and using @FCBarcelona_es in a Spanish written tweet) are usually maintained across languages. Hashtags are translated, often, depending on the popularity of a given hashtag in the target audience. A minimum number of user name 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

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

union. The threshold is empirically set to 0.76.

- **Longest Common Subsequence ratio (LCSR) (Cormen et al., 2001) between source and target tweets.** LCSR 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 closes. We empirically set a minimum threshold of 0.45.

As for the performance of the heuristics, publication date closeness is effective for filtering out wrong candidates, but it is not enough to find the correct parallel tweet, so it is applied first of the three. User and hashtag overlapping ration proved to be very successful, up to the point that the contribution of LCSR was minimal.

The output of this alignment is then corrected through manual checks by native speakers of their respective languages. The manual inspection showed a low error rate in the automatic alignment, especially for *ca-es*. For this language pair we found a 2% error rate, evaluated over a sample of 400 tweets on the development set. For *eu-es* the percentage increased to 15%, also evaluated over a sample of 400 tweets. Error rate over the collections manually reviewed to create the test-sets was 7% for the *ca-es* language pair (12500 tweets) and was 32% for the *eu-es* language pair (15045).

2.2 Crowdsourced Corpus Creation using Crowdflower

As we did not find bilingual Portuguese–Spanish or Galician–Spanish Twitter accounts, we used the CrowdFlower platform[3] to build the test data for this language pair. CrowdFlower provides a cheap and fast method for collecting annotations from a broad base of paid non-expert contributors over the Web. It works in a similar way to Amazon's Mechanical Turk (Snow et al., 2008) which cannot be used in our case because it requires to have an US address and credit card.

In the task we defined, the 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 in working tasks of 10 tweets each.

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 Portuguese/Galician hashtag.

The crowdsourcing platform allows to configure the jobs using a number of options. We used some of them with the aim of obtaining translations of a reasonable quality:

- **Geography.** One can select a set of countries from which workers are allowed to work on the job. We limited the countries to Spain for Galician and to Portugal and Brazil for Portuguese.[4]
- **Performance level.** Contributors of the platform fall into three levels, according to their performance. Our jobs were limited to contributors in level 3 (the top level), defined by Crowdflower as "the highest performance contributors who account for 7% of monthly judgments and maintain the highest level of accuracy across an even larger spectrum of CrowdFlower jobs [compared to contributors in levels 1 and 2]". In the case of Galician, we had to change this setting to level 1 as the tasks were getting completed too slowly.
- **Language capability.** It allows to restrict the contributors that can work in the job by their language skills. For translations into Portuguese, we restricted the contributors to those who are verified speakers of Portuguese. Galician is not in the list of languages provided in CrowdFlower, so this job was not configured in this case.

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

[4] Initially, the task to translate into Portuguese was only opened for users from Portugal as the focus is on Iberian Portuguese, but after we realized we were having no contributions, we broadened the geographical scope to Brazil as well, which helped to obtain contributions more swiftly.

- **Speed trap.** If set, contributors are automatically removed from the job if they take less than a specified amount of time to complete a task. Our jobs contained tasks of 10 translations each and the time trap was set to 150 seconds. Hence if a worker toke less than 15 seconds to translate per tweet he/she would be automatically removed from the job.

The task of translating into Portugese was completed by 40 different contributors, all of them from Brazil. The contributors were inquired about the quality of the task; they were asked to rank out of 5 the clarity of the instructions (average 4.12), ease of the job (3.78), pay (4.17) and overall satisfaction (4.05). The task to translate into Galician was carried out by 10 contributors. They ranked the task as follows: clarity of the instructions (4.90), ease of the job (3.59), pay (3.82) and overall satisfaction (4.00).

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. 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 the new spelling rules imposed since 2003. In the case of Portuguese, six errors (most of them lexical problems) were found from the 30 tweets evaluated.

2.3 Datasets post-processing

Before delivering the data sets to the participants the test was pre-processed. The development corpus includes the original tweets, neither @user nor URLs were normalized, but they are in the test corpus where @user and URLs are standardized to `IDIDID` and `URLURLURL`, respectively.

Datasets are distributed in tab separated format (tsv) files. For each language pair two files are provided, one for every translation direction. For the language pairs where the parallel corpus was gathered exclusively from Twitter —this includes *ca*, *es*, and *eu*— the files contains the tweetID, userID, date and the text of each tweet. For the language pairs where the corpus was obtained via crowdsourcing —*gl* and *pt*—, the file contains a segmentID and the text of the tweet.

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.

3 Evaluation Framework

The test sets just described were delivered to the participants which had to return the translations with the following tab separated format:

```
tweet_Id <tab> source_language_text
<tab> translation \n
```

The translated test would then be extracted, cut to a maximum length of 140 characters, and evaluated by automatic means.

The performance of the systems is assessed with lexical and syntactic automatic evaluation measures compared against a single reference. Lexical metrics which are mostly based on n-gram matching are available for all the language pairs under study. However, syntactic metrics are only available for Spanish and some of them for Catalan.

3.1 Evaluation Metrics

In order to study the quality of the translations at different levels we use a wide set of metrics as defined as follows:

Lexical evaluation measures

- PER (Tillmann et al., 1997), TER (Snover et al., 2006), WER (Nießen et al., 2000): Subset of metrics based on edit distances
- BLEU (Papineni et al., 2002), NIST (Doddington, 2002), ROUGE (RG): Based on n-gram matching (lexical precision: BLEU, NIST; and lexical recall: ROUGE). For ROUGE we use RGS*, i.e. a variant with skip bigrams without max-gap-length)

- GTM (Melamed, Green, and Turian, 2003), METEOR (Banerjee and Lavie, 2005) (MTR): Based on the F-measure. For GTM we use GTM2, with the parameter associated to long matches $e = 2$; for METEOR we use MTRex, i.e. using only exact matching.
- Ol (Giménez and Màrquez, 2008): Lexical Overlap is a measure based on the Jaccard coefficient (Jaccard, 1912) to quantify the similarity between sets. Lexical items associated with candidate and reference translations are considered as two separate sets of items. Overlap is computed as the cardinality of their intersection divided by the cardinality of their union.
- ULC (Giménez and Màrquez, 2008): *U*niform *L*inear *C*ombination. When applied to lexical metrics it includes WER, PER, TER, BLEU, NIST, RGS*, GTM2, MTRex.

Syntactic evaluation measures

- SP-O_{p}, SP-O_{c}, SP-pNIST (Giménez and Màrquez, 2007)[5]: Based on the lexical overlap according to the *p*art-of-speech or *c*hunk and the NIST score over these elements (Shallow Parsing)
- CP-O_{p}, CP-O_{c}, CP-STM9 (Giménez and Màrquez, 2007)[6]: Based on the lexical overlap among *p*art-of-speech or *c*onstituents of constituency parse trees (Constituency Parsing)
- ULC (Giménez and Màrquez, 2008): *U*niform *L*inear *C*ombination. When applied to syntactic metrics it includes the available metrics for the specific language.

All measures have been calculated with the Asiya toolkit[7] for MT evaluation (Giménez and Màrquez, 2010).

4 Shared Task Results

Participants were required to register[8] in order to obtain the development and test data-sets. Each participant had only 72 hours to work on the test set and to send the results.

4.1 Overview of the Systems Submitted

Out of the 5 initially registered participants, only 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., 2015); and UPC (Universitat Politècnica de Catalunya) for the *ca-es* track (Martínez-Garcia, España-Bonet, and Màrquez, 2015). In all, 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 related shared tasks that we organized in recent years (i.e., TweetNorm (Alegria et al., 2013) and TweetLID (Zubiaga et al., 2014)) attracted a higher number of participants. One of the reasons for this drop in number of participants might be the fact that English has not been considered this time as one of the languages included in the task; this could have made the task less appealing to some groups, which led to fewer participants from outside the Iberian Peninsula.

The main characteristics of the systems submitted are compiled in Table 2 and can be summarized as follows:

DCU This team submitted systems for three language pairs in both directions: Spanish from/to Catalan, Basque and Portuguese. They used a range of techniques including state-of-the-art SMT, morph segmentation (only for Basque as a morphologically rich language), data selection as a means of domain adaptation, available open-source rule-based systems and, finally, system combination to combine the strengths of the different systems that were built. DCU gathered vast amounts of tweets (from 11M for Basque to 130M for Spanish) to perform monolingual domain adaptation and complemented this with publicly available general-domain monolingual and parallel corpora. The first (DCU1), second (DCU2) and third (DCU3) systems submitted for each language

[5]Family of metrics only available for Catalan and Spanish.

[6]Family of metrics only available for Spanish.

[7]`http://nlp.cs.upc.edu/asiya/`

[8]`http://komunitatea.elhuyar.eus/tweetmt/participation/`

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.

direction were the individual systems or combinations that obtained the best, second best and third best result, respectively, on the development set.

EHU This team submitted systems for the Basque–Spanish pair. They have adapted previous MT engines for the *es–eu* and *eu-es* directions. For the translation into Basque RBMT and SMT were adapted whereas for the translation from Basque only a SMT based system was used. The main work was pre- and post-processing for adaptation to tweets and collecting new resources for training and tuning the systems. For RBMT, a small dictionary of hashtags was obtained from the development-set. For SMT, language models were improved using monolingual corpora from previous shared tasks and a new corpora of tweets in Basque.

UPC The team submitted two systems for the Catalan–Spanish language pair. The first one (UPC1) is a standard SMT system built with Moses (Koehn et al., 2007) and trained with 2,178,796 parallel sentences extracted from the *El Periódico* parallel corpus[9]. The second system (UPC2) uses a document-level decoder, Docent (Hardmeier et al., 2013), that takes UPC1 as a first step. Besides, the system uses as additional feature semantic models obtained with word2vec (Mikolov et al., 2013). Besides the parallel tweets available for the shared task, both systems use monolingual tweets for genre and domain adaptation. UPC2 was only submitted for Catalan-to-Spanish. The authors report some problems with this configuration and include both the official and new results in their paper. Here only the official results are shown.

[9] http://catalog.elra.info/product_info.php?products_id=1122

4.2 Results

Participants had a 72-hour window to work with the test set and submit up to three results per track. This section is a recap of the results of all the tracks and systems.

Table 3 and Table 4 show the results for the participants in the *ca-es* track. In Table 3, the lexical measures introduced in the previous section are shown and in Table 4 the syntactic ones. Five systems from two teams have been evaluated. DCU3 system was the best for the *ca-es* direction, a system combining two kinds of 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.

Tables 5, 6, 7 and 8 show the results for the participants in the *eu-es* and *pt-es* tracks. For the *eu-es* track four or five (depending on the direction) systems were presented by two teams. In general, the best translator for this language pair is the statistical system EHU1

Catalan to Spanish

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	15.24	12.49	13.25	76.73	12.09	72.75	83.8	83.37	83.7	77.84
DCU2	15.15	12.41	13.21	76.52	12.09	72.18	83.76	83.70	83.56	77.86
DCU3	**14.59**	**11.74**	**12.50**	**77.70**	**12.16**	**73.37**	**84.63**	**84.45**	**84.64**	**79.67**
UPC1	20.17	16.40	19.42	68.20	11.22	62.71	78.46	77.31	74.72	63.82
UPC2	25.10	17.09	22.25	63.12	10.93	57.92	76.44	75.56	73.76	57.45

Spanish to Catalan

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	16.70	12.42	14.46	75.79	11.88	70.45	52.08	82.65	82.88	66.23
DCU2	15.17	11.71	**13.21**	77.75	11.96	72.65	53.44	**83.32**	**83.96**	69.98
DCU3	17.09	13.08	14.70	75.25	11.85	70.34	51.73	82.26	82.46	64.94
UPC1	**14.35**	**11.25**	13.63	**77.93**	**12.04**	**72.69**	**53.98**	82.19	83.18	**70.56**

Table 3: Evaluation with a set of lexical metrics (see Section 3.1 for a description) for the participant systems on the Catalan–Spanish language pair. Results are obtained only considering the first 140 characters per tweet.

Catalan to Spanish

System	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	80.92	81.4	74.1	81.78	83.03	10.87	99.24
DCU2	80.71	81.5	74.19	81.66	82.8	10.90	99.22
DCU3	**81.64**	**82.27**	**74.48**	**82.40**	**83.73**	**10.92**	**100.00**
UPC1	68.52	70.93	58.74	70.95	71.97	9.36	84.47
UPC2	70.59	72.89	63.05	73.01	73.99	10.01	88.06

Spanish to Catalan

System	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	–	–	–	80.77	82.10	10.78	98.41
DCU2	–	–	–	**82.13**	**83.14**	10.88	**99.67**
DCU3	–	–	–	80.19	81.42	10.75	97.81
UPC1	–	–	–	81.59	82.02	**10.99**	99.33

Table 4: Evaluation with a set of syntactic metrics (see Section 3.1 for a description) for the Catalan–Spanish language pair. Results are obtained with the restriction of considering only the first 140 characters per tweet. Not all the syntactic metrics are available for Catalan.

in both directions. When translating from Spanish into Basque, however, DCU2 with the combination of 5 different systems gets very similar scores. Differences in this case are in general not statistically significant.

Finally, in the *pt-es* track DCU submitted the results of three systems. 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 their best system in development does not correspond to the best system in test.

Based on the previous figures as well as on the conclusions drawn by the authors of the papers submitted to the shared task (Toral et al., 2015; Alegria et al., 2015; Martínez-Garcia, España-Bonet, and Màrquez, 2015), we can emphasize the following conclusions:

- The results are in general very good when compared to previous results for

Basque to Spanish

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	62.19	44.72	56.37	25.30	6.46	32.70	45.71	34.20	44.48	59.78
DCU2	61.24	44.95	55.35	25.30	6.53	33.14	46.12	34.61	44.92	60.63
DCU3	**61.04**	44.78	54.99	25.44	6.56	33.34	46.32	35.31	45.50	61.31
EHU1	61.53	**38.17**	**52.96**	**28.61**	**6.94**	**34.53**	**50.57**	**40.80**	**51.12**	**69.13**

Spanish to Basque

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	61.48	47.56	55.81	23.22	5.96	32.45	40.00	29.92	42.87	66.27
DCU2	**61.06**	46.27	**55.17**	**24.44**	6.12	**33.17**	41.18	31.95	44.29	69.18
DCU3	61.77	47.30	56.07	23.42	5.96	32.48	40.12	30.38	43.00	66.56
EHU1	62.00	**45.04**	56.06	24.34	**6.14**	**33.17**	**41.98**	**32.22**	**45.07**	**69.63**
EHU2	66.43	50.13	62.46	19.54	5.29	29.29	36.36	23.30	38.15	55.33

Table 5: Evaluation with a set of lexical metrics for the Basque–Spanish language pair.

Basque to Spanish

System	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	36.82	38.54	29.67	40.94	43.43	5.24	87.99
DCU2	37.13	38.84	29.77	41.16	43.67	5.23	88.4
DCU3	37.71	39.32	30.11	41.69	44.20	5.27	89.45
EHU1	**43.26**	**45.19**	**33.59**	**47.42**	**49.8**	**5.48**	**100.00**

Table 6: Evaluation with a set of syntactic metrics for the Basque–Spanish language pair. These metrics are not available for Basque.

Portuguese to Spanish

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	40.51	33.22	37.39	43.36	8.70	42.69	58.85	52.59	58.77	65.48
DCU2	39.86	33.41	36.87	43.67	8.74	43.28	59.12	52.86	58.96	66.17
DCU3	**39.08**	**33.09**	**36.11**	**44.28**	**8.83**	**43.90**	**59.89**	**53.61**	**59.54**	**67.54**

Spanish to Portuguese

System	WER	PER	TER	BLEU	NIST	GTM2	MTRex	RGS*	Ol	ULC
DCU1	47.68	39.40	44.45	36.13	7.57	37.71	53.78	44.10	52.38	65.27
DCU2	**46.51**	36.67	**43.08**	**37.25**	**7.77**	**38.30**	**54.15**	**45.24**	**53.57**	**68.05**
DCU3	47.04	**36.51**	43.39	36.94	7.76	38.14	53.71	45.19	53.39	67.61

Table 7: Evaluation with a set of lexical metrics for the Portuguese–Spanish language pair.

Portuguese to Spanish

System	CP-Oc(*)	CP-Op(*)	CP-STM9	SP-Op(*)	SP-Oc(*)	SP-pNIST	ULC
DCU1	53.57	55.48	44.99	57.32	59.06	8.15	98.51
DCU2	53.85	55.66	45.30	57.48	59.24	8.17	98.89
DCU3	**54.53**	**56.28**	**45.96**	**58.06**	**59.92**	**8.23**	**100.00**

Table 8: Evaluation with a set of syntactic metrics for the Portuguese–Spanish language pair. These metrics are not available for Portuguese.

the same language pairs (Alegria et al., 2015).

- Combining techniques, including RBMT and SMT, can lead to improvements (Toral et al., 2015).
- Expanding the context by using a user's tweets within the same day can be of use to boost the performance of the machine translation system (Martínez-Garcia, España-Bonet, and Màrquez, 2015).

5 Conclusion

The shared task organized at TweetMT 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*. This has allowed participants to tune and compare their MT systems. The corpus developed for the shared task can in turn be downloaded from the workshop's website[10], which we expect that will enable further research in the field.

The participants of the shared task have applied and studied the suitability of state-of-the-art MT techniques. These techniques have been adapted to the specific features of tweets, including conventions such as hashtags, and user mentions, as well as considering the brevity of the texts. The study of the results achieved by the submitted systems enables us to draw conclusions to better inform future research in the field.

The results achieved by the participants of the shared task are surprisingly high, especially considering that we are dealing with tweets, whose brevity and specific characteristics make them more challenging to translate. Still, it is worthwhile noting that the tweets considered in this shared task can largely be deemed formal. Therefore, we could say that the task (translating formal tweets generated from multilingual Tweet IDs) was easier than usual tasks in MT.

A more thorough analysis of the task and performance of the participating systems will follow in an extended version of this paper, including the conclusions after the discussion in the workshop.

We should also emphasize that these results cannot be generalized to broader tasks of translating tweets. 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 machine translation task by including all kinds of tweets, to assess the ability of MT systems to translate informal tweets too. A second version of the TweetMT dataset would include:

- Tweets in English, so that we can attract a larger number of participants, comparing a larger number of MT systems.
- A more generalistic Twitter dataset including informal tweets as well, in order to test the result of MT to a large and diverse corpus like Twitter.

One of the main remaining challenges is the need to come up with a methodology to put together a gold standard corpus that encompasses the different types of tweets that one can find on Twitter, including more informal tweets than those we have considered here. To tackle such a process, we would first need to solve some open questions such as whether or not and how to translate words that are not written in its normalized form, as well as how to deal with multilingualism in a single tweet. We are confident that the discussion among attendees of the workshop, the presentations of accepted papers, as well as the invited talk (Gonzàlez, 2015) will help pave the way in this crucial task.

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

Acknowledgements

This work has been supported by the following projects: Abu-Matran (FP7-PEOPLE-2012-IAPP) PHEME (FP7, grant No. 611233), *Tacardi* (Spanish MICINN TIN2012-38523-C02-01), QTLeap (FP7, grant No. 610516), HPCPLN (Galician Gov, EM13/041), Celtic (Innterconecta program, 2012-CE138).

References

Alegria, Iñaki, Nora Aranberri, Víctor Fresno, Pablo Gamallo, Lluis Padró, Iñaki San Vicente, Jordi Turmo, and Arkaitz Zubiaga. 2013. Introducción a la tarea compartida tweet-norm 2013:

Normalización léxica de tuits en español. In *Tweet-Norm@SEPLN*, pages 1–9.

Alegria, Iñaki, Mikel Artetxe, Gorka Labaka, and Kepa Sarasola. 2015. EHU at TweetMT: Adapting MT engines for formal tweets. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Cormen, Thomas H., Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. 2001. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition.

Doddington, George. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology (HLT)*, pages 138–145, San Diego, CA, USA.

Giménez, Jesús and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic. Association for Computational Linguistics.

Giménez, Jesús and Lluís Màrquez. 2008. A Smorgasbord of Features for Automatic MT Evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio, June". The Association for Computational Linguistics.

Giménez, Jesús and Lluís Màrquez. 2010. Asiya: an Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94:77–86.

Gonzàlez, Meritxell. 2015. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Gotti, Fabrizio, Philippe Langlais, and Atefeh Farzindar. 2013. Translating government agencies tweet feeds: Specificities, problems and (a few) solutions. *NAACL 2013*, page 80.

Hardmeier, C., S. Stymne, J. Tiedemann, and J. Nivre. 2013. Docent: A document-level decoder for phrase-based statistical machine translation. In *Proceedings of the 51st Annual Conference of the Association for Computational Linguistics*, pages 193–198.

Jaccard, Paul. 1912. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50.

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. Association for Computational Linguistics.

Kaufmann, Max and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL07, pages 177–180.

Ling, Wang, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

Martínez-Garcia, Eva, Cristina España-Bonet, and Lluís Màrquez. 2015. The UPC TweetMT participation:

Translating formal tweets using context information. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Melamed, I. Dan, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 61–63, Edmonton, Canada.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*. http://code.google.com/p/word2vec.

Munro, Robert. 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.

Nießen, Sonja, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 39–45, Athens, Greece.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 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, Jānis and Raivis Skadiņš. 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.

Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. 2010. The edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, USA.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263.

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP Based Search for Statistical Translation. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece.

Toral, Antonio, Xiaofeng Wu, Tommi Pirinen, Zhengwei Qiu, Ergun Bicici, and Jinhua Du. 2015. Dublin city university at the tweetmt 2015 shared task. In *TweetMT@SEPLN, Proc. of the SEPLN 2015*.

Zubiaga, Arkaitz, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2014. Overview of tweetlid: Tweet language identification at sepln 2014. *TweetLID@SEPLN*.