# unimelb: Spanish Text Normalisation

## *unimelb: Normalización de texto en español*

**Bo Han,**[1,2] **Paul Cook**[1] **and Timothy Baldwin**[1,2]
[1] Department of Computing and Information Systems, The University of Melbourne
[2] NICTA Victoria Research Lab
hanb@student.unimelb.edu.au, paulcook@unimelb.edu.au, tb@ldwin.net

**Resumen:** El presente artículo describe una aproximación a la normalización de texto basada en léxico para tweets en español. En primer lugar se realiza una comparación entre la normalización de texto en español e inglés y se plantea la hipótesis de que se puede adaptar un enfoque similar ya planteado previamente para el inglés. Para ello, se construye un léxico de normalización a partir de un corpus, utilizando similaridad distribucional, y se combina con otros léxicos existentes (por ejemplo diccionarioes de jerga de Internet en español). Estos léxicos permiten una solución rápida basada en búsquedas. Los resultados experimentales indican que el léxico derivado del corpus complementa bien a los léxicos existentes, pero que la solución puede mejorarse con un mejor manejo de ciertos tipos de palabras, como las entidades con nombre.
**Palabras clave:** Twitter, español, normalización de texto

**Abstract:** This paper describes a lexicon-based text normalisation approach for Spanish tweets. We first compare English and Spanish text normalisation, and hypothesise that an approach previously proposed for English can be adapted to Spanish. A corpus-derived normalisation lexicon is built using distributional similarity, and is combined with existing lexicons (e.g., containing Spanish Internet slang). These lexicons enable a very fast, look-up based approach to text normalisation. Experimental results indicate that the corpus-derived lexicon complements existing lexicons, but that the approach could be improved through better handling of certain word types, such as named entities.
**Keywords:** Twitter, Spanish, Text Normalisation

## *1 Introduction*

A tremendous amount of user-generated text is produced on social media sites such as Twitter and Facebook, and can be leveraged for natural language processing (NLP) tasks such as sentiment analysis (Jiang et al., 2011) and event detection (Weng and Lee, 2011). However, this user-generated text is noisy, and contains various non-standard words, e.g., *jajaja* ("ja") and *queee* ("que"). These non-standard words are not recognised by off-the-shelf NLP tools, and may consequently degrade the utility of NLP on social media. One way to tackle this challenge is text normalisation — restoring these non-standard words to their canonical forms, e.g., transforming *jajaja* to "ja" and *queee* to "que" (Eisenstein, 2013; Han, Cook, and Baldwin, 2013).

This paper proposes a lexicon-based approach to Spanish text normalisation. In particular, we adapt the method of Han, Cook, and Baldwin (2012) to build a normalisation lexicon that maps non-standard words to their standard forms relative to a vocabulary, i.e., out-of-vocabulary (OOV) words are mapped deterministically to in-vocabulary (IV) words. This enables a very fast, look-up based approach to text normalisation. In our approach an OOV word is first looked up in an automatically-derived normalisation lexicon that is complemented with entries from Spanish Internet slang dictionaries and the development data. If the OOV word is found in this lexicon it is normalised according to its entry, otherwise it is left unchanged. During this normalisation step, OOV words and the resulting normalisations are down-cased, so a final case restoration step is performed to appropriately capitalise the lowercased normalisations.

Han, Bo, Paul Cook and Timothy Baldwin (2013) unimelb: Spanish Text Normalisation, In Proceedings of the Tweet Normalization Workshop at SEPLN 2013 (Tweet-norm), Madrid, Spain, pp. 67–71.

## 2 Comparing English and Spanish Text Normalisation

The lexicon-based normalisation approach of Han, Cook, and Baldwin (2012) was evaluated on English tweets. In this section we consider the plausibility of adapting their method from English to Spanish, and identify the following key factors:

**Orthography:** if we consider diacriticised letters as single characters, Spanish has more characters than English, and diacritics can lead to differences in meaning, e.g., *más* means "more", and *mas* means "but". The method of Han, Cook, and Baldwin (2012) uses Levenshtein distance to measure string similarity. We simply convert all characters to fused Unicode code points (treating *á* and *a* as different characters) and compute Levenshtein distance over these forms.

**Word segmentation:** Spanish and English words both largely use whitespace segmentation, so similar tokenisation strategies can be used.

**Morphophonemics:** Phonetic modeling of words — a component of the method of Han, Cook, and Baldwin (2012) — is available for Spanish using an off-the-shelf Double Metaphone implementation.[1]

**Lexical resources:** A lexicon and slang dictionary — key resources for the method of Han, Cook, and Baldwin (2012) — are available for Spanish.

Overall, English and Spanish text share important features, and we hypothesise that adapting a lexicon-based English normalisation system to Spanish is feasible.

One important component of this Spanish normalisation task is case restoration: e.g., *maria* as a name should be normalised to "Maria". Most previous English Twitter normalisation tasks have focused on lowercase words and ignored capitalisation.

## 3 System Description

The system consists of two steps: (1) downcase all OOVs and normalise them based on a normalisation lexicon which combines entries from existing lexicons (Section 3.1) and entries automatically learnt from a Twitter corpus (Section 3.2); (2) restore case for normalised words (Section 3.3).

[1] `https://github.com/amsqr/Spanish-Metaphone`

### 3.1 Resources

Our normalisation transforms OOV forms to IV words, and thus a Spanish lexicon is required to determine what is OOV. To this end, we use the `Freeling 3.0` Spanish dictionary (Padró and Stanilovsky, 2012) which contains 669k words.

We collected 146 Spanish Internet slang expressions and cell phone abbreviations from the web (SLANG LEXICON).[2] We further extracted normalisation pairs from the development data (DEV LEXICON).

Analysing the development data we noticed that many person names are not correctly capitalised. We formed NAME LEXICON from a list of 277 common Spanish names.[3] This lexicon maps lowercase person names to their correctly capitalised forms.

### 3.2 Corpus-derived Lexicon

The small, manually-crafted normalisation lexicons from Section 3.1 have low coverage over non-standard words. To improve coverage, we automatically derive a much larger normalisation lexicon based on distributional similarity (DIST LEXICON) by adapting the method of Han, Cook, and Baldwin (2012).

We collected 283 million Spanish tweets via the Twitter Streaming API[4] from 21/09/2011–28/02/2012. Spanish tweets were identified using `langid.py` (Lui and Baldwin, 2012). The tweets were tokenised using a simplified English Twitter tokeniser (O'Connor, Krieger, and Ahn, 2010). Excessive repetitions of characters (i.e., $\geq 3$) in words are shortened to one character to ensure different variations of the same pattern are merged. To improve coverage, we removed the restriction from the original work that only OOVs with $\geq 4$ letters were considered as candidates for normalisation.

For a given OOV, we define its *confusion set* to be all IV words with Levenshtein distance $\leq 2$ in terms of characters or $\leq 1$ in terms of Double Metaphone code. We rank the items in the confusion set according to their distributional similarity to the OOV. Han, Cook, and Baldwin (2012) considered many configurations of distributional similarity for normalisation of English tweets. We

[2] `http://goo.gl/wgCFSs` and `http://goo.gl/xsYkDe`, both accessed on 26/06/2013

[3] `https://en.wikipedia.org/wiki/Spanish_naming_customs`

[4] `https://dev.twitter.com`

| Rank | *callendo* | | *guau* | |
|---|---|---|---|---|
| 1 | cayendo | 0.713 | y | 1.756 |
| 2 | saliendo | 3.896 | que | 1.873 |
| 3 | fallando | 4.303 | la | 2.488 |
| 4 | rallando | 6.761 | a | 2.649 |
| 5 | valiendo | 6.878 | no | 3.206 |

Table 1: The KL divergence for the top-five candidates for *callendo* and *guau*.

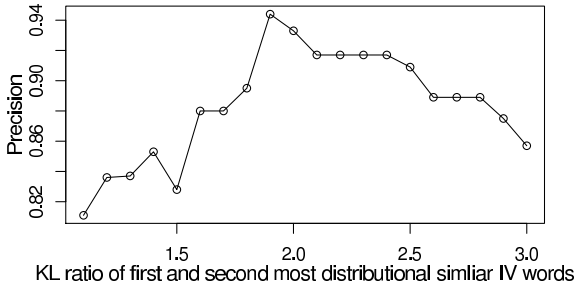


Figure 1: KL divergence ratio cut-off vs. precision of the derived normalisation lexicon on the development data and Slang Lexicon.

use the same settings they selected: context is represented by positionally-indexed bigrams using a window size of ±2 tokens; similarity is measured using KL divergence. An entry in the normalisation dictionary then consists of the OOV and its top-ranked IV.

From development data, we observe that in many cases when a correct normalisation is identified, there is a large difference in KL divergence between the first- and second-ranked IVs. Conversely, if the KL divergence of the first- and second-ranked normalisation candidates is similar, the normalisation is often less reliable. As shown in Table 3.2, *callendo* ("cayendo") is a correctly-derived (OOV, IV) pair, but *guau* ("y") is not.

Motivated by this observation, we filter the derived (OOV, IV) pairs by the KL divergence ratio of the first- and second-ranked IV words for the OOV. Setting a high threshold on this KL divergence ratio increases the reliability of the derived lexicon, but reduces its coverage. This ratio was tested for values from 1.0 to 3.0 with a step size of 0.1 over the development data and Slang Lexicon. As shown in Figure 1, the best precision (94.0%) is achieved when the ratio is 1.9.[5] We directly use this setting to derive the final lexicon, instead of further re-ranking the (OOV,IV) pairs using string similarity.

[5]Here precision is defined as $\frac{\#\text{correct normalisations}}{\#\text{normalisations}}$.

| Lexicon | Accuracy |
|---|---|
| Combined Lexicon | 0.52 |
| − Slang Lexicon | 0.51 |
| − Dev Lexicon | 0.46 |
| − Dist Lexicon | 0.42 |
| − Name Lexicon | 0.51 |
| + Edit distance | 0.54 |
| Baseline | 0.20 |

Table 2: Accuracy of lexicon-based normalisation systems. "−" indicates the removal of a particular lexicon.

### 3.3 Case Restoration

We set the case of each token that was normalised in the previous step (which is down-cased at the current stage) to its most-frequent casing in our corpus of Spanish tweets. We also capitalise all normalised tokens occurring at the beginning of a tweet, or following a period or question mark.

## *4 Results and Discussion*

We evaluated the lexicons using classification accuracy, the official metric for this shared task, on the `tweet-norm` test data. This metric divides the number of correct proposals — OOVs correctly normalised or left unchanged — by the number of OOVs in the collection. This is termed "precision" by the task organisers, but a true measure of precision would be based on the number of OOVs that were actually normalised. We therefore use the term "accuracy" here.

We submitted two runs for the task. The first, Combined Lexicon (Table 4), uses only the combination of lexicons from Section 3, and achieves an accuracy of 0.52. The second run builds on Combined Lexicon but incorporates normalisation based on character edit distance for words with many repeated characters. We observed that such words are often non-standard, and tend not to occur in the lexicons because of their relatively low frequency. For words with $\geq 3$ repeated characters, we remove all but one of the repeated characters, and then select the most similar IV word according to character-based Levenshtein distance. The accuracy of this run is 0.54 (+ Edit distance, Table 4).

We further consider an ablative analysis of the component lexicons of Combined Lexicon. As shown in Table 4, when Slang Lexicon (− Slang Lexicon) or Name Lexicon (− Name Lexicon) are excluded,

accuracy declines only slightly. Although this suggests that existing resources play only a minor role in the normalisation of Spanish tweets, this is likely due in part to the relatively small size of SLANG LEXICON, which is much smaller than similar English resources that have been effectively exploited in normalisation — i.e., 145 Spanish entries versus 5k English entries used by Han and Baldwin (2011). Furthermore, SLANG LEXICON might have little impact due to differences between Spanish Twitter and SMS, the latter being the primary focus of SLANG LEXICON.

On the other hand, normalisation lexicons derived from tweets — whether based on the development data (DEV LEXICON) or automatically learnt (DIST LEXICON) — substantially impact on accuracy (− DEV LEXICON and − DIST LEXICON). These findings for the automatically derived DIST LEXICON are in line with previous findings for English Twitter normalisation (Han, Cook, and Baldwin, 2012) that indicate that such lexicons can substantially improve recall with little impact on precision.

We considered an experiment in which we used COMBINED LEXICON, but ignored case in the evaluation; the accuracy was 0.56. This corresponds to the upper-bound on accuracy if our system performed case restoration perfectly, and suggests that improving the case restoration of our system would not lead to substantial gains in accuracy.

In the final row of Table 4 we show results for a baseline method which makes no attempt to normalise the input. All lexicon-based methods improve substantially over this baseline.

To further analyse our lexicon-based normalisation approach, we categorise the errors for both false positives (OOVs that were normalised, but incorrectly so) and false negatives (OOVs that were not normalised, but should have been). As shown in Table 4, 37% of false positives are incorrect lexical forms, e.g., *algerooo* is normalised to "algero" and not its correct form "alegra". Further examination shows that 23% of these cases are incorrectly normalised to "que", suggesting that distributional similarity alone is insufficient to capture normalisations for some non-standard words.

Surprisingly, we found some OOVs included in the test data, but excluded from the gold-standard annotations (due to tweet deletions), or present in the test data, but not found in the tweets, and excluded in the gold standard. These error types are denoted as "Not available" in Table 4, and account for the second largest source of false positives.

| Error type | Number | Percentage |
|---|---|---|
| Incorrect lexical form | 22 | 37% |
| Not available | 19 | 32% |
| Accent error | 10 | 17% |
| Case error | 5 | 8% |
| One to many | 2 | 3% |
| Annotation error | 1 | 2% |

Table 3: Categorisation of false positives.

Incorrect accents and casing account for 17% and 8% of false positives, respectively. In both of these cases, contextual information, which is not incorporated in the proposed approach, could be helpful. Finally, we identified two one-to-many normalisations (which are outside the scope of our normalisation system), and one case we judged to be an annotation error.

We analysed a random sample of 20 of the 280 false negatives, and found irregular character repetitions and named entities to be the main sources of errors, e.g., *uajajajaa* ("ja") and *Pedroo* ("Pedro").[6] The lexicon-based approach could be improved, for example, by using additional regular expressions to capture repetitions of character sequences. Errors involving named entities reveal the limitations of using the `Freeling 3.0` Spanish dictionary as the IV lexicon, as it has limited coverage of named entities. A corpus-derived lexicon (e.g., from Wikipedia) could help improve the coverage.

## 5 Summary

In this paper, we applied a lexicon-based approach to normalise non-standard words in Spanish tweets. Our analysis suggests that the corpus-derived lexicon based on distributional similarity improves accuracy, but that this approach is limited in terms of flexibility (e.g., to capture accent variation) and lexicon coverage (e.g., of named entities). In future work, we plan to expand the IV lexicon, and incorporate contextual information to improve normalisation involving accents and casing.

[6] *Pedro* is not in our collected list of Spanish names.

### *Acknowledgements*

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme. The authors would like to thank the anonymous reviewers for their valuable feedback and language expertise.

### *References*

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 359–369, Atlanta, USA.

Han, Bo and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 368–378, Portland, Oregon, USA.

Han, Bo, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432, Jeju Island, Korea. Association for Computational Linguistics.

Han, Bo, Paul Cook, and Timothy Baldwin. 2013. Lexical normalisation of short text messages. *ACM Transactions on Intelligent Systems and Technology*, 4(1):5:1–5:27.

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 151–160, Portland, Oregon, USA.

Lui, Marco and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

O'Connor, Brendan, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pages 384–385, Washington, USA.

Padró, Lluís and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2473–2479, Istanbul, Turkey.

Weng, Jianshu and Bu-Sung Lee. 2011. Event detection in Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain.