Combining Phonology and Morphology for the Normalization of Historical Texts

Izaskun Etxeberria, Iñaki Alegria, Larraitz Uria
IXA taldea, UPV-EHU
{izaskun.etxeberria,i.alegria,larraitz.uria}@ehu.es

Mans Hulden
Department of Linguistics
University of Colorado
mans.hulden@colorado.edu

Abstract

This paper presents a proposal for the normalization of word-forms in historical texts. To perform this task, we extend our previous research on induction of phonology and adapt it to the task of normalization. In particular, we combine our earlier models with models for learning morphology (without additional supervision). The results are mixed: induction of the segmentation of morphemes fails to directly offer significant improvements while including known morpheme boundaries in standard texts do improve results.

1 Introduction and scenario

1.1 Normalization of historical documents

Historical documents are usually written in ancient languages which exhibit a number of differences in comparison with modern text, all of which have a significant impact on Natural Language Processing (NLP) (Piotrowski, 2012).

Carrying out a form of normalization before indexing historical texts makes it possible to perform queries against the text using standard (modern-day) words or lemmas and find their historical variants. This offers a method to make ancient documents more accessible to non-expert users. In addition, NLP tools developed for working with standard word forms perform better after normalization, in turn allowing for deeper processing such as information extraction for the identification of historical events.

1.2 The scenario

In this paper, we propose an approach for the normalization of historical texts. It is assumed that the corpus operated upon is digitized and that optical character recognition (OCR) has been carried out. A unique book—or a collection of them in case they are available from the same historical period or dialect—will be the processing unit. Under this scenario, long parallel texts are not available and statistical machine translation (SMT) approaches are therefore excluded.

For the normalization of historical texts, we develop an approach based on the induction of phonology and morphology. It is a lightly supervised model motivated by the need to achieve reasonable performance without requiring unrealistic amounts of manual annotation effort. In our previous work (Etxeberria et al., 2016) we have obtained good results using only induced phonological weighted finite state transducers (WFSTs). However, we have conjectured that additional lexicon and morphological paradigm information could serve to complement the phonological model (Beesley and Karttunen, 2003), and so we have sought to combine the two types of information in the normalization task. In this paper we present our work and results trying to demonstrate that additional lexical/morphological information can be advantageous in the normalization task.

In our setting the type of supervised data available is restricted to a limited number of annotated pairs of non-standard and standard (modern) word forms in a short piece of text. Availability of such annotations presumes an annotator with expertise in historical texts, but not necessarily in NLP.

2 Related work

Techniques for normalization can be roughly divided into two groups that take advantage of either. (1) hand-written morphophonological grammars and (2) machine-learning based techniques

Unsupervised techniques are also often used as a baseline for addressing the problem of normalization. Using edit-distance (Levenshtein dis-

tance) or a measure of phonetic distance (e.g. Soundex) are some of the more popular simple solutions.

In the realm of rule-based methods, Jurish (2010) compares a linguistically motivated context-sensitive rewrite rule-system with several unsupervised solutions in an information retrieval task in a corpus of historical German verse, reducing errors by over 60%.

Porta et al. (2013) presents a system for the analysis of Old Spanish word forms using weighted finite-state transducers.

Using machine learning techniques, Kestemont et al. (2010) documents a system that carries out lemmatization in a Middle Dutch literary corpus and presents a language-independent system that can 'learn' intra-lemma spelling variation.

Mann and Yarowsky (2001) presents a method for inducing translation lexicons based on transduction models of cognate pairs via bridge languages. Bilingual lexicons within language families are learned using probabilistic string edit distance models.

More recently, Scherrer and Erjavec (2015) presents a language-independent word normalization method which is tested on the problem of modernizing historical Slovene words. The method relies on supervised data and employs a model of character-level statistical machine translation (CSMT). Pettersson et al. (2014) also proposes a similar method and applies it to several languages.

As we want to obtain a morphological segmentation of variants, we have studied the state-of-the-art on unsupervised and semi-supervised morphology learning. Paradigms or morphological segmentations can be inferred from historical texts without supervision. Hammarström and Borin (2011) presents an interesting survey on unsupervised methods in morphology induction. *Morfessor* (Creutz and Lagus, 2002) is probably the most popular out-of-the-box tool for this task. (Bernhard, 2006) proposes an alternative solution to *Morfessor*.

In our previous work (Etxeberria et al., 2016) we have mainly used the *Phonetisaurus* tool,[1] a WFST-driven phonology tool (Novak et al., 2012) which is commonly used to map grapheme sequences to phoneme sequences under a noisy channel model. It is a solution that relies on some amount of supervision in order to achieve adequate performance, without however, requiring large amounts of manual development. We evaluated the system on the same corpus used in this paper using the usual parameters: precision, recall and F_1-score.

In the same paper we showed that the method works language-independently as we employed the same setup for both Spanish and Slovene and obtained similar or stronger results than that of previous systems reported by Porta et al. (2013) and Scherrer and Erjavec (2015). For Spanish our results are comparatively high, even with a small training set. For Slovene our method, without tuning, improves or equals the performance of the rest of the methods.

3 Corpus

As in our prior experiments, our main corpus is a 17th century literary work in Basque (*Gero*, written by Pedro Agerre "Axular" and published in 1643).

After a very simple process to clean up the noise in the corpus, 10% and 5% of the text was randomly selected for training and testing. Table 1 elaborates on the details of each slice.

Corpus	Tokens	OOVs	Types	OOVs
Training	8,223	1,931	3,025	1,032
Test	4,386	1,105	1,902	636

Table 1: Training and test corpora for Basque.

The training and test parts of the corpus were analyzed by a morphological analyzer of standard Basque. This way, words to be set aside for manual checking—e.g. out-of-vocabulary (OOV) items—were detected and after annotating these, a small parallel corpus was built.

The *BRAT* annotation tool (Stenetorp et al., 2012) was used for manual revision and annotation of the OOV words. Each OOV item was annotated as either "variation", "correct", or "other". For words in the first class, the corresponding standard word form was provided.

Finally, two lists of pairs (variant-standard) were obtained, one for training/tuning and the second one for testing. The test was carried out on the set of OOVs from the list.

[1] https://github.com/AdolfVonKleist/Phonetisaurus

4 Methods

4.1 Basic WFSTs

In order to learn the changes that occur within the selected word pairs, the previously mentioned *Phonetisaurus* tool was used. This tool is a WFST-driven grapheme-to-phoneme (g2p) framework suitable for rapid development of high quality g2p or p2g systems. It is a new alternative for such tasks; it is open-source, easy-to-use, and its authors report promising results (Novak et al., 2012). As the results obtained with this tool were the best ones in our previous research, we decided to focus only on using and improving our *Phonetisaurus*-based model for this task. In essence, we are leveraging a grapheme-to-phoneme tool in order to address the more general problem of word-form to word-form mappings.

After training a model with *Phonetisaurus*, a WFST is obtained which can be used to generate correspondences between previously unseen words and their matching standard forms. It is possible to change the number of transductions that the WFST returns for each input word and we have carried out a tuning process to choose the best value for this parameter.

Whenever we obtain multiple answers for a corresponding historical variant, some filtering becomes necessary. In our case, the answers that do not correspond to any accepted standard words are eliminated immediately. From among the rest of the words, the most probable answer (according to *Phonetisaurus*) is then selected.

To test if adding information about morpheme-boundaries helps in the task, our previous experiments in learning from word-pairs was complemented with a method of using word/morpheme-sequence pairs.

In our earlier approach, the tool was given complete plain word-form pairings to learn from. For example:

bekhaturik → *bekaturik*
emaiteak → *emateak*

In the augmented experiment, we use a different dictionary for generating training data. That is, we provide the morphological segmentation of the standard word instead of simply using the word itself. The result is the concatenation of the morphemes in their canonical forms:

bekhaturik → *bekatu+rik*
emaiteak → *eman+te+ak*

4.2 First extension: getting unsupervised morphological segmentation

Our hypothesis is that providing such morphological segmentations as given above together with morphological paradigms generated automatically from the original and annotated corpora could improve the previous results.

At this point, a problem is how to obtain the morpheme sequence of the corresponding historical forms as our morphological analyzer does not recognize historical variants found in the corpus. To address this, we have performed an automatic segmentation of the data using the *Morfessor* tool (Creutz and Lagus, 2005).

Morfessor is a program that takes as input a corpus of unannotated text and produces a segmentation of the word forms observed in the text. It is a state-of-the-art tool, language independent, and the number of segments per word is not restricted as in other existing morphology learning models.

After the tuning phase (using standard Basque) we input the entire historical corpus to *Morfessor*. Using this text, *Morfessor* creates a model which is then used to obtain the segmentation of any word forms annotated in the corpus. This way, we can produce a new dictionary for *Phonetisaurus* consisting of segmented pairs of historical/standard forms. Following the previous example, the output would be:

bekha+turik → *bekatu+rik*
emai+te+ak → *eman+te+ak*

4.3 Second extension: morphological inference from the parallel corpus

Another alternative approach to expanding the training data is to identify new lemmas and affixes among the historical forms by taking advantage of the (limited) parallel entries. For example, from the entries

bertzetik → *beste+tik*
dadukanak → *dauka+n+ak*
beranduraiño → *berandu+raino*

it can be inferred that *bertze*/*beste* and *daduka*/*dauka* are equivalent lemmas and *raiño*/*raino* equivalent suffixes.

With such equivalences, we built, using the finite-state tool *foma* (Hulden, 2009), an enhanced morphological analyzer that recognizes, in addition to the standard Basque, historical variants, including the identified new morphemes and also links the variants to the corresponding stan-

dard word-forms. With such an enhanced analyzer previously unseen historical words can be identified and linked to the corresponding standard word-form. Considering the previous example *bertzeraiño* and *dadukanetik* (non-standard forms); these can now be analyzed because the non-standard lemmas (*bertze* and *daduka*) and non-standard suffixes (*raiño*) are recognized by the new analyzer.

Because of possible noise in the data we use a threshold of two for the minimum number of times a morpheme/affix needs to be seen before it is included in the new analyzer.[2] As the resulting analyzer is strong on precision (98.17%) but weak on recall (37.99%), we combine it with the first WFST in a hierarchical way: by first applying the enhanced analyzer, and that failing to give results, passing the word on to the WFST from the first experiment.

5 Evaluation

We evaluated the quality of the different approaches using the standard measures of precision, recall and F_1-score. We have also analyzed how the different options in each approach affect the results.

The baseline for our experiments is a simple method based on using a dictionary of equivalent words with the list of word pairs learned. This approach involves simply memorizing all the distinct word pairs of historical and the standard forms, and subsequently applying this knowledge during the evaluation task.

5.1 Results

The first three different runs corresponding to the three possible representations were tuned using cross-validation and increasing the number of retrieved answers (5, 10, 20 or 30). Retrieving more answers yields a better F_1-score in the WFST model until an upper limit is reached. 20 answers were selected for the last two experiments and 5 for the first. After tuning, a new evaluation was carried out using the test corpus (shown in Table 2).

The results for the model that uses the morphological segmentation are slightly worse than the ones obtained using only the phonological induction (full word-form pairs) from the parallel corpus, but they are quite close. When the enhanced morphological analyzer is applied before the word/word WFST a slight improvement is seen. We believe that if we were able to improve the quality of the inferred morphological segmentation the overall results could also be improved.

System	Prec.	Recall	F-score
Baseline	**94.87**	39.22	55.50
Word/word	91.53	78.27	84.34
Word/morph	91.08	77.56	83.78
Morph/morph	90.68	75.62	82.47
Supervis. morph & and word/word	91.94	**78.62**	**84.76**

Table 2: Results on the test corpus for the baseline and the four proposed systems

5.2 Combination and Oracle

In order to detect if the behaviors of the two systems are complementary we looked for words that were well normalized in only one system, as in the following words: *arintkiago(arinkiago)*, *autsikizetik(ausikitzetik)*, *baillezakete(bailezakete)*, *bereganik(beregandik)*, *dathorrenean(datorrenean)*, *etzedilla(ez_zedila)*, *fintkiago(finkiago)*, *lothu(lotu)*, *zeikan(zitzaion)* and *zuetzaz(zuetaz)* are correctly normalized by the first system (word/word); *baiteraku(baitigu)*, *erraxten(errazten)*, *fariseoek(fariseuek)*, *hilzaileak(hiltzaileak)* and *lekhukok(lekukok)* only by the second (word/morph); and *ezterauet(ez_diet)*, *konsideratzeak(kontsideratzeak)* and *malizia(malezia)* by the third (morph/morph).[3]

Due to this complementarity we decided to combine the first three systems. In a first (simple) attempt we applied a voting system: if two systems offer the same proposed output, we choose that, else we choose the output of the first system. This yields a slight improvement.

We also calculated an oracle score using the same three systems—i.e. hypothetically always picking the best output. While we observe that a simple voting system improves slightly over the single-answer methods, examining the oracle re-

[2] Better single results were obtained using the threshold only for the suffixes, but the best combination is obtained using the threshold for both suffixes and lemmas

[3] It may also be observed that some non-phonological cases of variation (i.e. zeikan/zitzaion) can be solved by the first system which does not use morphological information

sults (table 3), we conclude that there is indeed room for improvement.

System	Prec.	Recall	F-score
Voting	91.94	78.62	84.76
Oracle	95.48	82.16	88.32

Table 3: Results on the test corpus.

6 Conclusions and future work

We have extended previous work on normalization of historical texts and tested the new methods against 17th century literary work in Basque.

Some extensions for taking advantage of morphological information have been proposed; this includes using morphological segmentation as a source of information as well as expanding a morphological analyzer. The results are somewhat limited because segmentation of morphemes only improves the results slightly over a purely phonological model.

We expect to further develop and test these techniques on more languages and corpora (additional historical texts in Basque and Spanish in a first step).

In the near future, our aim is to improve the results by taking advantage of a more precise and wider morphological segmentation and to attempt to combine the various models in a more effective way. Based on the oracle results we surmise that there is much opportunity for improvement.

Acknowledgments

The research leading to these results was carried out as part of the TADEEP project (Spanish Ministry of Economy and Competitiveness, TIN2015-70214-P, with FEDER funding) and the ELKA-ROLA project (Basque Government funding).

References

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*.

Delphine Bernhard. 2006. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of the Pascal Challenges Workshop on the Unsupervised Segmentation of Words into Morphemes*, pages 19–23.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pages 21–30. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Izaskun Etxeberria, Iñaki Alegria, Larraitz Uria, and Mans Hulden. 2016. Evaluating the noisy channel model for the normalization of historical texts: Basque, spanish and slovene. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Bryan Jurish. 2010. Comparing canonicalizations of historical German text. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 72–77. Association for Computational Linguistics.

M. Kestemont, W. Daelemans, and G. De Pauw. 2010. Weigh your words—memory-based lemmatization for Middle Dutch. *Literary and Linguistic Computing*, 25(3):287–301.

Gideon S Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, NAACL '01, pages 1–8. Association for Computational Linguistics.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, Donostia–San Sebastian, July. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. *Proceedings of LaTeCH*, pages 32–41.

Michael Piotrowski. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 5(2):1–157.

Jordi Porta, José-Luis Sancho, and Javier Gómez. 2013. Edit transducers for spelling variation in old Spanish. In *Proc. of the workshop on computational historical linguistics at NODALIDA 2013. NEALT Proc. Series*, volume 18, pages 70–79.

Yves Scherrer and Tomaž Erjavec. 2015. Modernising historical Slovene words. *Natural Language Engineering*, pages 1–25.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.