Mining Lexical Variants from Microblogs: An Unsupervised Multilingual Approach

Alejandro Mosquera
University of Alicante
San Vicente del Raspeig s/n - 03690
Alicante, Spain
amosquera@dlsi.ua.es

Paloma Moreda
University of Alicante
San Vicente del Raspeig s/n - 03690
Alicante, Spain
moreda@dlsi.ua.es

Abstract

User-generated content has become a recurrent resource for NLP tools and applications, hence many efforts have been made lately in order to handle the noise present in short social media texts. The use of normalisation techniques has been proven useful for identifying and replacing lexical variants on some of the most informal genres such as microblogs. But annotated data is needed in order to train and evaluate these systems, which usually involves a costly process. Until now, most of these approaches have been focused on English and they were not taking into account demographic variables such as the user location and gender. In this paper we describe the methodology used for automatically mining a corpus of variant and normalisation pairs from English and Spanish tweets.

1 Introduction

User-generated content (UGC), and specially the microblog genre, has become an interesting resource for Natural Language Processing (NLP) tools and applications. Many are the advantages of exploiting this real-time stream of multilingual textual data. Popular applications such as Twitter has an heterogeneous user base of almost 600 million users that generate more than 60 million new tweets every day. For this reason, Twitter has become one of the most used sources of textual data for NLP with several applications such as sentiment analysis (Tumasjan et al., 2010) or realtime event detection (Sakaki et al., 2010). Recent advances on machine translation or information retrieval systems have been also making an extensive use of UGC for both training and evaluation purposes. However, tweets can be very noisy and sometimes hard to understand for both humans (Mosquera et al., 2012) and NLP applications (Wang and Ng, 2013), so an additional preprocessing step is usually required.

There have been different perceptions regarding the lexical quality of social media (Rello and Baeza-Yates, 2012) (Baldwin et al., 2013) and even others suggested that 40% of the messages of Twitter were "pointless babble" (PearAnalytics, 2009). Most of the out of vocabulary (OOV) words present in social media texts can be catalogued as lexical variants (e.g. "See u 2moro" → "See you tomorrow"), that are words lexically related with their canonic form.

The use of text normalisation techniques has been proven useful in order to clean short and informal texts such as tweets. However, the evaluation of these systems requires annotated data, which usually involves costly human annotations. There are previous works about automatically constructing normalisation dictionaries, but until now, most of these approaches have been focused on English and they were not taking into account demographic variants. In this paper we describe the methodology used for automatically mining lexical variants from English and Spanish tweets associated to a set of headwords. These formal and informal pairs can be later used to train and evaluate existing social media text normalisation systems. Additional metadata from Twitter such as geographic location and user gender is also collected, opening the possibility to model and analyse gender or location-specific variants.

This paper is organised as follows. We describe the related work in Section 2. We then describe our variant mining methodology in Section 3. The obtained results are presented in Section 4. Section 5, draws the conclusions and future work.

Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) @ EACL 2014, pages 1–7,
Gothenburg, Sweden, April 26-30 2014.

2 Related Work

One way to handle the performance drop of NLP tools on user-generated content (Foster et al., 2011) is to re-train existing models on these informal genres (Gimpel et al., 2011), (Liu et al., 2011b). Another approaches make use of preprocessing techniques such as text normalisation in order to minimise the social media textual noise (Han et al., 2013), (Mosquera and Moreda, 2012) where OOV words were first identified and then substituted using lexical and phonetic edit distances. In order to enhance both precision and recall both OOV detection and translation dictionaries were used. Moreover, the creative nature of informal writing and the low availability of manually-annotated corpora can make the improvement and evaluation of these systems challenging.

Motivated by the lack of annotated data and the large amount of OOV words contained in Twitter, several approaches for automatically constructing a lexical normalisation dictionary were proposed; In (Gouws et al., 2011) a normalisation lexicon is generated based on distributional and string similarity (Lodhi et al., 2002) from Twitter. Using a similar technique, a wider-coverage dictionary is constructed in (Han et al., 2012) based on contextually-similar (OOV, IV) pairs. More recently, (Hassan and Menezes, 2013) introduced another context-based approach using random walks on a contextual similarity graph.

Distributional-based methods can have some drawbacks: they rely heavily on pairwise comparisons that make them computationally expensive, and as the normalisation candidates are selected based on context similarity they can be sensitive to domain-specific variants that share similar contexts. Moreover, these approaches were focusing on extracting English lexical variants from social media texts, but due the heterogeneity of its users, lexical distributions can be influenced by geographical factors (Eisenstein et al., 2010) or even gender (Thomson and Murachver, 2001).

To the best of our knowledge, there are not multilingual approaches for mining lexical variants from short, noisy texts that also take into account demographic variables. For this reason, we present an unsupervised method for mining English and Spanish lexical variants from Twitter that collects demographic and contextual information. These obtained pairs can be later used for training and evaluating text normalisation and inverse text normalisation systems.

3 Lexical Variant Mining

Lexical variants are typically formed from their standard forms through regular processes (Thurlow and Brown, 2003) and these can be modelled by using a set of basic character transformation rules such as letter insertion, deletion or substitution (Liu et al., 2011a) e.g. ("tmrrw" → "2morrow") and combination of these ("2moro"). The relation between formal and informal pairs is not always 1-to-1, two different formal words can share the same lexical variant ("t" in Spanish can represent "te" or "tú") and one formal word can have many different variants (e.g. "see you" us commonly shortened as "c ya" or "see u"). As a difference with previous approaches based on contextual and distributional similarity, we have chosen to model the generation of variant candidates from a set of headwords using transformation rules. These candidates are later validated based on their presence on a popular microblog service, used in this case as a high-coverage corpus.

3.1 Candidate Generation

We have defined a set of 6 basic transformation rules (see Table 1) in order to automatically generate candidate lexical variants from the 300k most frequent words of Web 1T 5-gram (English) (Brants and Franz, 2006) and SUBTLEX-SP (Spanish) (Cuetos et al., 2011) corpora.

Rule	Example
a) Character duplication	"goal" → "gooal"
b) Number transliteration	"cansados" → "cansa2"
c) Character deletion	"tomorrow" → "tomrrw"
d) Character replacement	"friend" → "freend"
e) Character transposition	"maybe" → "mabye"
f) Phonetic substitution	"coche" → "coxe"
g) Combination of above	"coche" → "coxeee"

Table 1: Transformation rules.

As modelling some variants may need more than one basic operation, and lexically-related variants are usually in an edit distance t where $t <= 3$ (Han et al., 2013), the aforementioned rules were implemented using an engine based on stacked transducers with the possibility to apply a maximum of three concurrent transformations:

(a) Character duplication: For words with n characters, while n>19 each character were

duplicated n times (∀ n>0, n<4), generating n^3 candidate variants.

(b) Number transliteration: Words and numbers are transliterated following the language rules defined in Table 2.

Rule	Lang.
"uno" → "1"	SP
"dos" → "2"	SP
"one" → "1"	EN
"two" → "2"	EN
"to" → "2"	EN
"three" → "3"	EN
"for" → "4"	EN
"four" → "4"	EN
"eight" → "8"	EN
"be" → "b"	EN
"a" → "4"	EN
"e" → "3"	EN
"o" → "0"	EN
"s" → "5"	EN
"g" → "6"	EN
"t" → "7"	EN
"l" → "1"	EN

Table 2: Transliteration table for English and Spanish.

(c) Character deletion: The candidate variants from all possible one character deletion combinations plus the consonant skeleton of the word will be generated.

(d) Character replacement: Candidate variants are generated by replacing n characters (∀ n>0, n<7) by their neighbours taking into account a QWERTY keyboard and an edit distance of 1.

(e) Character transposition: In order to generate candidate lexical variants the position of adjacent characters are exchanged.

(f) Phonetic substitution: A maximum of three character n-grams are substituted for characters that sound similar following different rules for Spanish (Table 3) and English (Table 4).

3.2 Candidate Selection

We have explored several approaches for filtering common typographical errors and misspellings, as these are unintentional and can not be technically considered lexical variants, in order to do this we have used supervised machine learning techniques. Also, with aim to filter uncommon or low quality variants, the Rovereto Twitter corpus (Herdagdelen, 2013) was initially used in order to rank the English candidates present in the corpus by their frequencies. The 38% of the variants generated by one transformation were successfully found, however, performing direct Twitter search API queries resulted to have better coverage than using a static corpus (90% for English variants).

Rule
"b"→["v" or "w"]
"c"→["k"]
"s"→["z"]
"z"→["s"]
"c"→["s"]
"x"→["s"]
"ñ"→["ni"]
"ch"→["x"]
"gu"→["w"]
"qu"→["k"]
"ll"→["y"]
"ge"→["je"]
"gi"→["ji"]
"ll"→["i"]
"hue"→["we"]

Table 3: Phonetic substitution table for Spanish.

3.2.1 Intentionality Filtering

Given an OOV word *a* and its IV version *b* we have extracted character transformation rules from *a* to *b* using the longest common substring (LCS) algorithm (See Table 5). These lists of transformations were encoded as a numeric array where the number each transformation counts were stored. We have used NLTK (Bird, 2006) and the SequenceMatcher Python class in order to extract those sets of transformations taking into account also the position of the character (beginning, middle or at the end of the word).

A two-class SVM (Vapnik, 1995) model has ben trained using a linear kernel with a corpus composed by 4200 formal-variant pairs extracted from Twitter [1], SMS[2] and a corpus of the 4200 most common misspellings [3]. In table 6 we show the k-fold cross-validation results (k=10) of the model, obtaining a 87% F1. This model has been used in order to filter the English candidate variants classified as not-intentional.

To the best of our knowledge there are not similar annotated resources for Spanish, so this classifier was developed only for English variants. However, would be possible to adapt it to work for

[1] $http://ww2.cs.mu.oz.au/\ hanb/emnlp.tgz$
[2] $http://www.cel.iitkgp.ernet.in/\ monojit/sms$
[3] $http://aspell.net/test/common-all/$

Rule
”i”→[“e”]
“o”→[“a”]
“u”→[“o”]
“s”→[“z”]
“f”→[“ph”]
“j”→[“ge” or “g”]
“n”→[“kn” or “gn”]
“r”→[“wr”]
“z”→[“se” or “s”]
“ea”→[“e”]
“ex”→[“x”]
“ae”→[“ay” or “ai” or “a”]
“ee”→[“ea” or “ie” or “e”]
“ie”→[“igh” or “y” or “i”]
“oe”→[“oa” or “ow” or “o”]
“oo”→[“ou” or “u”]
“ar”→[“a”]
“ur”→[“ir” or “er” or “ear” or “or”]
“or”→[“oor” or “ar”]
“au”→[“aw” or “a”]
“er”→[“e”]
“ow”→[“ou”]
“oi”→[“oy”]
“sh”→[“ss” or “ch”]
“ex”→[“x”]
“sh”→[“ss” or “ch”]
“ng”→[“n”]
“air”→[“ear” or “are”]
“ear”→[“eer” or “ere”]

Table 4: Phonetic substitution table for English.

another languages if the adequate corpora is provided. Because of the lack of this intentionality detection step, the number of generated candidate variants for Spanish was filtered by taking into account the number of transformations, removing all the variants generated by more than two operations.

3.2.2 Twitter Search

The variants filtered during the previous step were searched on the real time Twitter stream for a period of two months by processing more than 7.5 million tweets. Their absolute frequencies n were used as a weighting factor in order to discard not used words ($n > 0$). Additionally, variants present in another languages rather than English or Spanish were ignored by using the language identification tags present in Twitter metadata.

There were important differences between the final number of selected candidates for Spanish, with 6 times less variant pairs and English (see Table 7). Spanish language uses diacritics that are commonly ignored on informal writing, for this reason there is a higher number of possible combinations for candidate words that would not generate valid or used lexical variants.

Formal/Informal pair	Transf.	Pos.
house → h0use	o → 0	middle
campaign → campaing	n → ∅	end
	∅ → n	middle
happy → :)	happy → :)	middle
embarrass → embarass	r→ ∅	middle
acquaintance →	∅ → q	middle
aqcuaintance	q → ∅	middle
virtually → virtualy	l→ ∅	middle
cats → catz	s→ z	end

Table 5: Example of formal/informal pairs and the extract transformations.

Method	Precision	Recall	F1
SVM	0.831	0.824	0.827
SVM+Pos.	0.878	0.874	0.876

Formal/Informal pair	Verdict
you → yu	intentional
accommodate → acommodate	unintentional
business → bussiness	unintentional
doing → doin	intentional
acquaintance → aqcuaintance	unintentional
basically → basicly	unintentional
rules → rulez	intentional

Table 6: Cross-validation results of intentionality classification with examples.

4 Results

Besides the original message and the context of the searched variant, additional metadata has been collected from each tweet such as the gender and the location of the user. In Twitter the gender is not explicitly available, for this reason we applied an heuristic approach based on the first name as it is reported in the user profile. In order to do this, two list of male and female names were used: the 1990 US census data [4] and popular baby names from the US Social Security Administration's statistics between 1960 and 2010 [5].

We have analysed the gender and language distribution of the 6 transformation rules across the mined pairs (see Figure 1). On the one hand, lexical variants generated by duplicating characters were the most popular specially between female

[4] $census.gov/genealogy/www/data/1990surnames$
[5] $ssa.gov/cgi-bin/popularnames.cgi$

Candidates	Selected	Lang.
2456627	48550	EN
1374078	8647	SP

Table 7: Number of generated and selected variants after Twitter search.

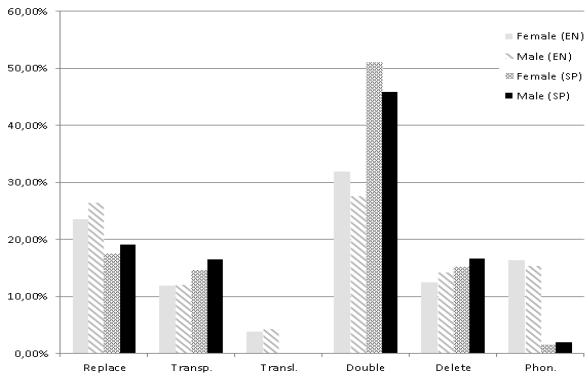

Figure 1: Transformation trends by gender.

users with a 5% more than their male counterparts. On the other hand, variants generated by character replacement and deletion were found a 2% more on tweets from male users. The differences between English and Spanish were notable, mostly regarding the use of transliterations, that were not found on Spanish tweets, and phonetic substitutions, ten times less frequent than in English tweets.

For the distribution of transformations across geographic areas, we have just taken into account the countries where the analysed languages have an official status. Lexical variants found in Tweets from another areas are grouped into the "Non-official" label (see Figure 2). The biggest differences were found on the use of transliterations (higher in UK and Ireland with more than a 5%) and phonetic substitutions (higher in Pakistani users with more than a 22%). Transformation frequencies from non-official English speaking countries were very similar as the ones registered for users based on United States and Canada.

Spanish results were less uniform and showed more variance respect the use of character duplication (57% in Argentina), character replacement (more than 24% in Mexico and Guatemala) and character transposition (with more than a 19% for users from Cuba, Colombia and Mexico) (see Figure 3).

5 Conclusions and Future Work

In this paper we have described a multilingual and unsupervised method for mining English and Spanish lexical variants from Twitter with aim to close the gap regarding the lack of annotated corpora. These obtained pairs can be later used for the training and evaluation of text normalisation systems without the need of costly human annotations. Furthermore, the gathered demographic and contextual information can be used in order to model and generate variants similar to those that can be found on specific geographic areas. This has interesting applications in the field of inverse text normalisation, that are left to a future work. We also intend to explore the benefits of feature engineering for the detection and categorisation of lexical variants using machine learning techniques.

Acknowledgments

This research 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 author, and the Commission cannot be held responsible for any use which may be made of the information contained therein. Moreover, it has been partially funded by the Spanish Government through the project "Análisis de Tendencias Mediante Técnicas de Opinión Semántica" (TIN2012-38536-C03-03) and "Técnicas de Deconstrucción en la Tecnologías del Lenguaje Humano" (TIN2012-31224).

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1. Technical report, Google Research.

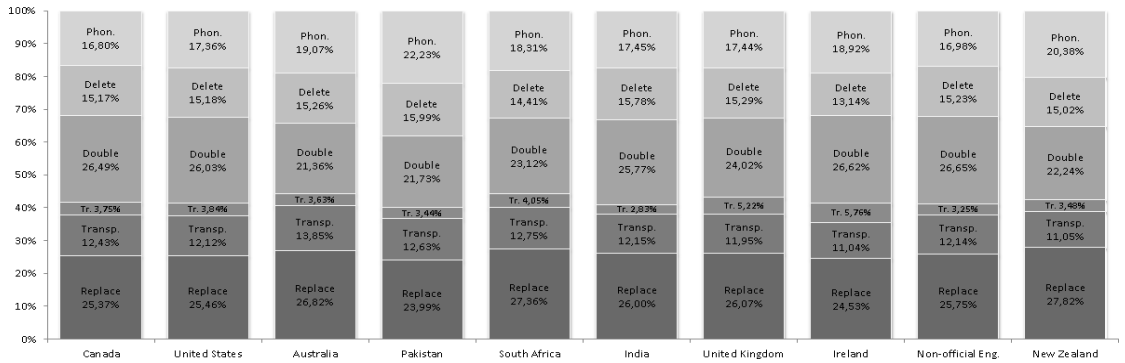

Figure 2: Transformation trends by English-speaking countries.

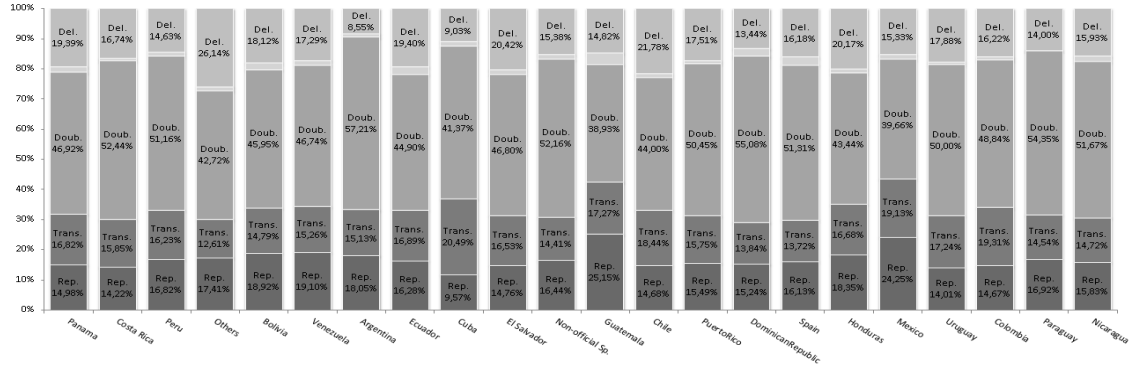

Figure 3: Transformation trends by Spanish-speaking countries.

Fernando Cuetos, Maria Glez-Nosti, Anala Barbn, and Marc Brysbaert. 2011. Subtlex-esp: Spanish word frequencies based on film subtitles. *Psicolgica*, 32(2).

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1277–1287, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. 2011. #hardtoparse: Pos tagging and parsing the twitterverse. In *Analyzing Microtext*, volume WS-11-05 of *AAAI Workshops*. AAAI.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Gouws, D. Hovy, and D. Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, page 82–90.

Bo Han, 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 (EMNLP-CoNLL 2012)*, pages 421–432, Jeju Island, Korea.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, 4(1):5:1–5:27, February.

Hany Hassan and Arul Menezes. 2013. Social text normalization using contextual graph random walks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1577–1586, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ama Herdagdelen. 2013. Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, 47(4):1127–1147.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011a. Insertion, deletion, or substitution?: Normalizing text messages without pre-categorization

nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 71–76, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011b. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 359–367, Stroudsburg, PA, USA. Association for Computational Linguistics.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, March.

Alejandro Mosquera and Paloma Moreda. 2012. Tenor: A lexical normalisation tool for spanish web 2.0 texts. In *Text, Speech and Dialogue - 15th International Conference (TSD 2012)*. Springer.

Alejandro Mosquera, Elena Lloret, and Paloma Moreda. 2012. Towards facilitating the accessibility of web 2.0 texts through text normalisation. In *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA) ; Istanbul, Turkey.*, pages 9–14.

PearAnalytics. 2009. Twitter study. In *Retrieved December 15, 2009 from http://pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf*.

Luz Rello and Ricardo A Baeza-Yates. 2012. Social media is not that bad! the lexical quality of social media. In *ICWSM*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Robert Thomson and Tamar Murachver. 2001. Predicting gender from electronic discourse.

Thurlow and Brown. 2003. Generation txt? the sociolinguistics of young people's text-messaging.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178–185.

Vladimir N. Vapnik. 1995. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Pidong Wang and Hwee Tou Ng. 2013. A beam-search decoder for normalization of social media text with application to machine translation. In *HLT-NAACL*, pages 471–481.