MT @HanBaldwin: Fightin OOVs in German #twitter

Max Mustermann, Otto Normalverbraucher, and John Doe

University of Universe, Milky Way City,
max@example.com, otto@example.com, john@example.com,
WWW home page: http://www.example.com

Abstract. This article gives an overview of existing approaches to the problems of out-of-vocabulary tokens and noisiness effects in natural language texts. Additionally, quantitative and qualitative analyses of unknown words in German tweets are conducted in order to assess how widely this textual genre differs from standard language texts with regard to noisiness phenomena. Subsequently, we present a set of ad-hoc techniques which are supposed to tackle some of the most prominent disturbing effects and show how this set of techniques allows us to lower the average rate of out-of-vocabulary tokens in Twitter messages.

Keywords: twitter, social media, text normalization, spelling correction

1 Introduction

When Jack Dorsey, the present CEO of Twitter Inc., was sending the very first tweet on March 21, 2006 (Dorsey, 2006), he probably could not imagine that a few years later presidents and government officials would use this service to communicate with their voters and the Pope would be posting short messages holding an iPad in his hand (Pianigiani, 2012). Yet another thing that Jack Dorsey was apparently not aware of at that moment, was the fact that his message – "just setting up my twttr" – already contained a word which was unknown to the majority of NLP applications existing at that time, and that there would be many of such words in future causing a lot of headache to natural language specialists.

Though the problem of out-of-vocabulary (OOV) words and its related task of text normalization have been extensively studied in computational linguistics since as early as the late 1950s (cf. Petersen, 1980) and were certainly anything but new at the time when mobile communication emerged, it were small messages that revived interest in this area in the past two decades.

In the next Section we will give a short overview of recent scientific approaches to the problem of tackling out-of-vocabulary words in non-standard texts. After that, in Section 3 we will analyze which types of text noisiness phenomena are especially characteristic for German Twitter. Section 4 will subsequently describe an automatic procedure for mitigating some of the most obvious of those effects. In a concluding step, we will perform an evaluation of the results of this procedure and give further suggestions for future research.

2 Related Work

Before proceeding with the description of existing methods for noisy text normalization (NTN) task, we first would like to define the criteria by which these methods could be classified. It should be noted that there already exist several classifications of NTN methods including, for example, Kukich (1992), Kobus et al. (2008), and Sproat et al. (2001). But all of these classification attempts seem to have one significant shortcoming, namely they try to involve several independent criteria within one classification scheme.

Kukich (1992), for example, divided NTN techniques into six classes:

1. minimum edit distance techniques;
2. similarity key techniques;
3. rule-based techniques;
4. *n*-gram based techniques;
5. probabilistic techniques;
6. neural nets.

Kobus (2008), on the contrary, referred to existing NTN methods as *metaphors* and split them into the following three groups:

1. "spell checking" metaphor;
2. "translation" metaphor;
3. "speech recognition" metaphor.

Though both divisions seem to be justified to some extent, it is difficult to determine using them whether an NTN approach that detects and restores incorrectly spelled words on the basis of phonetical *n*-gram statistics should fall into *n*-gram based, probabilistic, spell checking or speech recognition class.

We instead suggest using two separate types of criteria which are unrelated to each other and characterize NTN approaches from different points of view.

The first criterion is *segmentation level*. It depends primarily on the maximal length of syntagmatic segments which are used *a*) to infer in-vocabulary (IV) equivalents for OOV tokens and *b*) to choose the most probable variant among multiple possible suggestions[1]. For this criterion we propose division into following classes:

1. graphematic[2];
2. lexical;

[1] For the cases, when segments of different lengths are used for tasks a and b, we note it explicitly in our classification on which segmentation length each subtask relies.

[2] Depending on whether phonetical information is involved or not at this level, this class could be further divided into a phonographematic and purely graphematic subclasses.

3. phrasal.

Each broader level of this hierarchy is also supposed to either incorporate or ignore information provided by its narrower subsegments. In this way, we only need to mention one (the broadest) hierarchical class for the cases when multiple segmentation levels are involved by some techniques.

The second classification criterion regards the *type of information induction* that was used to devise the correction directives. This leads us to the usual NLP-taxonomy which divides all approaches into:

1. rule-based;
2. statistical[3];
3. and hybrid ones.

Equipped with these two criteria, we could now proceed to the description of main scientific works on NTN task which were done in recent years.

It is worthwile noting that many approached to NTN, especially the earlier ones, were heavily influenced by the idea of noisy channel model. The underlying notion of this model was that observed sentences had to be considered as corrupted versions of some original language signals. In order to restore the undistorted signal, one had to come up with two probabilistic submodels: the error model (one which caused the corruption) and the language model (one which reflected the nature of initial signal source). To these works belong Brill and Moore (2000), Sproat et al. (2001), Toutanova and Moore (2002), Clark (2003), Choudhury et al. (2007), Cook and Stevenson (2009), Beaufort et al. (2010) etc. The majority of these methods used either purely graphematic (Brill and Moore, 2000; Sproat et al., 2001; Clark, 2003) or phonographematic segmentation (Toutanova and Moore 2002; Choudhury et al., 2007; Cook and Stevenson, 2009) for error models. For language model, normally phrasal segments (word n-grams) were used. From the point of view of information induction almost all of these techniques could be characterized as supervised statistical, with the exception of Cook and Stevenson (2009) who claimed to use an unsupervised technique.

The raising influence and improved quality of machine translation tools and applications in the 2000s lead to the development of NTN technologies which used broader levels of segmentation. In 2006, Aw et al. suggested a supervised statistical system for normalization of SMS-messages. Their system operated exclusively on phrasal level. A few years later, Clark and Araki (2011) described a purely rule-based phrasal technique.

As stated by Kobus et al. (2008), NTN methods relying on either graphematic or phrasal segments usually revealed complementary strengths and weaknesses. This notion led NLP scientists to the idea that incorporating multiple levels of language in one NTN system could improve the total quality of the system as a whole since different levels would benefit from each other. As a

[3] Depending on the type of training data required, this class is usually divided into unsupervised, semi-supervised, and supervised groups.

consequence of this notion, a wealth of combined techniques emerged in the past few years. Among these we should especially mention Kobus et al. (2008), Kaufmann (2010), Han and Baldwin (2011), and Oliva et al. (2013). The system suggested by Kobus et al. (2008), for example, used a hybrid phrasal approach in a pre-processing step and subsequently fed the output of this pre-processing into a finite state transducer (FST). The FST performed a phonographematic segmentation of data and derived normalization equivalents on the basis of statistical inference. Another approach proposed by Kaufmann (2010) first used unambigous lexical mappings and straightforward graphematic correction rules to reduce the noisiness effects and then redirected pre-normalized input to a statistical phrasal MT system.

Eventually, in 2011, an article called "Lexical Normalization of Short Text Messages: Makn Sens a #twitter" was published by Han and Baldwin. In this article the authors separated the tasks of identification of ill-formed words and finding appropriate correction for them. For the former problem, they first generated a confusion set (CS) for each word unknown to `GNU aspell`. Based on this set, the decision was made whether a particular word had to be corrected or regarded as vlid. Subsequently, for words identified as ill-formed the most probable restoration candidate was chosen from CS by combining features resulting from dictionary lookup, analysis of surrounding context, and estimating word similarity to each proposed correction variant. According to authors' estimations, this combination allowed them to outperform most of the NTN methods existing at that time.

In our next section will also perform quantitive and qualitative analyses of unknown words in German Twitter texts in order to assess whether and to what extent the problem of ill-formed words is relevant for them. Relying on these analyses we then will suggest a set of ad-hoc techniques which could help one lower the average rate of incorrectly spelled words in this kind of text.

3 Analysis of Unknown Tokens

In order to estimate the percentage of unknown words in Twitter messages, we randomly selected 10,000 tweets from a previously collected corpus, split them into sentences and tokenized using social media aware tokenizer by Christopher Potts [4]. After skipping all words which did not contain any alphabetic character or consisted only of a single letter, we obtained a list of 129,146 tokens. As reference systems for dictionary lookup we used open source spell checking program `hunspell`[5] and publicly available part-of-speech tagger `TreeTagger`[6] (Schmid, 1994).

Out of this token list, 26,018 tokens (20.15 %) were regarded as unknown by `hunspell` and 28,389 tokens (21.98 %) were considered as OOV by `TreeTagger`. We also performed similar estimations after leaving only unique words without

[4] `http://sentiment.christopherpotts.net/code-data/happyfuntokenizing.py`

[5] Ispell Version 3.2.06 (Hunspell Version 1.3.2); dictionary de_DE.

[6] Version 3.2 with German parameter file UTF-8.

taking into account their frequencies. This allowed us to shrink our initial token list by four times to 32,538 unique tokens. The relative rate of unknown words raised as expected and run up to 46.96 % for `hunspell` and 58.24 % for `TreeTagger`.

Here once again the question of classification arose but this time with regard to the reasons, why different tokens could have been omitted from corresponding applications' dictionaries. In this respect, division into following groups seemed to be appropriate for us:

1. **Objective limitedness of machine-readable dictionary (MRD)**. To this group we counted words of basic vocabulary which did not get into applications' MRD either because they supposedly were rare or because they did not exist at the time when dictionaries were created. Another reason for inclusion in this type was the belonging of a word to an open lexical or part-of-speech class (like, for example, named entities or interjections) which are usually deliberately not included in MRDs due to impossibility to cover these classes fully;
2. **Unintended sloppiness of user's input**. In the scope of this group we considered unintended typos including unexpected truncation of words at the end of Twitter messages[7];
3. **Stylistic specifics of text genre**. This group comprised words which could be considered as illegal from the point of view of standard language texts but were perfectly valid terms in the domain of web discourse and Twitter communication.

In order to see how out-of-vocabulary words were distributed among and within these 3 major groups, we manually analyzed all OOV tokens which appeared in text more than once and also looked at 1,000 randomly selected hapax legomena. The results of these estimations are shown and explained below.

We subdivided group 1 into the following subgroups:

1. regular German words, e.g. *aufm*, *losziehen*;
2. compounds, e.g. *Altwein*, *Amtsapothekerin*;
3. abbreviations, e.g. *NBG*, *OL*;
4. interjections, e.g. *aja*, *haha*;
5. named entities, with subclasses:
   (a) persons, e.g. *Ahmadinedschad*, *Schweiger*;
   (b) geographic locations, e.g. *Biel*, *Limmat*;
   (c) companies, e.g. *Apple*, *Facebook*;
   (d) product names, e.g. *iPhone*, *MacBook*;
6. neologisms, with subclasses:
   (a) newly coined German terms, e.g. *entfolgen*, *geskypt*;
   (b) loanwords, e.g. *Community*, *Stream*;

[7] As is generally known, Twitter imposes a strong restriction on the length of posted messages which can be no longer than 140 characters. Upon exceeding this length, tweets get automatically truncated.

7. and, finally, foreign words like *is* or *now* which in contrast to 6b were not mentioned in any existing German lexicons nor they complied with inflectional rules of German grammar.

Though this division is admittedly arbitrarily to a certain degree and involves different linguistic criteria simultaneously, the underlying notion for it was simple. Valid words could have been omitted from an MRD either due to the limitation of developers' capacities (group 1), active word formation or lexical productivity of the language (groups 2 through 6a) or also due to language's openness to foreign language systems (groups 6b and 7).

In Table 1 percentage figures for each of the above subgroups are shown. We have considered OOV-distribution for both `hunspell` and `TreeTagger`. For both of them we estimated the percentage of particular subclass with regard to the total number of occurrences of OOV-tokens (column "% of total OOVs") as well as with regard to their percentage in the list of unique OOVs (column "% of unique OOVs") without taking into account their frequencies.

Table 1. Distribution of OOV words belonging to the class "objective limitedness of MRD"

OOV subclass	`hunspell`		`TreeTagger`	
	% of total OOVs	% of unique OOVs	% of total OOVs	% of unique OOVs
regular German words	7.88	8.85		
compounds	1.22	2.42		
abbreviations	4.12	4.91		
interjections	5.93	4.47		
person names	4.79	6.45		
geographic locations	1.51	2.53		
company names	2.26	2.8		
product names	2.08	2.49		
newly coined terms	1.22	1.24		
loanwords	3.72	4.05		
foreign words	11.68	14		
total	46.41	54.21		

Similarly to class 1 we subdivided the group of misspellings into the following subgroups:

1. insertion;
2. deletion;
3. substitution;

according to the kind of operation which led to a particular spelling mistake. The statistics on distribution of these subgroups is shown in Table 2.

Table 2. Distribution of OOV words belonging to the class "unintended sloppiness of user's input"

OOV subclass	hunspell		TreeTagger	
	% of total OOVs	% of unique OOVs	% of total OOVs	% of unique OOVs
insertion				
deletion				
substitution				
total				

Table 3. Distribution of OOV words belonging to the class "stylistic specifics of text genre"

OOV subclass	TreeTagger		hunspell	
	% of total OOVs	% of unique OOVs	% of total OOVs	% of unique OOVs
hashtags				
@-tokens				
links				
smileys				
slang				
total				

4 Text Normalization Procedure

4.1 Replacement of Twitter-Specific Phenomena

4.2 Restoration of Umlauts

4.3 Squeezing of Elongated Words

4.4 Translation of Slang Idioms

5 Evaluation

6 Conclusion

This article provided an overview of exisiting approaches to noisy text normalization task. Additionally, all mentioned methods were classified on the basis of two independent criteria. In section 3, we performed qulitative and quantitive analyses of out-of-vocabulary words in German tweets and suggested a set of ad-hoc techniques for mitigating their potential negative influence on natural language processing. This procedure allowed us to reduce the total OOV rate by ... % for `hunspell` and by ... % for `TreeTagger`.

Nevertheless, we should honestly admit that our system still has potential for development and research, since it mainly addresses only one of three main groups of OOV tokens. Future directions should certainly include a more thorough tackling of unintentional spelling mistakes and especially their most prominent types – deletions and substitutions.

Furthemore, a better evalution technique as well as comparison with other systems are needed for our normalization procedure. On the one hand, an *extrinsic* evaluation should be performed (cf. Sparck Jones and Galliers, 1996) which means that we not only have to show how the rate of OOV words goes down but much more how this lower OOV rate affects the work of the whole NLP system. On the other hand, we need to assess the quality of our procedure's work on the basis of metrics used by other researchers.

One possible estimation criterion which was used by Aw et al. (2006), Kaufmann (2010), Beaufort et al. (2010), and Oliva et al. (2013) is the BLEU score (Papineni et al., 2002). Another possibility would be to use the Word (WER) and Sentence Error Rates (SER) as suggested by Kobus et al. (2008). However, an obvious difficulty that we already encountered here is that both metrics highly rely on a subjective notion of the look of a "normalized" message. While the BLEU score could be an approriate criterion for normalization of messages like "i luv ma mather and wd do evrythin 4 her" which in fact looks like as if it were not English. But such highly distorted tweets are rather atypical for German. In this regard, WER and SER could be considered as more appropriate measurement criteria. But these metrics once again are rather dealing with spelling mistakes and would highly depend on whether one, for example, would consider the hash sign in a hashtag as an error.

Acknowledgement.

This work was financially supported by ... as part of collaborative project "...". The authors are also thankful to ... for help with the analysis of data.

References

Aw, A., Zhang, M., Xiao, J., Su, J.: A Phrase-based Statistical Model for SMS Text Normalization. COLING/ACL (2006) 33–40

Bangalore, S., Murdock, V., Riccardi, G.: Bootstrapping Bilingual Data using Consensus Translation for a Multilingual Instant Messaging System. COLING (2002) 33–40

Beaufort, R., Roekhaut, S., Cougnon, L. A., Fairon, C.: A Hybrid Rule/Model-Based Finite-State Framework for Normalizing SMS Messages. ACL (2010) 770–779

Brill, E., Moore, R. C.: An improved model for noisy channel spelling correction. ACL (2000) 286–293

Clark, A.: Pre-processing very noisy text. In Proceedings of Workshop on Shallow Processing of Large Corpora. (2003) 12–22

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., Basu A.: Investigation and Modeling of the Structure of Texting Language. International Journal of Document Analysis and Retrieval: Special Issue on Analytics of Noisy Text. **10** (2007) 157–174

Clark, E., Araki, K.: Text Normalization in Social Media: Progress, Problems and Applications for a Pre-processing System of Casual English. PACLING. Procedia - Social and Behavioral Sciences **27** (2011) 2–11

Cook, P., Stevenson, S.: An unsupervised model for text message normalization, Proceedings of the Workshop on Computational Approaches to Linguistic Creativity. CALC '09. (2009) 71–78

Dorsey, J.: "just setting up my twttr". `https://twitter.com/jack/status/20` Accessed February 26, 2013. (2006)

Han, B., Baldwin, T.: Lexical Normalization of Short Text Messages: Makn Sens a #twitter. ACL HLT. (2011) 368–378

Jurafsky, D., Martin, J. H.: Speech and Language Processing. 2nd Edition. Prentice Hall (2008) 129, 323

Kaufmann, M.: Syntactic normalization of twitter messages. The 8-th International Conference on Natural Language Processing. (2010)

Kobus, C., Yvon, F., Damnati, G.: Normalizing SMS: are Two Metaphors Better than One? COLING (2008) 441–448

Kukich, K.: Techniques for Automatically Correcting Words in Text. ACM Computing Surveys **24/4** (1992) 378–439

McVeigh, T.: Text messaging turns 20. The Observer (December 1, 2012)

Oliva, J., Serrano, J. I., and Del Castillo, M. D., and Igesias, .: A SMS normalization system integrating multiple grammatical resources. Natural Language Engineering. (2013) 121–141

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a Method for Automatic Evaluation of Machine Translation. ACL (2002) 311–318

Petersen, L. J.: Computer Programs for Detecting and Correcting Spelling Errors. Communications of the ACM **23/ 12** (1980) 676–687

Pianigiani, G., Donadio, R.: Twitter Has A New User: The Pope. The New York Times. Page A6. (December 4, 2012)

Sproat, R., Black, A. W., Chen, S. F., Kumar, S., Ostendorf, M., Richards, Ch.: Normalization of non-standard words. Computer Speech & Language, **15/3** (2001) 287–333

Schmid, H.: Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing. (1994)

Sparck Jones, K., Galliers, J. R.: Evaluating Natural Language Processing Systems. An Analysis and Review. Lecture Notes in Computer Science 1083, Springer. (1996)

Terdiman, D.: Report: Twitter hits half a billion tweets a day. `http://timmurphy.org/2009/07/22/line-spacing-in-latex-documents/`,Accessed February 26, 2013. (2012)

Toutanova, K., Moore, R. C.: Pronunciation Modeling for Improved Spelling Correction. ACL (2002) 144–151