INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case

Anna Björk Nikulásdóttir, Jón Guðnason

Reykjavik University, Iceland
annabn@ru.is, jg@ru.is

Abstract

Text normalization is an important part of many natural language applications, in particular for text-to-speech systems. Text normalization poses special challenges for highly inflected languages since the correct morphological form for the normalization is not evident from the non-standard word, e.g. a digit.

In this paper we report on ongoing work on a text normalization system for Icelandic, a highly inflected North Germanic language. We describe experiments on the normalization of numbers and address the problem of choosing the correct morphological form of number names. We use language models trained on texts containing number names and inspect effects of different LMs on domain specific texts with a high ratio of digits. A partially class based LM, replacing number names with their part-of-speech tags, shows the best results in all domains. We further show that testing normalization on texts where number names have been converted to digits does not show representative results for texts originally containing digits: while a test set similar to the language model training data shows an error rate of 10.1% on inflected cardinals from 1-99, test sets originally containing digits show 45.3% and 55% error rates.
Index Terms: text normalization, inflected languages, Icelandic

1. Introduction

Normalization of non-standard words (NSWs) is the process of identifying NSW tokens in text, like digits and abbreviations, and converting them to their standard word representation. As an example, *5* would be converted to *five* in English text, and *etc.* to *et cetera*. Text normalization is an important part of many natural language applications, and is particularly important for text-to-speech systems (TTS) where a misread word can be perceived as a serious flaw in the system. In TTS, text normalization has the role of preparing input text for grapheme-to-phoneme conversion, as well as to deliver information on prosody, e.g. phrase brake prediction.

Even if text normalization has recently gained somewhat increased interest, especially in the context of normalizing social media texts, it can still be considered an "understudied area in NLP" [1]. With speech applications becoming outspread real-world applications that need to deal with all kinds of text input and still meeting high demands of users, high-accuracy text normalization systems are of outermost importance. As with general development in speech and language technology, low-resourced languages are lagging behind when it comes to text normalization. On the other hand, methods already developed for larger languages can be utilized to speed up development for new languages.

Two main challenges can be defined for text normalization: First, the identification and classification of NSWs in text. The simplest definition of a NSW is, that every word, not found in the dictionary of the processing system, should be treated as a non-standard word [2, 3]. Each NSW belongs to a *semiotic class*, a term established in the context of text normalization in [4], further developed to a taxonomy of NSWs in [2] and [5]. As noted in [5], the classes might not always be clear-cut and the taxonomy is unlikely to cover all possible cases.

Second, once a NSW has been classified, it has to be converted to its written-out standard word form. Depending on the semiotic class and the language in question, the NSW might have multiple possible verbalizations. The ambiguity can be semantic, as for the abbreviation *St.* in English ('Street' or 'Saint') or morphological. In highly inflected languages, verbalizations of common NSWs, like numbers and abbreviations, can have different morphological forms. When e.g. a cardinal number *2* has been identified and classified, the correct verbalization has to be derived from context. This is the case e.g. in Russian and other Slavic languages, and in Icelandic.

The disambiguation problem has been attacked using language models, see e.g. [1, 2]. However, to train a language model for this purpose, texts containing verbalized NSWs have to be available. One problem in the development of text normalization systems in general has been the lack of training and test data. Many NSWs are rarely found in text in their verbalized form, and it is unclear whether the context of verbalized forms of NSWs is similar enough to the context in which they occur in their non-standard word form. When approaching the normalization task with neural networks the data problem gets even larger, as very large amount of labeled data (NSWs labeled with their verbalized form in context) has to be constructed.

In this paper we report on ongoing work on a text normalization system for Icelandic TTS [6]. The aim is to develop a general normalization module, as the strived-for application of the TTS system is to read general web content[1]. Here, we focus on the problem of choosing the correct morphological form of number names when verbalizing cardinal and ordinal numbers, and in particular how different language models influence accuracy. Further, we will demonstrate results on different test sets, both domain specific test sets of different type than the language model training data, as well as on a similar type of text.

2. Related Work

While text normalization has traditionally been approached in an ad-hoc way for the application at hand, a systematic approach was established in [2]. They introduced a taxonomy of non-standard words (further expanded in [5]). This taxonomy defines eleven broad categories of NSWs: basic numbers, word-like tokens, dates and times, etc., each with concrete subcategories. The *basic numbers* category thus defines *cardinal numbers, ordinal numbers, number as digits*, and *decimal*.

[1]The Textahaukur text normalization system can be found at https://github.com/cadia-lvl/icelandic-textnorm

http://dx.doi.org/10.21437/Interspeech.2019-2367

Further development, introducing a two level approach to normalization, was introduced in [3]. The first level is concerned with identifying and classifying NSWs into semiotic classes, and can in many cases be done language and application independent. That way, a NSW *$2.50* can be labeled as being a *Money* expression, containing the symbol for USD, an integer part *2* and a fractional part *50*. The verbalization stage can be designed to verbalize the expression to *two dollar fifty* or *two dollars and fifty cents* or e.g. in German: *zwei Dollar fünfzig*. [7] exploits the method of using one covering grammar for different languages at the classification stage, finding that the performance differences need not be too large compared to a language-specific grammar.

When developing the systematic approach to text normalization, [2] inspected NSWs in four different text domains: news, real estate ads, communication on technical issues, and recipes. The NSWs in the texts were annotated in a semi-automatic way, using manual annotations for unclear or ambiguous cases. This corpus was used to train expansion models, resulting in mostly context independent algorithmic expanders. For ambiguous cases, domain dependent trigram language models were used to determine the most likely expansion. Due to the small size of the hand labeled corpus used for language model training, the improvement in error rates were not as large as one would expect when using a large training corpus. To somewhat decrease the data sparsity problem, partially class-based language models were used. The authors suggested that this approach used more extensively could prove useful.

Russian, and other Slavic languages, as well as Icelandic, have a complex number name system. In [1] a lightly supervised method to learn Russian number names is described. An overgenerating grammar is used to generate lattices of possible verbalizations of numbers. Combined with an n-gram language model, the most likely verbalization in context can be chosen. In contrast to [2], the training and test data are retrieved from texts originally containing verbalized versions of numbers. The real task, however, is to verbalize digits in texts. As [1] notes, the verbalized forms of numbers might not occur in the same way as digits in texts. They find, however, the frequency distribution of verbalized numbers and digits to be similar enough, such that the assumption can be made, that the results are reasonable representative for a real-world task.

Another research addressing the issue of text normalization in an inflected language is [8], where abbreviations are the core subject. The hypothesis is stated, that morphosyntactic tags can be predicted based on the morphosyntactic information of words in the context. Based on this hypothesis, a deep neural network is trained to infer morphosyntactic tags of abbreviations, and the tags then used to find the correct expansion form, using a lookup dictionary.

Context information is also used in [9] to learn expansions of abbreviations. In contrast to [8], the aim is to also deal with unknown abbreviations and to learn their expansions looking at common context of a potential abbreviation and its expansion.

Like in [8], recent research has been investigating neural networks in text normalization. One such approach was demonstrated in [10] and [11]. A DNN-based training of a normalization system relies on large amount of labeled data, available only through an established normalization system as described in [3]. Current results show that at the moment these established methods based on classification and verbalization grammars are more feasible for the development of new text normalization systems for low-resource languages (see also [12]). Apart from the training data problem, DNN-based systems have been found to make more severe errors than grammar based approaches. While grammar based systems might choose the wrong morphological form of a number name, an error in a system based on neural networks is more likely to show as a verbalization of a wrong number, thus changing the meaning of the text [10, 13]. Research on DNNs in text normalization will, however, certainly continue. An important contribution of the work in [10] to that development was to publish large labeled data sets for English and Russian text normalization, with a data set for Polish also currently available.

3. Inflected number names

3.1. The inflected number system in Icelandic

In Icelandic, nominals are inflected by case (nominative, accusative, dative, genitive), gender (masculine, feminine, neuter), and number (singular, plural). This applies to some number words as well: ordinals mostly follow the weak inflection paradigm of adjectives and the cardinals from 2 to 4 are inflected by case and gender, as do all numerals ending with those numbers (*24, 322, 10,373*)[2].

The ordinals *1.* and *2.* as well as the cardinal *1*, pose special challenges when it comes to verbalization. Not only do they have from 12 to 15 different forms (the cardinal 1 has both singular and plural forms, as do all ordinals), but their verbalized forms do not necessarily represent numerals at all. The word for 1, *einn*, can also be an adjective or a pronoun, and certain forms from the paradigm can be adverbs as well (*eins og* 'like', *eins konar* 'a kind of', *hver og einn* 'each and everyone', *svo fátt eitt sé nefnt* 'to name a few', etc.).

The verbalization of the ordinal *1.* is the only numeral that follows both the weak and strong inflection paradigm of adjectives. The verbalization for *2.* only follows the strong paradigm, and all other ordinals follow the weak paradigm only, resulting in three different forms for ordinals from *3.* onwards.

As with the cardinal 1, the first two ordinals also intervene with other word classes. For example, the superlative of *snemma* 'early' is *fyrst*, which belongs to the inflection paradigm of *fyrstur*, very commonly used in combinations like *sem fyrst* ('as soon as possible'). The verbalization of *2., annar* is also a common pronoun: *einhver annar* ('someone else').

3.2. Distribution of numbers in text

As noted in [1], digits and their verbalized forms might not appear in the same context in text. In Icelandic, as shown in 3.1, we additionally have to take the other word classes into account, the number word forms might represent.

In a part-of-speech (POS) tagged collection of texts on soccer, about 50% of the word forms from the inflection paradigm of *1 - einn* were found to be tagged as numerals, in general news text the ratio was even smaller, 43% tagged as numerals and 57% as pronouns or adjectives. All in all 77 different POS-tag strings were found for the forms of *einn*.

To further inspect the distribution of digits and number names in text, we computed word-embeddings for the cardinals from 1 to 10 in both forms using 30 million tokens of news text from the Icelandic Gigaword Corpus (IGC) [14]. We trained a CBOW model with 150-dimensional vectors and a window size of 7 [15, 16].

[2]Icelandic also has number adjectives for 2 to 4 denoting that the counted items are pairs or plural nouns. These forms are not included in the current experiments

When plotting the embeddings in two dimensions (dimensionality reduction being computed using the t-SNE algorithm [17]), clear groups can be identified, as shown in Figure 1: The digits build one cluster, as do the non-inflected number names from *fimm* ('five') to *tíu* ('ten'). Six clusters contain each one form of the number names from *tveir* ('two') to *fjórir* ('four'), each representing a certain combination of grammatical categories. The verbalization forms for *1 - einn* build an own cluster. This can be explained due to the various word classes these word forms can represent, meaning that they do not only occur in a number context.

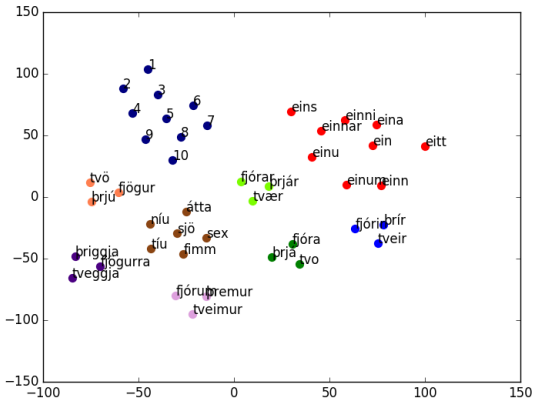

Figure 1: *Similarity clusters of cardinals from 1 to 10, as digits and numerals*

4. Text normalization system

The ongoing work on a text normalization system for Icelandic TTS follows the two level approach described in [3]. For the disambiguation of inflected number names, three different corpora for language model training were constructed.

4.1. Grammar

The normalization system contains a two level grammar for the verbalization of digits: a classification grammar and a verbalization grammar. This is an approach adapted from the Kestrel text normalization system [3], which has also proven a useful approach in developing text normalization for low-resource languages [12].

The classification grammar classifies input containing digits into several semiotic classes, a selection from [5]: basic numbers (cardinals, ordinals, decimals, numbers as digits), identifiers (telephone), dates and times, percentages, and temperatures. Further, abbreviations and acronyms are dealt with to some extent. The classification grammar is the same for all experimental setups.

There are two different versions of the verbalization grammar: a) a version containing actual verbalizations of the classified numbers, and b) a grammar where POS-tags are attached to the verbalizations of inflected numerals.

Both the classifier and the verbalization grammars are written for the Thrax Grammar Compiler ([18]). The core number mappings and expansion rules were adapted from the one-level text normalization grammar developed for the Icelandic Althingi ASR system ([19]), and extended to a two-level grammar.

4.2. Language Modeling

Three text collections with approx. 10 million tokens each were extracted from the Icelandic Gigaword Corpus (IGC). The IGC contains modern Icelandic texts from news papers, news sites on the web, etc., a total of about 1.3 billion tokens. It is annotated in TEI-format[3], POS-tagged and lemmatized[4].

As described in Section 3.1, some number names in Icelandic have homographs in other word classes. We therefore want to inspect if choosing sentences for a language model training corpus where it is ensured that each number name is tagged as a number (and not, e.g. as a pronoun), makes a difference compared to a corpus where no information on part-of-speech is available. Further, given the clustering based on grammatical categories (see Figure 1), a language model that utilizes POS-information in that the number names are replaced by their POS-tag string was to be trained, related to the partially class-based model as described in [2]. Thus, sentences were extracted from IGC as follows:

1. **Baseline:** each sentence contains a surface form corresponding to a number verbalization.
2. **Raw:** the same criteria as for the baseline, but an additional constraint is that each number name has to be POS-tagged as such.
3. **Raw+tags:** the same collection as *Raw*, where each number name is substituted by its POS-tag string.

A trigram language model with Witten-Bell smoothing was trained on each of the text collections, using the OpenGrm-based NGram library [18]

4.3. Normalization system

The text normalization system is based on the idea described in [3]. The Pynini Python library [20] is used for the compilation and processing of WFSTs. The open-source version of Kestrel, Sparrowhawk[5], was used as a reference for the implementation.

Input strings are compiled to FSTs and non-standard words classified using the FSTs compiled from the classifier grammar. The classified objects are again compiled into FSTs which are composed with the verbalization grammar FSTs. Finally, for ambiguous verbalizations, an intersection with the language model is performed, choosing the shortest path as the most likely verbalization.

5. Experiments

We tried three different system setups for the experiments, one for each language model described in Section 4.2 with the corresponding verbalization grammar (see Section 4.1)

The experiments aim at inspecting differences in normalization accuracy for inflected number names, depending on differences in the training data of the language models used for disambiguation. All language model training data, however, is based on occurrences of number names in text, whereas the real task of normalization is to verbalize digits, that might not occur in the same or similar context as verbalized number names.

[3] https://tei-c.org/

[4] The IGC is available at http://www.malfong.is/index.php?lang=enpg=rmh

[5] https://github.com/google/sparrowhawk

5.1. Test data

Different e.g. to [1], the main test data were not chosen as a portion of the language model training data. Instead, four domains were chosen from the IGC, all known to contain a high ratio of digits, and all the kind of texts a general TTS-system for the web should be able to read: a) general news texts, b) sports news, c) weather reports, and d) food recipes. The texts contain from ~2,500 to ~8,000 tokens each. They were normalized using the baseline system and then manually corrected to serve as a gold standard.

A fifth test set was extracted in the same way as the *Raw* LM training corpus, described in Section 4.2, from a different section of the IGC. The texts were from general news sites, but happened to contain a large portion of sports news. Number names were manually converted to digits to prevent erroneous converting of number name homographs.

6. Results

For all domains, the system setup using the *Tags+Raw* language model, i.e. where all number names are substituted with their POS-tags, shows the lowest error rate. Table 1 shows error rates of all number normalizations in each of the four text domains. Using POS-tag information to select number names as in the *Raw* model does show some benefits over the baseline, but shows slightly worse results for the weather domain (the weather domain in general also shows the least differences in error rates between the models).

Table 1: *Error rates in normalization of numbers in Icelandic text. The percentages describe the ratio of digit tokens containing a verbalization error.*

Domain	Baseline	Raw	Tags+Raw
NEWS	20.36%	18.56%	**15.14%**
SPORTS	27.46%	27.17%	**19.10%**
WEATHER	14.81%	15.64%	**14.40%**
RECIPES	15.67%	12.10%	**7.97%**

The errors are almost all due to wrong morphological forms of number names. There are also a few classification errors, like when the ambiguous NSW *500-1000* is classified as a telephone number instead of a *from-to* relation. The system seems to deal well with some categories of number names, e.g. numbers followed by *ára* ('years old'), denoting age: *32 ára - þrjátíu og tveggja ára* ('thirty two years old'). It is common in text to write out the ages of toddlers, e.g. *tveggja ára* ('two years old'), i.e. this context is represented in the language model. Other numbers, followed by common words like *prósent/prósenta/prósentum* ('percent'), or *milljarðar/milljarða/milljörðum* etc. ('billions') show errors in the verbalization, especially decimal numbers:

input: *43,5 prósent*
normalizer: *fjörutíu og* ***þrír*** *komma fimm prósent*
correct: *fjörutíu og* ***þrjú*** *komma fimm prósent*

Decimal numbers rarely occur as written-out number names in text, and thus it is not possible for the language model to learn these forms from standard text.

We compared the results of three test sets regarding normalization of inflected cardinals, i.e. cardinals from 1-4, from 21-24, etc. up to 91-94, using the *Tags+Raw* language model. Two sets reported on in Table 1, news and sports, and the general test set, containing similar data to that of the LM training, set were used for the comparison. As shown in Table 2, the error rate for the training data related test set (LM_SIMILAR) is only ~10%, while the results for the news and sports domain show 45% and 55% error rates respectively. The normalization of *1* and *2* is particularly difficult in the NEWS and SPORTS test sets, where it shows up to 100% error rate for *1* in the news domain. In the LM_SIMILAR set on the other hand, the normalization of *2* has the worst result, 14.5% error rate.

Table 2: *Error rates in normalization of digits from 1-4, 21-24, etc. up to 91-94. All experiments use the Tags+Raw LM. The percentages describe the ratio of digit tokens containing a verbalization error.*

Test set	All numbers	Inflected cardinals
NEWS	15.14%	45.3%
SPORTS	19.10%	55%
LM_SIMILAR	–	10.1%

7. Conclusions and future work

The goal of this work was to create a bootstrapping text normalization system for numbers in Icelandic using a general language model based on texts originally containing number names. With this system, large amount of text data were to be normalized, which then in turn could be used to train new language models based on texts originally containing digits.

The high error rates of the system on texts from several target domains, however, show that this general approach does not achieve sufficient results. The main problem is due to the many different forms of number names in Icelandic, especially for the digits 1 and 2, which also share homographs with other word classes. The context of the number names seems to be far too different from the context of digits in text, and thus a language model based on number names context does not generalize well enough to help choosing the correct form when verbalizing digits in context.

The next steps will include a more thorough analysis of the context of digits in the tested text domains. A more fine grained grammar will be developed, e.g. also taking standard words into account at the classification stage, like 'percent' and not only NSWs like '%'. Combining the refined grammars with larger, domain specific language models, domain specific texts will be normalized, which in turn will be added to the language model training data for further improvements of the system.

8. Acknowledgements

This work is part of the project No 160083-7001 *Environment for building text-to-speech synthesis for Icelandic* which is funded by The Icelandic Language Technology Fund.

9. References

[1] R. Sproat, "Lightly supervised learning of text normalization: Russian number names," in *IEEE Workshop on Spoken Language Technology (SLT)*, Berkeley, U.S.A., December 12-15 2010.

[2] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech & Language*, vol. 15, no. 3, pp. 287 – 333, 2001.

[3] P. Ebden and R. Sproat, "The Kestrel TTS text normalization system," *Natural Language Engineering*, vol. 21, pp. 333–353, 2015.

[4] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[5] D. van Esch and R. Sproat, "An expanded taxonomy of semiotic classes for text normalization," in *Proceedings of Interspeech*, Stockholm, Sweden, 2017, pp. 4016–4020.

[6] A. B. Nikulásdóttir, J. Guðnason, and E. Rögnvaldsson, "An Icelandic pronunciation dictionary for TTS," in *IEEE Workshop on Spoken Language Technology (SLT)*, Athens, Greece, December 18-21 2018.

[7] A. H. Ng, K. Gorman, and R. Sproat, "Minimally superviesed written-to-spoken text normalization," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, December 16-20 2017.

[8] P. Żelasko, "Expanding abbreviations in a strongly infletcted language: Are morphosyntactic tags sufficient?" in *Proceedings of LREC*, Myasaki, Japan, May 7-12 2018.

[9] B. Roark and R. Sproat, "Hippocratic abbreviation expansion," in *Proceedings of ACL*, Baltimore, U.S.A., June 23-25 2014, pp. 364–369.

[10] R. Sproat and N. Jaitly, "RNN approaches to text normalization: A challenge," *CoRR*, vol. abs/1611.00068, 2016.

[11] N. Jaitly and R. Sproat, "An RNN model of text normalization," in *Interspeech*, Stockholm, Sweden, August 20-24 2017.

[12] K. Sodimana, P. D. Silva, R. Sproat, A. Theeraphol, C. F. Li, A. Gutkin, S. Sarin, and K. Pipatsrisawat, "Text normalization for Bangla, Khmer, Nepali, Javanese, Sinhala, and Sundanese TTS systems," in *6th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, Gurugram, India, August 29-31 2018, pp. 147–151.

[13] K. Gorman and R. Sproat, "Minimally supervised number normalization," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 507–519, 2016. [Online]. Available: https://www.transacl.org/ojs/index.php/tacl/article/view/897/213

[14] S. Steingrímsson, S. Helgadóttir, E. Rögnvaldsson, S. Barkarson, and J. Guðnason, "Risamálheild: A very large Icelandic text corpus," in *Proceedings of LREC*, Miyazaki, Japan, May 7-12 2018.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013, pp. 3111–3119.

[17] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data usint t-sne," *Journal of Machine Learning Research*, vol. 9(Nov), pp. 2579–2605, 2008.

[18] B. Roark, R. Sproat, C. Allauzen, M. Riley, J. S. Sorensen, and T. Tai, "The OpenGrm open-source finite-state grammar software libraries," in *Proceedings of the ACL 2012 System Demonstrations*, July 2012, pp. 61–66.

[19] I. Helgadóttir, R. Kjaran, A. Nikulásdóttir, and J. Guðnason, "Building an ASR corpus using Althingi's parliamentary speeches," in *Proceedings of Interspeech*, Stockholm, Sweden, August 20-24 2017, pp. 2163–2167.

[20] K. Gorman, "Pynini: A Python library for weighted finite-state grammar compilation," in *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*, 2016, pp. 75–80.