SINAI at Twitter-Normalization 2013

SINAI en Twitter Normalization 2013

Arturo Montejo Ráez, M. Carlos Diaz Galiano, Eugenio Martínez Cámara, M. Teresa Martín Valdivia, Miguel A. García Cumbreras, L. Alfonso Ureña López
Universidad de Jaén
Campus Las Lagunillas 23071
{amontejo, mcdiaz, emcamara, maite, magc, laurena}@ujaen.es

Resumen: Este artículo presenta el sistema de normalización de tweets desarrollado por el grupo SINAI. Realizamos una serie de conversiones a partir de lexicones de traducción y un corrector ortográfico. Nuestro sistema obtiene un resultado de accuracy bajo, un 37.6%, y analizando los resultados necesita mejorarse en varios aspectos tales como diminutivos y superlativos, tratamiento de entidades o abreviaturas.
Palabras clave: normalización en Twitter, tokenización, traducción, corrector

Abstract: In this paper, we present the Twitter-normalization system developed by the SINAI group. Our system performs a series of conversions on the text by the use of translation lexicons and a spell checker. We obtain a poor result, only 37.6% of accuracy, and after the analysis of these results our system should be improved in areas such as the treatment of diminutives and superlatives, entities or abbreviations.
Keywords: Twitter normalization, tokenization, translation

1 Introduction and objectives

Twitter is a popular medium for broadcasting news, staying in touch with friends, and sharing opinions. Several researches have been focused on this new microblogging platform that is changing the communication way among people. However, tweets often contain highly irregular syntax and nonstandard use of a language. In addition, Twitter posts frequently include URLs, as well as markup syntax, which further decreases the amount of characters available for content. Because of these limits, users have created a novel syntax to communicate their messages with as much brevity as possible. While this brevity allows tweets to contain more information, it makes them harder to mine and analyze the information due to its lack of standardization.

Several works have studied the normalization problem for short text. For example, (Kaufmann and Kalita, 2010) describe a novel system which normalizes these Twitter posts standard form of English by taking a two step-approach, first preprocess tweets to remove as much noise as possible and then feed them into a machine translation model to convert them into standard English. (Han and Baldwin, 2011) target out-of-vocabulary words in short text messages and propose a method for identifying and normalizing ill-formed words that doesn't require any annotations. They use a classifier to detect ill-formed words, and generate correction candidates based on morphophonemic similarity.

On the other hand, most of the studies on short text normalization only deal with English tweets while more and more, other languages are increasingly used on Twitter. For example, there are some works dealing with Spanish tweets (Moreno-Ortiz and Hernández, 2013) but very few are focused on the normalization process.

This paper describes a system which normalizes Spanish Twitter posts, converting them into a more standard form and so natural language processing (NLP) techniques can be more easily applied to them. Next section describes our approach based on the use of translation lexicons and spell checking. Then, the evaluation process is commented and, in addition, we have accomplished an analysis of the obtained results.

2 System Architecture

Our system performs a series of conversions on the text, which is, step by step, trans-

formed into a final normalized form. We have not considered annotation-based approaches like those followed by well-known systems like GATE[1] or proposed by recommendations like the UIMA specification[2]. Instead, we have chosen a straightforward solution, where first the text is tokenized with special attention on Twitter related items (like emoticons, mentions or hashtags) and then each token is converted into some sort of canonical form by the use of translation lexicons and a spell checker. Details of each module are given in the following subsections.

2.1 Tokenization

Tokenization allows the segmentation of texts into their most simple units of meaning: terms. In our case, multi-word forms are not considered, so each term is either related to a word or to other type of information like: emoticons, HTML tags, telephone numbers, mentions, hashtags, dates, URLs, e-mail addresses and some other minor items. Case is preserved during the tokenization process and, as result, we obtain a list of strings to feed next modules.

2.2 Translation tables

A translation table allows for the replacement of certain forms of strings into other forms. In this way, we can recognize some expressions and translate them to more convenient representations. In this step, the following translation tables have been considered:

1. **Abbreviations**. Expressions like "*a2*" are translation into "*adiós*", "*q*" into "*que*" and so on up to twelve possible Spanish abbreviations commonly used in "texting" communication.
2. **Laughings**. This translation table make intensive use of regular expressions in order to capture most possible forms of laughing expressions found in text. In this way, "*aajajajaaj*" would be replaced by "*ja*", for example.

2.3 Spell checking

For this module we have used the GNU Aspell[3] spell checker and its binding for Python, *aspell-python*[4]. GNU Aspell is an open source spell checker that works well with Unicode strings, which makes it very suitable for multilingual texts. Also, it allows multiple dictionaries to be used concurrently and the addition of further vocabularies to be considered as correct forms, so we can integrate more lexicons. Aspell works by converting the misspelled word (that is, the word not included in their dictionaries) into a *sounds like* equivalent. Then proposes a list of words with one or two *edit distances* from the original words sounds like. An edit distance is one replacement, insertion or deletion of one single character.

We have added into Aspell the following lexicons:

- **Main provinces and cities in Spain**, extracted from the INE (Statistics National Institute of Spain)[5]
- **Interjections** like "*ajá*", "*jolín*" or "*puf*" among others. This list is a selection from the ones proposed in Wiktionary[6]
- **Twitter jargon and neologisms**, with terms like "*Facebook*" or "*tuiteo*", selected from an on-line glossary[7].
- **Named entities**, generated from Wikipedia and containing more than 650 different named entities. Also, political parties and main political leaders have been added to this list manually.

2.4 Automatic spelling correction

After receiving a list of possible spelling corrections from the previous module, the system selects the most common term, according to a list of words sort by frequency generated by (Vega et al., 2011). Although more sophisticated solutions could be used here (like considering surrounding words as context for candidate selection), our attempts applying techniques taken from word sense desambiguation approaches did not lead to significant improvements.

To consider surrounding words as context, first we have calculated a table with normalized pointwise mutual information (NPMI) of lemmatized words in the same sentence.

[1] http://gate.ac.uk/
[2] http://uima.apache.org
[3] http://aspell.net/
[4] http://0x80.pl/proj/aspell-python/
[5] http://www.ine.es/daco/daco42/codmun/cod_provincia.htm
[6] http://es.wiktionary.org/wiki/Categor%C3%ADa:ES:Interjecciones
[7] http://estwitter.com/glosario/

To calculate this table we have used a dump of Spanish Wikipedia[8] articles and calculated the NPMI values of the first 10.000 lemmas most frequents. Second, we have computed the sum of NPMI values of a candidate with each word of the context. Finally, we have selected the candidate with the best sum of NPMI.

3 Evaluation and results

The performance reached by the system showed above are not good according with the results published by the organization. After a deep analysis of the results we have realized that we have to improve the following issues:

1. Diminutives and superlatives: We have followed an approach based on a Spanish lemma dictionary. The Spanish lemma dictionary used was the offered by the project LingPipe[9]. This dictionary does not include a great amount of diminutives and superlatives, so one of the weaknesses of our system is the detection of this kind of words and a set of the errors are caused by them.
2. New words: Aspell is a dictionary bases on spell checker. It is also possible to add more list of words to Aspell with the aim of enlarging the tool coverage. However the coverage of all the Spanish language is not easy. Other problem is the new Spanish words included in the RAE (Royal Spanish Language Academy) because those ones are difficult to find out in the classic spell checker tools. Although, we have appended to the Aspell dictionaries new Spanish words, they have not been enough and the system has failed in words such as "flipante" or "sobao".
3. Entities: The misclassification of entities has been other error of our system. The entities without any error must be classified as 1 (CORRECT, NO VARIATION) but our system considered them correct. Also the entity recognition power of our system is not strength, so some of the errors are related with this problem. A clear example is the entity Vallecas, which was not recognized by our system as an entity, so it was replaced by the word Vacas.
4. Abbreviations: Although we have compiled a bag of abbreviations, after the publications of the results we have realized that they are not enough and we need to add more abbreviations.

We have detected some errors in the organization results. Laughing expressions like "jajaja" have been normalized in some tweets but other have not, so we do not know if our right normalization of some laughing expressions have been considered as correct. Other example of words that we think they have to be normalized is "que" that some users write as "q". In some tweets like "#Escorpio Puedes sentir q el camino es muy oscuro, será mejor q busques q alguien te ayude a iluminarlo puede ser algun amigo.", the organizers considered "q" well written and we are not agree. The organizers also think that the word "días" without accent is well written and it is not. Due to that, we think that the test corpus have to be improved for future editions of the workshop.

Those are some of the reasons because our system has reached only 37.6% of accuracy.

4 Conclusions and ongoing work

In this paper, we have proposed a normalization system for tweets that performs a series of conversions on the text by the use of translation lexicons and a spell checker. We found that most illformed words are based on morphophonemic variation and proposed a cascade method to convert each tweet. Our system has reached only 37.6% of accuracy.

Our future work will be focused on resolve some problems discovered such as the treatment of diminutives and superlatives, entities or abbreviations. Furthermore, we want to adapt our normalization system for subsequent processes such as sentiment analysis or text classification.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), TEXT-COOL 2.0 project (TIN2009-13391-C04-02) and ATTOS project (TIN2012-38536-C03-0) from the Spanish Government. The project AORESCU (TIC - 07684) from the regional government of Junta de Andalucía partially

[8] http://dumps.wikimedia.org/eswiki/

[9] http://alias-i.com/lingpipe/

supports this manuscript. Also,this paper is partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

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

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

Moreno-Ortiz, Antonio and Chantal Pérez Hernández. 2013. Lexicon-based sentiment analysis of twitter messages in spanish. *Procesamiento del Lenguaje Natural*, 50(0).

Vega, Fernando Cuetos, María González Nosti, Analía Barbón Gutiérrez, and Marc Brysbaert. 2011. Subtlex-esp: Spanish word frequencies based on film subtitles. *Psicológica: Revista de metodología y psicología experimental*, 32(2):133–143.