Towards Shared Datasets for Normalization Research

Orphée De Clercq, Sarah Schulz, Bart Desmet and Véronique Hoste
LT^3, Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
firtname.lastname@ugent.be

Abstract

In this paper we present a Dutch and English dataset that can serve as a gold standard for evaluating text normalization approaches. With the combination of text messages, message board posts and tweets, these datasets represent a variety of user generated content. All data was manually normalized to their standard form using newly-developed guidelines. We perform automatic lexical normalization experiments on these datasets using statistical machine translation techniques. We focus on both the word and character level and find that we can improve the BLEU score with ca. 20% for both languages. In order for this user generated content data to be released publicly to the research community some issues first need to be resolved. These are discussed in closer detail by focussing on the current legislation and by investigating previous similar data collection projects. With this discussion we hope to shed some light on various difficulties researchers are facing when trying to share social media data.

Keywords: user generated content, text normalization, resource sharing

1. Introduction

In the current age, social media are omnipresent online and represent a rapidly evolving aspect of human communication to research. In NLP, working with user generated content (UGC) is becoming increasingly popular. Workshops – NLP4UGC at LREC2012 (Melero, 2012), LASM (Farzindar et al., 2013) and WASSA (Balahur and Montoyo, 2013) at NAACL2013 – shared tasks – SemEval2013 Task 2 (Kozareva et al., 2013) – and special journal issues – "Analysis of short texts on the web" (LRE 42:1) – or books (Moens et al., 2014) are devoted to this subject. The bulk of these studies focuses on issues such as performing opinion mining on UGC (Maynard et al., 2012) and on coping with the noise present in this type of data (Beaufort et al., 2010). Especially this latter topic is important because it serves as a basis for all other research performed on UGC. State-of-the-art NLP tools that are trained on standard data, reveal a crucial drop in performance when applied to noisy data (Ritter et al., 2011). Essentially, there are two ways to overcome this problem: the tools can either be adapted, i.e. domain adaptation (Daume, 2007) or the noisy text can be brought closer to a standard form, i.e. text normalization (Liu et al., 2011a; Han et al., 2013).

We have chosen to normalize Dutch and English data. Before performing this automatic step it is crucial to collect data coming from these social media and enrich these corpora with annotations. We describe our corpus collection and normalization efforts and present a basic set of experiments to illustrate the usefulness of such a corpus for normalization purposes. We show that by using a standard statistical machine translation system we can already improve the BLEU score with 20% for both languages. As in all scientific disciplines, research is only valuable when it can be reproduced by others working in the same field. This, however, brings up the issue of resource sharing. In this paper we particularly want to draw attention to the necessity of solving the therewith connected legal issues. This is complicated in the case of social media where the large companies all want their share and where legislation provides various restrictions.

The remainder of this paper is structured as follows. In Section 2 we describe our datasets and how these were annotated. The lexical normalization experiments that were performed using these data are presented in closer detail in Section 3. In Section 4 we introduce the problem of collecting and sharing social media resources. Section 5 concludes this paper and offers some prospects for future work.

2. Normalization Datasets

In order to fully represent the domain of UGC we decided to include three social media genres for each language: text messages (SMS), message board posts from a social networking site (SNS) and tweets (TWE).

For our Dutch corpus we sampled 1,000 text messages from the Flemish part of the SoNaR corpus (Treurniet et al., 2012), aiming at a balanced spread of two characteristics: age and region. In order to also include longer streams of UGC, 1,505 message board posts were randomly selected from the social networking site Netlog, which is popular amongst Belgian teenagers. In order to take into account the vast amount of normalization research done on Twitter data, we also included 248 randomly selected tweets and 600 tweets accompanying a popular Flemish TV show (tvvv[1]). For our English corpus, we followed a similar approach: 574 text messages were sampled from the NUS SMS corpus collected by Chen and Kan (2012) and Netlog posts and tweets were again randomly selected. Some data statistics of these corpora are presented in Table 1. These numbers show that the genres are not equally noisy. For both languages, the posts from social networking sites seem most noisy, i.e. we observe an increase in tokens after

[1]The Voice of Flanders

	DUTCH				ENGLISH			
	# items	before	after	%	# items	before	after	%
SMS	1000	14739	15364	4.2	574	10329	10539	2.0
SNS	1505	29986	31341	4.5	817	18245	18882	3.5
TWE	848	11500	11657	1.4	356	13273	13571	2.2

Table 1: Data statistics of our Dutch and English corpus representing the amount of data and the number of tokens before and after normalization. For all datasets the number of tokens increases after normalization. We present this increase in percentages.

normalization of 4.5% for Dutch and of 3.5% for English. What also draws the attention is that the Dutch tweets contain almost no noise, i.e. a token increase of only 1.4%.

All data was normalized to their standard Dutch or English form following normalization guidelines adapted to each language's characteristics (De Clercq et al., 2014). For Dutch, these guidelines have been drawn up in close collaboration with the developers of the Chatty Corpus (Kestemont et al., 2012) and for English findings from previous studies (Baron, 2003) were included. The guidelines can roughly be divided into two parts. The first part consists of the actual text normalization and comprises three steps: clearing all obvious tokenization problems, stating the different normalization operations and writing down the full normalized version. We allow four different operations: insertions, deletions, substitutions and transpositions. Examples of tokens requiring these operations are given below.

- INS: spoke (spoken), sis (sister)
- DEL: baaaaabyyyy (baby)
- SUB: iz (is), stoopid (stupid)
- TRANS: liek (like)

Insertions allow to indicate missing characters in a string. Deletions are used when characters should be deleted from a certain string. Substitutions are used when a character has been replaced with another similar one. Finally, transpositions are used when a combination of characters should be switched within one string. The second part of the guidelines consists of flagging additional information that might be useful for further automatic processing purposes. Within each utterance the annotators were asked to indicate the end of a thought (to account for missing punctuation), regional words, foreign words and named entities. They could also flag words that are ungrammatical, stressed, part of a compound, used as interjections or words that require consecutive normalization operations.

Using these guidelines, all data has been annotated by two annotators for each language. For Dutch, we were able to check the reliability by computing the word error rate (WER) between 1,000 text messages that were annotated by two linguists independently from each other. The resulting WER was 0.048 which reveals an almost perfect inter-annotator agreement.

3. Normalization Experiments

One well-known problem when dealing with UGC text is that the traditional NLP tools do not perform well on this type of text. The Stanford NER, for example, drops from 90.8% tot 45.88% when applied to tweets (Liu et al., 2011b). Also part-of-speech tagging, chunking and other techniques reveal a significant drop in performance when applied to UGC (Ritter et al., 2011). This is where our corpora come in as useful data for performing normalization experiments. Currently, there are three dominant approaches to transfer noisy into standard text: using spell-checking (Choudhury et al., 2007; Cook and Stevenson, 2009; Beaufort et al., 2010), speech recognition (Kobus et al., 2008) or machine translation techniques (AiTi et al., 2006; Liu et al., 2011a). For the present study we focus on applying statistical machine translation (SMT).

The underlying idea is to perceive normalization as a translation task from noisy text in one language to normalized standard text in the same language. Applying SMT to text normalization can be done at different levels of granularity. Previous work in this field has mostly focused on the word level (AiTi et al., 2006; Kobus et al., 2008; Raghunathan and Krawczyk, 2009). However, if we consider normalization, the task intuitively has a lot in common with transliteration tasks for which character-based SMT systems have proven adequate (Vilar et al., 2007). Pennell and Liu (2011) were the first to study character-based normalization. They, however, limited their approach by only focusing on abbreviations whereas we studied the added value by processing all the words on the character level for Dutch (De Clercq et al., 2013). For the present study, we focus on both Dutch and English. Important to note is that we only focus on lexical normalization, thus not resolving grammatical or other issues.

In the current set-up, two levels are compared: the token or word level and the character-unigram level. The aim of our experiments is to find out which level yields the best results. The assumption is that the different levels account for different normalization problems. The token level model is supposed to find frequent abbreviations (such as *lol* for *laughing out loud*) whereas the character-unigram model should allow us to generalize over the mapping of characters. This generalization yields the opportunity to correct frequently appearing normalization problems on the character level more effectively which is especially promising for UGC since there, problems like fusion of words or omission of word endings can be encountered.

The English ending *-ing* is for instance often realized as *-in* like in *goin*. Moreover, we also experiment with a cascaded approach which is a combination of the token and unigram-based models. i.e. in a first round we perform SMT at the token level and that output is then split into characters and fed to the character-unigram model.

	DUTCH		ENGLISH	
	# train	# test	# train	# test
SMS	6878	1109	7221	1110
SNS	5782	829	6997	1128
TWE	6833	1121	7069	1164

Table 2: Number of tokens in the train and test datasets for both languages that were used for the normalization experiments

For all experiments we use the Moses SMT-system (Koehn et al., 2007) and use equally large train and test datasets for both languages in order to achieve comparable results. See Table 2 for the exact number of tokens. As background corpus for our language models we use the *Corpus Gesproken Nederlands* (Oostdijk, 2000) for Dutch and the *British National Corpus* (Aston and Burnard, 1998) for English together with all available training data. All language models are built using KenLM (Heafield et al., 2013) since preliminary experiments revealed that this worked best for our type of data. For the token level model we build a language model with grams up to an order of 5 whereas for the character-unigram model we use up to an order of 10. We each time train on a combination of all three genres and test the performance for each genre individually and on their combination. To evaluate, we calculate the BLEU metric (Papineni et al., 2002) which has been specifically designed for measuring machine translation quality. It measures the n-gram overlap between the translation being evaluated and a set of target translations.

The results for both languages are presented in Table 3. The original, unnormalized data is used as a baseline (A). If all genres are combined we see that the BLEU score lies around 50. If we then look at our various translation models we see that the token level model (B) accounts for the highest improvements, i.e. around 25%, for both languages closely followed by the cascaded-unigram model (D), i.e. around 22%. Especially for English, we see that the character-unigram model does not perform well, an increase in BLEU of less than 10%. If we have a closer look at the testing on the individual genres we again notice that B and D perform best across all genres. This intuitively makes sense, because at the token level highly frequent normalization problems like reoccurring abbreviations will be resolved and if we look at the actual data we see that this is actually the case. The more frequent those problems appear in the data, the better the token approach works. After investigating the data we saw that this might also explain the very large increase in performance on the English SMS data using the token model, i.e. an increase of 91.5%. We noticed that this data is quite homogeneous with regards to the distribution of normalization problems. It is probable that if we would only train and test on this individual genre we would yield an even higher performance. If we analyze the robustness of our approach with respect to genre, we observe that for Dutch the same developments can be perceived across the various genres. The results on the English data are overall much lower, especially the SMS data seems to require many normalizations (BLEU of only 28.14 on the original data), if we look at the data statistics in Table 1, however, we see that many normalizations are not necessary (increase in tokens of only 2%) which was also confirmed after investigating the data. A result that also draws the attention is that on the English twitter data the unigram model seems to perform worse (-3.4%). A possible explanation for this could be a diversity of normalization problems in the TWE dataset. These findings for English led us to believe that for we might need to collect and annotate additional English UGC data.

The experiments presented here show the validity of using both the Dutch and English dataset for normalization purposes. Of course applying only SMT techniques to our data is not enough. Normalization is a much more complex task and an error analysis on our first Dutch experiments already revealed that errors remain due to abbreviation, phonetic and ortographic issues (De Clercq et al., 2013). This is why we are currently building a system that includes other modules besides machine translation. In order to solve the orthographic issues we are including a spell checker, whereas for the phonetic ones we are using a grapheme-to-phoneme converter.

4. Sharing UGC Data

We are convinced that the accessibility of normalization data for the research community is important in order to advance in the field of automatic social media processing. However, looking at the current legislation it seems extremely difficult to distribute these data without fearing legal repercussions. When collecting data from social media, it should be borne in mind that, in addition to the rules of copyright, other legislation may come into the picture, particularly in Europe. For this reason, a tendency can be observed to spread data underhandedly within the research community. As far as the availability of the data presented here is concerned we should bear in mind the following issues. In order to make this data inabusive of copyright laws it should first of all be anonymized, i.e. replace all named entities, URLs, etc. with special characters. This, however, is only a first step since the remainder of the process is dependent on many other factors. First, the data found in social network profiles may directly or indirectly identify natural persons. This will require permission from the individuals concerned. Secondly, the social network may – if it is established in the EU or has strong economic links with an EU Member State – qualify as a database protected by law. Thus, not only substantial extraction of data, but also insubstantial but frequent extractions may become unlawful. While EU law allows Member States to make exceptions in copyright, database and data protection

	DUTCH				ENGLISH			
	all genres	SMS	SNS	TWE	all genres	SMS	SNS	TWE
A. Baseline	51.81	47.45	51.25	57.06	46.50	28.14	49.79	56.38
B. Token level	65.04	60.77	62.72	71.58	59.30	53.90	56.70	66.21
	(25.5%)	*(28.1%)*	*(22.4%)*	*(25.4%)*	*(27.5%)*	*(91.5%)*	*(13.9%)*	*(17.4%)*
C. Character-unigram	61.27	57.34	61.61	65.30	50.98	47.38	50.10	54.48
	(18.3%)	*(20.8%)*	*(20.2%)*	*(14.4%)*	*(9.6%)*	*(68.4%)*	*(0.6%)*	*(-3.4%)*
D. Cascaded-unigram	63.48	60.76	62.16	67.50	56.95	51.56	54.85	63.41
	(22.5%)	*(28.1%)*	*(21.3%)*	*(18.3%)*	*(22.5%)*	*(83.2%)*	*(10.2%)*	*(12.5%)*

Table 3: Results of training on all data and testing on all data as well as the individual genres. The results are expressed in BLEU measure together with their improvement in percentages in between brackets.

legislation for scientific research, the boundaries of this exception are not clear, and differ across the EU. Moreover, the "terms" of use imposed by social networks are binding contracts, and most of them explicitly prohibit crawling or copying data. This even overrules the scientific research exception in most Member States.[2]

Currently, the two most popular genres to collect user generated content from are text messages (SMS) and tweets. With SMS, the largest problem is the private character of this type of data. Three notable SMS collection projects have, however, made this type of information available for research purposes. The sms4science project (Cougnon and François, 2011) is probably the most successful one. It was started up in Belgium to collect French text messages and over the years the same techniques have been carried out in other countries (Switzerland, France, Greece, Spain and Italy). The technique used for this was to lower the barrier for donation by letting people forward their messages directly to a central number free of charge The NUS SMS project (Chen and Kan, 2012) focused on collecting English and Mandarin text messages whereas the SoNaR project (Oostdijk et al., 2013) focussed on collecting Dutch text messages. These two latter approaches used the same technique to collect messages, namely by developing an application on the Google Android platform that allowed users to automatically send messages to the corpus.[3] Most of the time data can be acquired from these projects for research purposes when mentioning the source or after signing a license agreement. This is also valid for our SMS data which originates from two of the above-mentioned corpora.

Twitter data, on the other hand, is more difficult to release. Previous tweet collections, of which the Edinburgh Corpus (Petrović et al., 2010) is a well-known example, have been collected but are no longer available to the general public because Twitter changed its "terms" of use for crawling through the API during the summer of 2013. Following these new rules it now seems good practice to provide a third party with the user and message id and a download script that only downloads tweets from users that are still publicly available and that does not download tweets which have been removed by the user. This is for example how the data for the SemEval 2013 Task 2 (Kozareva et al., 2013) were spread to the research community. However, as soon as Twitter again changes its rules the data might become unavailable once more. Recently, Twitter has reacted to this problem by launching the pilot project *Twitter Data Grants*[4] where research institutions could submit a proposal to receive both public and historical Twitter data without violating their terms. No exact numbers on how many applications were submitted and granted are available at the time of writing but this is already a step in the right direction.

Closely related to Twitter is other data coming from popular social networking sites (SNS). The large companies such as Facebook and Google are known for not giving access to their valuable data. As far as the smaller companies are concerned, such as Netlog with which we worked together to get our data, these mostly ask you yo sign a non-disclosure agreement (NDA) which explicitly prohibits you to spread the data to other, third parties.

Based on these findings, we can state with certainty that all SMS data will be made publicly available for download at the LT3 website[5]. As far as the tweets and SNS data are concerned we are investigating the possibilities of releasing our data for research under terms satisfactory to all parties involved. Currently, the European Commission has also taken up a leading role in investigating these issues for copyright and database legislation, in order to foster non-commercial research in Europe which we hope will become clearer in the next few months.

5. Conclusion and Future Work

In this paper we have presented a Dutch and English corpus containing three genres of user generated content – text messages, message board post coming from social network sites and tweets – which have been manually normalized. Using social media data is a fast growing research topic in NLP and by performing experiments on these data we can achieve deeper insights which is important for advances

[2]For a more detailed overview of the current legislation we refer to Truyens and Van Eecke (2014).

[3]For a complete overview of techniques used to collect user generated content such as text messages we refer the reader to Treurniet et al. (2012).

[4]https://engineering.twitter.com/research/data-grants-closed

[5]www.lt3.ugent.be

in this field. We demonstrated the validity of our datasets for normalization purposes by applying statistical machine translation techniques. We found that for both languages we are able to improve the overall BLEU measure with up to 20%. However, the results of our normalization experiment also indicate that our English dataset, especially the SMS, might require some additional data collection and annotation.

After having studied the current legislation and other similar data collection projects we see that some legal issues remain which must be removed in order to support a praxis of shared data, which in turn leads to comparable results and faster progress in scientific research. For now we are only available to share our SMS dataset to the wide public without fearing legal repercussions and we hope that the same will become possible for our other datasets in the near future.

6. Acknowledgements

The work presented in this paper was carried out in the framework of the PARIS project which is part of the SBO Program of the IWT (IWT- SBO-Nr. 110067). We would like to thank the annotators, Maarten Truyens for his legal advice and the reviewers for their valuable comments.

7. References

AiTi, A., Min, Z., Juan, X., and Jian, S. (2006). A Phrase-based Statistical Model for SMS Text Normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney, Australia.

Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Balahur, A. van der Goot, E. and Montoyo, A. (2013). *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. ACL.

Baron, N. S. (2003). Language of the internet. *The Stanford Handbook for Language Engineers*, pages 59–127.

Beaufort, R., Roekhaut, S., Cougnon, L.-A., and Fairon, C. (2010). A hybrid rule/model-based finite-state framework for normalizing sms messages. In *Proceedings of ACL*, pages 770–779.

Chen, T. and Kan, M.-Y. (2012). Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, 47(2):299–335.

Choudhury, M., Saraf, R., Jain, V., Mukherjee, A., Sarkar, S., and Basu, A. (2007). Investigating and modeling the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157–174.

Cook, P. and Stevenson, S. (2009). An unsupervised model for text message normalization. In *Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*.

Cougnon, L.-A. and François, T. (2011). Étudier lécrit sms. un objectif du projet sms4science. In *La communication par SMS en Suisse. Usages et variétés linguistiques*. Linguistik Online (Themenheft).

Daume, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of ACL2007*, pages 256–263.

De Clercq, O., Schulz, S., Desmet, B., Lefever, E., and Hoste, V. (2013). Normalization of Dutch User-Generated Content. In *Proceedings of RANLP2013*.

De Clercq, O., Desmet, B., and Hoste, V. (2014). Guidelines for Normalizing Dutch and English User Generated Content. Technical Report LT3 Technical Report - LT3 14.01, Language and Translation Technology Team.

Farzindar, A., Gamon, M., Nagarajan, M., Inkpen, D., and Danescu-Niculescu-Mizil, C. (2013). *Proceedings of the NAACL Workshop on Language Analysis in Social Media*. The Association for Computational Linguistics.

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

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of ACL2013*, pages 690–696, Sofia, Bulgaria, August.

Kestemont, M., Peersman, C., De Decker, B., De Pauw, G., Luyckx, K., Morante, R., Vaassen, F., van de Loo, J., and Daelemans, W. (2012). The netlog corpus. a resource for the study of flemish dutch internet language. In *Proceedings of LREC2012*, Istanbul, Turkey, may.

Kobus, C., François, Y., and G., D. (2008). Normalizing SMS: are two metaphors better than one? In *Proceedings of Coling 2008*, pages 441–448, Manchester, UK.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Kozareva, Z., Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation. ACL*.

Liu, F., Weng, F., Wang, B., and Liu, Y. (2011a). Insertion, deletion or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proceedings of ACL2011*, pages 71–76, Portland, Oregon.

Liu, X., Zhang, S., Wei, F., and Zhou, M. (2011b). Recognizing named entities in tweets. In *Proceedings of ACL2011*, pages 359–367, Portland, Oregon.

Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. In *Proceedings of the LREC workshop: NLP can u tag #usergeneratedcontent?!*, Istanbul, Turkey.

Melero, M. (2012). *Proceedings of the LREC workshop: NLP can u tag #user_generated_content?!* European Language Resources Association (ELRA).

Moens, M.-F., Li, J., and Tat-Seng, C., editors. (2014). *Mining User Generated Content*. CRC Press - Francis Tailor Group.

Oostdijk, N., Reynaert, M., Hoste, V., and Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essen-*

tial Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing, Theory and Applications of Natural Language Processing, pages 219–247. Springer, New York.

Oostdijk, N. (2000). The spoken Dutch corpus. Outline and rst evaluation. In *Proceedings of LREC 2000*, pages 887–894, Athens, Greece.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL2002*, pages 311–318.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media 2010*.

Raghunathan, K. and Krawczyk, S. (2009). CS224N: Investigating SMS Text Normalization using Statistical Machine Translation. Technical report, Stanford University: Department of Computer Science.

Ritter, A., Clark, S., and Etzioni, M., O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of EMNLP 2011*, pages 1524–1534.

Treurniet, M., De Clercq, O., van den Heuvel, H., and Oostdijk, N. (2012). Collection of a Corpus of Dutch SMS. In *Proceedings of LREC 2012*, pages 2268–2273, Istanbul, Turkey.

Truyens, M. and Van Eecke, P. (2014). Legal aspects of text mining. In *Proceedings of LREC 2014*, Reykjavik, Iceland.