TweetNorm_es Corpus:
an Annotated Corpus for Spanish Microtext Normalization

Iñaki Alegria[1], Nora Aranberri[1], Pere R. Comas[2], Víctor Fresno[3], Pablo Gamallo[4], Lluis Padró[2], Iñaki San Vicente[5], Jordi Turmo[2], Arkaitz Zubiaga[6]

(1) IXA. UPV/EHU, (2) UPC, (3) UNED, (4) USC, (5) Elhuyar, (6) City University of New York
Ixa group. Informatika Fakultatea. UPV/EHU. 649 PK. 20080 Donostia. Basque Country
tweet-norm@elhuyar.com

Abstract

In this paper we introduce TweetNorm_es, an annotated corpus of tweets in Spanish language, which we make publicly available under the terms of the CC-BY license. This corpus is intended for development and testing of microtext normalization systems. It was created for Tweet-Norm, a tweet normalization workshop and shared task, and is the result of a joint annotation effort from different research groups. In this paper we describe the methodology defined to build the corpus as well as the guidelines followed in the annotation process. We also present a brief overview of the Tweet-Norm shared task, as the first evaluation environment where the corpus was used.

Keywords: Microtext normalization, Twitter, phonology

1. Introduction

With the evergrowing usage of Twitter as a microblogging service, it has become an ubiquitous platform where users share vast amounts of information in a real-time fashion. Information posted by users in the form of *tweets* is characterized by their brevity restricted by Twitter's 140 character limit, which often lack of correct grammar and/or spelling. This makes the tweet normalization task a key initial step for subsequently running NLP tools such as machine translation and sentiment analysis on tweets. While normalization of SMS and tweets in English has attracted the interest of numerous researchers recently, little has been studied for this kind of short texts written in Spanish.

In order to promote research in this field, we organized the Tweet-Norm 2013 shared task held at the SEPLN conference in Madrid. The goal of the shared task was to create a benchmark for microtext normalization in Spanish. Participants were asked to normalize a set of tweets containing "ill-formed" word forms. We created a corpus of annotated tweets that was shared as a benchmark for the task. It is our believe that the shared task provided a common testing ground, where the most recent algorithms and techniques were used. By making the corpus publicly available, we would like to enable researchers and practitioners to make use of a common evaluation setting.

In this paper, we describe the methodology followed for the generation of the annotated corpus, and the resulting Tweet-Norm_es (TN_es) corpus.[1] We also present an overview of the results of the shared task, as part of the evaluation of the resource. The corpus can be used, modified and redistributed under the terms of the CC-BY license.

2. Related Work

A good general introduction to the problem of running NLP techniques on the "bad language" found in social media is offered by (Eisenstein, 2013). The author surveys the two most popular solutions to deal with it: normalization and domain adaptation. The work of (Han and Baldwin, 2011) is a well-known reference to the task of lexical normalization of tweets, albeit their study focused on English tweets. We relied on the methodology defined in their paper to design the corpus and the shared task. Yet, some differences must be pointed out when comparing this approach with Tweet-Norm 2013 shared task. In (Han and Baldwin, 2011), the evaluation focused on the candidate selection step, since the authors assumed perfect ill-formed word detection. This means that only ill-formed OOV words were taken as input of the candidate selection process. The task defined at Tweet-Norm 2013 also considers the detection of ill-formed OOV. Another significant difference between Tweet-Norm 2013 and the method described in (Han and Baldwin, 2011) lies in the use of multiwords in lexical normalization. Tweet-Norm 2013 did consider one-to-several correspondences (e.g., $imo \rightarrow in\_my\_opinion$), while they were disregarded by (Han and Baldwin, 2011). So, multiwords had to be taken into account by the participants at the workshop.

There is also research that has incorporated a tweet normalization step before performing the subsequent task. For instance, (Wei et al., 2011) perform a 4-step normalization of English tweets before running their topic detection system: (i) OOV word detection, (ii) slang word translation, (iii) candidate set generation, and (iv) candidate selection.Instead, we want to tackle tweet normalization starting from the reduced availability of training data, focusing on the algorithms and external resources that can be of help.

Others have relied on large-scale training data to perform the normalization afterward. For instance, (Beaufort et al., 2010) deal with SMS normalization making use of the noisy channel model, very common in speech processing. In (Kaufmann and Kalita, 2010), the authors make use of the Statistical Machine Translation framework.

To the best of our knowledge, no work has focused before

[1] Further information can be found at `http://komunitatea.elhuyar.org/tweet-norm/`

on tweets written in Spanish. As a related effort for tweets in Spanish, (Villena Román et al., 2013) organized a shared task focused on sentiment analysis.

3. Corpus

3.1. Tweet Dataset

Among Twitter API's[2] choices to track and collect tweets, we opted for the geolocated tweets, whose metadata include the coordinates of the location each tweet was sent from. Twitter's API allows in turn to filter the tweets sent from the geographic area specified as a parameter. Making use of this feature, we chose an area within the Iberian Peninsula, taking out regions where languages other than Spanish are also spoken. We found this approach to be highly effective when it comes to gathering large numbers of tweets in Spanish. Thus, the selected geographic area forms a rectangle with Guadalajara (coordinates: 41, -2) as the northeasternmost point and Cádiz (coordinates: 36.5, -6) as the southwesternmost point. The collection of tweets we gathered on April 1-2, 2013 amounts to 227,855 tweets. From this large dataset, we created two random subsets of 600 tweets each, which were shared with participants, one as a training set, and the other as a test set for final evaluation purposes. The rest of the dataset was also shared with participants, with no manual annotations, which they could use to feed their normalization system.

3.2. Preprocessing

We used FreeLing's morphological analyzer (Padró and Stanilovsky, 2012) to identify out-of-vocabulary words (OOV) in tweets. We used the basic modules (dictionary, affix analysis, number and dates detection, etc.) to analyze tokens in tweets, and when a token had no match in any of the modules, it was ultimately considered as an OOV.

The first step of the preprocessing consisted in tokenizing tweets. The tokenizer's rules were tuned to keep usernames (`@user`), hashtags (`#hashtag`), e-mail addresses, URLs, and the most common smileys as a single token. We also utilized the *usermap* module to apply a set of regular expressions to each token, and to assign an analysis to those matching any of them. In this way, these tokens were discarded from being considered as OOVs, since they did obtain an analysis.

In the second step of the preprocessing, we applied a basic morphological analyzer, with the default modules, except the multiword detector (to avoid agglutination of several tokens into a single one), the named entity detector (since we want to keep them as OOVs), and the lexical probabilities module (which includes a guesser that would assign at least one analysis to every word).

After this preprocessing, words that did not receive an analysis from any module were considered OOVs.

4. Annotation Methodology

During the annotation process, experts were set to annotate the OOV words. They tagged each OOV word either as *correct*, *variant* or *NoES* (not in Spanish). For those cases deemed variant, they also provided the normalized spelling of the word along with the annotation.

Three experts independently annotated each OOV word for the development set, and two of them participated in the annotation of the test corpus. We put together the annotations from the different experts by majority voting when possible, and by further discussing the correct annotation among the experts in case of ties. So, no metric to measure inter-annotator agreement was needed, since all decisions were taken by all annotators. To facilitate the annotation process and subsequent discussions, we defined the following guidelines for each OOV word, which include the most controversial cases:

- When the word is included in RAE's[3] dictionary: correct.
- When a well-formed word refers to a Named Entity (e.g., Zaragoza) or a loanword (e.g. Twitter): correct.
- When a word incorporates an emphatic or dialectal variation, it is misspelled, or lacks or misuses the acute accent: mark as variation and provide the standard spelling (e.g., muuuuuuucho/mucho, kasa/casa, cafe/café)
- When more than a word are written together with no separation: mark as variation and provide the standard spelling.
- When a single word is split into smaller strings: mark all of them as variation and provide the standard spelling.
- When a word is unintelligible, a foreign word, or others (e.g., XD): NoES.

These guidelines include the most common cases, but some of the cases we found were still not covered. In these cases, we met to further discuss each case in search of the most suitable solution.

Examples of uncommon cases that are not considered by the guidelines above include:

- `que estafa de tablet` [what a scam is this tablet] (in this case *tablet* is a loanword that is not included in the RAE dictionary yet, but the Spanish alternative *tableta* will incorporate this new meaning in the next release).

One of the most challenging cases we identified during the annotation process was the normalization of abbreviations. The context surrounding the abbreviated word in question is not sufficient in some cases so as to disambiguate its meaning and identify the intention of the user. For instance:

- `cariiii k no te seguia en twitter!!!mu fuerte!!!..yasoy tu fan....muak....se te exa d menos en el` **`bk`**`....sobreto en los cierres jajajajas` [sweetieeee i wasn't following you on twitter!!no way!!i'm your fan from now on....kisses... we miss you in the **bk**.... especially when closing hahaha]

[2] https://dev.twitter.com/docs/api

[3] RAE, or *Real Academia Española*, is the institution responsible for regulating the Spanish language.

where it is difficult to know what *bk* refers to with certainty. This user talking to whom seemingly was his/her colleague at a place called *bk* provides little evidence that, without further research, makes hard to grasp its exact meaning. The annotators ultimately chose *Burger King* as the variant, as the most likely choice for the acronym. In a few cases that could not be disambiguated, the annotators provided two alternatives. This includes cases where the gender could not be disambiguated from the abbreviated form –e.g., a tweet from the corpus contained with the word *her*, which may have referred to either *hermano* (brother) or *hermana* (sister).

The specific use of some onomatopoeias was also hard to grasp in some cases, which needed further discussion to come to an agreement among annotators. For instance:

- `me da igual JUUUM!!` [i don't care huuum!!]

In this case, as *JUUUM* cannot be normalized using any standard Spanish onomatopeia, it is not considered for evaluation.

Finally, non-textual emoticons, such as :), ;), :-(, etc, are not considered since their normalization is not trivial and relies on the specific criteria defined by further applications, namely, sentiment analysis.

5. Development and test corpora

Two collections have been generated from the initial corpus described in Section 3.1.: the development corpus and the test corpus, which consist of 600 tweets each. A total of 775 and 724 OOV words were manually annotated respectively in both corpora.

As required by Twitter API's terms of use,[4] we do not release the content of the tweets, but provide instead the user names and tweet IDs that enable to download the content of the tweets by using *Twitid*. *Twitid*[5] is a script that retrieves the content of tweets from the list of user names and tweet IDs.

Since we distributed the lists of tweets to participants by following the method above, chances are that some tweets might become unavailable from the date we collected them, to the date when participants did it. Some tweets may become unavailable with the time as some users remove their accounts or make them private. This may lead to different participants having slightly different collections of tweets, which would affect the evaluation process. We solved this by identifying the subset of tweets that were still available after all participants submitted their results. We found that 562 of the 600 tweets in the original test set were still accessible at the time. Thus, the initial set of 724 OOV words found in the initial test corpus were reduced to 662 due to the unavailable tweets. We relied on this slightly reduced set of tweets for the final evaluation.

Table 1 shows the distribution of the three OOV word categories (0, correct; 1, variant; 2, NoES) in both the development corpus and the test corpus. Note that the distribution of the three categories is similar in both corpora. This fact allowed the participants to develop their systems with a corpus that is similar to the test corpus.

Corpus	**#OOV**	**0**	**1**	**2**
Development	775	600	107	68
Test	662	531	98	33

Table 1: Statistics of the corpora

6. Tweet-Norm shared task

The Tweet-Norm shared task consisted in normalizing words unknown to the analyzer at the preprocessing step, such as abbreviations, incorrectly spelled words, words with repeated characters, etc. Following the line of work by (Han and Baldwin, 2011) we focus on lexical normalization, other phenomena such as syntactical or stylistic variants are left out of this task.

The goal of the task is to measure how useful a system is to normalize OOV words found in tweets. This goal does not involve the classification of the OOV words into different categories (0, 1 and 2, as described in previous section). Therefore, the task rather focuses on identifying whether an OOV word needs to be corrected, and providing the correct alternative when necessary. Participants had to determine whether to deem an OOV word correct (e.g., new named entities, words in other language, etc.) or to assign a normalized variation. We defined the following criteria that we considered when performing the final evaluation measuring the accuracy of each system:

- **Correct**: if the OOV word is correct (category 0) or NoES (category 2) and the system does not provide any correction, or if the OOV word is a variant (category 1) and the word suggested by the system to normalize the OOV word is correct.
- **Incorrect**: otherwise.

In order to measure the performance of the systems, we relied on the precision score, defined as the number of correct responses of a system over the whole set of OOV words in the test corpus:

$$P(system_i) = \frac{\text{\#correct suggestions}}{\text{\#OOV words}}. \qquad (1)$$

Recall is not taken into account in the evaluation because all participants in the shared task must use the same OOV identifier, provided by FreeLing.

6.1. Results

Table 2 shows the accuracy results obtained by the 13 participants.[6] The table includes an extra column with a second precision value for participants who submitted two runs. Besides the results of the participants, we also show two more results as references. On one hand, the *Baseline* would be the result of deeming all OOV words correct, therefore without suggesting any changes at all from

[4] `https://dev.twitter.com/terms/api-terms`

[5] `http://komunitatea.elhuyar.org/tweet-norm/iles/2013/06/download_tweets.py`

[6] Out of 20 initially registered participants, 13 groups sent results.

Rank	System	Prec #1	Prec #2
—	*Upperline*	0.927	—
1	RAE	0.781	—
2	Citius-Imaxin	0.663	0.662
3	UPC	0.653	—
4	Elhuyar	0.636	0.634
5	IXA-EHU	0.619	0.609
6	Vicomtech	0.606	—
7	UniArizona	0.604	—
8	UPF-Havas	0.548	0.491
9	DLSIAlicante	0.545	0.521
10	UniMelbourne	0.539	0.517
11	UniSevilla	0.396	—
12	UJaen-Sinai	0.376	—
13	UniCoruña	0.335	—
—	*Baseline*	0.198	—

Table 2: Precision of the Tweet-Norm 2013 evaluation participants

the input –this would achieve a precision of 0.198. On the other hand, the *Upperline* is the aggregated precision value of words that were correctly guessed by one of the participants. With a precision of 0.927, 7.3% of the OOV words were missed by all of the participants.

A more detailed description of the systems can be found on the papers of the workshop.[7]

6.2. Discussion of the results

The system presented by RAE clearly outperformed the rest of the systems, with 11.8% gain over the runner-up, Citius-Imaxin. Most of the other systems achieved intermediate precision values that range from 54% to 67%. We believe that one of the features that stand out from the winners' systems is the careful integration of different components that consider a number of misspelling cases, as well as the quality and coverage of the components utilized.

7. Conclusions and future work

The TweetNorm_es corpus, and the Tweet-Norm 2013 shared task that enabled evaluation of systems from 13 participants, served as an initial step toward encouraging implementation of new methods for and approaches to Spanish microtext normalization in the research community. The high number of participants has proven the task important, and posited a number of issues to be considered in future research.

The work presented in this paper paves the way for future development and research on Spanish microtext normalization, setting forth a methodology to create a corpus for these purposes, as well as releasing the corpus we created following such methodology. The corpus provides a gold-standard for development and evaluation of microtext normalization tools. The corpus is available under the terms of the CC-BY license for both researchers and practitioners interested in the task, and can be found at the website of the workshop.[8]

This work has also brought to light a number of issues that remain unresolved and are worth studying in future work. Here we have performed *in vitro* evaluations of the normalization systems. We believe that *in vivo* evaluations by incorporating normalization into other NLP systems, such as sentiment analysis or machine translation will enable deeper study of the task, as well as to quantify the actual effect of processing normalized outputs. Additionally, we would like to broaden the task by not only dealing with lexical normalization, but also addressing complementary tasks such as normalization of syntax and/or real-word errors. Last but not least, we are aware that the size of the corpus is limited. Extending the corpus and considering the different OOV categories would enable to perform more detailed evaluation, especially for machine learning purposes.

Acknowledgments

We would like to thank all the members of the organizing committee. This work has been supported by the following projects: Spanish MICINN projects *Tacardi* (Grant No. TIN2012-38523-C02-01), *Skater* (Grant No. TIN2012-38584-C06-01), *TextMESS2* (TIN2009-13391-C04-01), *OntoPedia* (Grant No. FFI2010-14986) and *Holopedia* (TIN2010-21128-C02-01); *Xlike* FP7 project (Grant No. FP7-ICT-2011.4.2-288342); UNED project (2012V/PUNED/0004); *ENEUS-Marie Curie Actions* (FP7/2012-2014 under REA grant agreement n^{o}302038); *Celtic* CDTI FEDER-INNTERCONECTA project (Grant No. ITC-20113031); Research Network MA2VICMR (S-2009 / TIC-1542); and *HPCPLN* (Grant No. EM13/041, Xunta de Galicia).

8. References

Beaufort, Richard, Roekhaut, Sophie, Cougnon, Louise-Amélie, and Fairon, Cédrick. (2010). A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *48th Annual Meeting of the Association for Computational Linguistics*, pages 770–779, Uppsala, Sweden.

Eisenstein, Jacob. (2013). What to do about bad language on the internet. In *Proceedings of the NAACL-HLT*, pages 359–369.

Han, Bo and Baldwin, Timothy. (2011). Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the ACL*, pages 368–378.

Kaufmann, J. and Kalita, J. (2010). Syntactic normalization of twitter messages. In *Conference on Natural Language Processing*, Kharagpur, India.

Padró, Lluís and Stanilovsky, Evgeny. (2012). Freeling 3.0: Towards wider multilinguality. In *Proceedings of the LREC.*

Villena Román, Julio, Lana Serrano, Sara, Martínez Cámara, Eugenio, and González Cristóbal,

[7] SEPLN2013 conference: workshops `http://www.congresocedi.es/images/site/actas/ActasSEPLN.pdf`

[8] `http://komunitatea.elhuyar.org/tweet-norm/`

José Carlos. (2013). TASS-workshop on sentiment analysis at SEPLN. *Proceedings of the SEPLN*.

Wei, Zhongyu, Zhou, Lanjun, Li, Binyang, Wong, Kam-Fai, Gao, Wei, Wong, Kam-Fai, and of Education, Ministry. (2011). Exploring tweets normalization and query time sensitivity for twitter search. In *TREC*.