Evaluating the Noisy Channel Model for the Normalization of Historical Texts: Basque, Spanish and Slovene

Izaskun Etxeberria, Iñaki Alegria, Larraitz Uria, Mans Hulden
University of the Basque Country, University of Colorado
{izaskun.etxeberria i.alegria larraitz.uria}@ehu.eus, mans.hulden@colorado.edu

Abstract

This paper presents a method for the normalization of historical texts using a combination of weighted finite-state transducers and language models. We have extended our previous work on the normalization of dialectal texts and tested the method against a 17th century literary work in Basque. This preprocessed corpus is made available in the LREC repository. The performance of this (semi-)supervised method for learning relations between historical and contemporary word forms is evaluated against resources in three languages. The method we present learns to map phonological changes using a noisy channel model; it is a solution that uses a limited amount of supervision in order to achieve adequate performance without the need of an unrealistic amount of manual effort. The model is based on techniques commonly used for phonological inference and producing Grapheme-to-Grapheme conversion systems encoded as weighted transducers and produces F-scores above 80% in the task for Basque. A wider evaluation shows that the approach performs equally well with all the languages in our evaluation suite: Basque, Spanish and Slovene. A comparison against other methods that address the same task is also provided.

Keywords: computational morphology, phonology, lexical normalization, historical texts, digital libraries

1. Introduction and scenario

Historical documents are usually written in ancient languages that exhibit a number of differences in comparison to modern languages, all of which have a significant impact on Natural Language Processing (NLP) (Piotrowski, 2012). Both historical and dialectal texts present similar problems from an NLP point of view in that NLP tools developed for contemporary standard language often fail in handling the linguistic varieties encountered in such texts.

A majority of NLP tools are designed to process newspaper texts written in contemporary language. The processing of standardized modern languages exhibits some characteristics that are not found in historical and dialectal corpora. For example, (i) a *standard variant* is used for writing communication which is well documented by dictionaries and grammars; (ii) these languages have *standard orthographies* and the majority of published texts adhere to these orthographic norms; (iii) large amounts of text are electronically available and can be used for developing NLP tools and resources. On the contrary, most of these characteristics are not shared by historical and dialectal text resources and, therefore, standard NLP tools can often not be directly applied to such corpora.

Traditionally, some degree of lexical normalization is performed when working with historical and dialectal texts in order to link each variant to its corresponding standard form. Once the texts are normalized, standard NLP and IR (Information Retrieval) tools can be applied to the corpora with reasonably high performance. *Canonicalization* is the term used in this area, a term which referring to mapping each non-standard variant to a canonical one (Jurish, 2010).

Accurate normalization can be very useful: by carrying out such normalization before indexing historical texts, for example, it is possible to perform queries against texts using standard words or lemmata and find their historical counterparts. Normalization has the potential to make ancient documents more accessible for non-expert users.

NLP tools for standard languages also work better after normalization, which in turn allows for subsequent deeper processing to be carried out, e.g. information extraction for the purpose of identification of historical events and other applications.

In this paper, we propose and evaluate an approach based on a model that is often used for similar tasks such as the induction of phonology and learning grapheme-to-phoneme conversion models.

Our working hypothesis is that, as in the case of dialectal variants, the differences between ancient and current standard Basque seem to be mainly phonological, and therefore, we have reapplied the best method used in our previous work with dialects (Etxeberria et al., 2014). This method uses *Phonetisaurus*,[1] a Weighted Finite State Transducer (WFST) driven phonology tool (Novak et al., 2012) which learns to map phonological changes using a noisy channel model. It is a solution that uses a limited amount of supervision in order to achieve adequate performance without the need of an unrealistic amount of manual effort. This technique has been also performed for normalization of non-standard texts in social media (Alegria et al., 2013).

Experiments for Basque were carried out using a corpus of old Basque (Section 3 in this article). In order to compare our results with the systems used for Spanish (Porta et al., 2013) and for Slovene (Scherrer and Erjavec, 2015), we have also evaluated a supervised learning method. Using these resources, we can examine how performance varies with the size of the supervised corpus.

2. Related work

The foremost techniques currently used for the normalization or canonicalization of historical texts can be roughly divided into three groups:

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

- Rule-based methods (hand-written phonological grammars) are the most habitual solution; however, these techniques do not fit into our scenario because of the amount of manual work required.
- Machine-learning based techniques: systems that learn from examples of standard-variant pairs. These are our primary concern in this paper.
- Unsupervised techniques: systems that work without supervision. Applying edit-distance (Levenshtein distance) or phonetic distance (by i.e. the *Soundex* algorithm) are popular solutions. Such approaches are often used as a baseline for testing new systems (Jurish, 2010).

2.1. Rule-based methods

Most of the systems found in the literature report handwritten phonological rules which are compiled into finite-state transducers.

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

Porta et al. (2013) present a system for the analysis of Old Spanish word forms using rules compiled into weighted finite-state transducers. The system makes use of previously existing resources such as a modern lexicon, a phonological transcription system and a set of rules that model the evolution of the Spanish language from the Middle Ages onward. The results obtained in all datasets show significant improvements, both in accuracy and in the trade-off between precision and recall with respect to the baseline and the Levenshtein edit distance.

2.2. Learning phonological changes

Kestemont et al. (2010) carries out lemmatization in a Middle Dutch literary corpus, presenting a language-independent system that can 'learn' intra-lemma spelling variation. This work employs a novel string distance metric to better detect spelling variants. The semi-supervised system attempts to re-rank candidates suggested by the classic Levenshtein distance, leading to substantial gains in lemmatization accuracy.

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

Inspired by that paper, Scherrer (2007) makes use of a generate-and-filter approach quite similar to the method we initially used for phonological induction on dialectal corpora (Etxeberria et al., 2014; Hulden et al., 2011). In this previous work we tested two approaches:

- Based on the work by Almeida et al. (2010), differences between substrings in distinct word-pairs are obtained and phonological rules are learned in the format of so-called phonological replacement rules (Beesley and Karttunen, 2003; Hulden, 2009) transformation patterns. These rules are then applied to novel words in the evaluation corpus. To prevent overgeneration, the output of the learning process is later subject to a morphological filter where only actual standard-form outputs are retained.
- An Inductive Logic Programming-style (ILP) (Muggleton and De Raedt, 1994) learning algorithm where phonological transformation rules are learned from word-pairs. The goal is to find a minimal set of transformation rules that is both necessary and sufficient to be compatible with the learning data, i.e. the word pairs seen in the training data.

More recently, Scherrer and Erjavec (2015) have developed a language-independent word normalization method and tested it on a task of modernizing historical Slovene words. Their method relies on supervised data, and employs a model of character-level statistical machine translation (CSMT), using only shallow knowledge. Pettersson (2016) proposes a similar method and applies it on several languages.

In the following, we compare our results with those reported by Porta et al. (2013) for Spanish and Scherrer and Erjavec (2015) for Slovene.

3. Basque historical corpus: annotation and experimental set-up

In order to test the applicability of the noisy channel method, we chose the classical book *Gero*, written by Pedro Agerre "Axular" and published in 1643. Several reasons led us to choose this particular work, most important of which are that (i) it is a classical Basque work, (ii) it is old enough (from the 17th century), (iii) it is not too short (around 100,000 words) and (iv) a digitized version is readily available (at the `www.armiarma.com` website).

After an initial cleaning of the noise in the corpus, the corpus was divided into three parts, each containing 85%, 10% and 5% of the text.

The unit used to make the division was a paragraph, and paragraphs were randomly selected to obtain the splits described. Following this, a small parallel corpus of historical and standard Basque was built semi-manually for training and tuning (from the part containing 10%, *Gero_10*) and other one for testing (from the part containing 5%, *Gero_5*). The *Gero_10* and *Gero_5* parts of the corpus were analyzed by the morphological analyzer of standard Basque. This way, words to be set aside for manual checking—i.e. Out Of Vocabulary (OOV) items—were detected and after annotating these, a small parallel corpus was built.

In the two files *Gero_10* and *Gero_5* each paragraph was divided into sentences and a text categorization tool named "*TextCat*" was used to determine the language of each sentence in order to get rid of those citations written in latin.

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.

Corpus	Tokens	OOVs	Word-forms	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.

Finally, two lists of pairs (variant-standard) were obtained, one for training/tuning and the second one for testing. Figures can be consulted in Table 1. The test was carried out on the set of OOVs from the list.

4. Training Process and Results

Once the *Gero_10* and *Gero_5* files were annotated, we applied the methods described below for learning phonological changes. The number of different word pairs (word-list) corresponding to the OOVs were 956 for learning and 566 for testing.[2] To avoid having to account for possible case mismatches, only lower case letters were used for all word pairs.

As discussd above, we apply the strongest method found in our previous work with dialect normalization (Etxeberria et al., 2014). This approach uses *Phonetisaurus*, a Weighted Finite State Transducer (WFST) driven phonology tool (Novak et al., 2012), based on OpenFST (Allauzen et al., 2007), which learns mapping of phonological changes using a noisy channel model.

After data preparation, where we collect pairs into a dictionary, the application of the tool includes three major steps:

1. Sequence alignment. The alignment algorithm is based on the algorithm proposed in Jiampojamarn et al. (2007) and includes some minor modifications to it.

2. Model training. An n-gram language model is trained using the aligned data and then converted into a WFST. For producing the language model, we used the Language Model training toolkit *NGramLibrary* for our experiments, although several alternative similar tools exist that all cooperate with *Phonetisaurus*: *mitlm, NGramLibrary,* SRILM, SRILM *MaxEnt extension, CMU-Cambridge SLM*.

3. Decoding. The default decoder used in the WFST-based approach finds the best hypothesis for the input words given the WFST obtained in the previous step. It is also possible to extract the k-best output hypotheses for each word.

In practice, the application of the tool is straightforward. We have used the *Phonetisaurus* tool to learn the changes that occur within the selected word pairs, which by itself produces a grapheme-to-grapheme system. In our case, the data consist of a dictionary that contains the 956 word pairs to learn in the *Gero_10* file.

Once this model is trained and converted to a WFST format, it can be used to generate correspondences between previously unseen words and modern standard forms, (i.e. for the 566 words in the test set obtained from the *Gero_5* file). The WFST model provides the possibility of retrieving multiple candidate transductions for each input word and we carried out a tuning process to choose the best value for the number of candidates to generate.

When multiple possibilities for a corresponding historical variant exist, some filtering becomes necessary. The first filter is obvious: the transductions that do not correspond to any accepted standard word form are eliminated. For selecting standard words a morphological analyzer of Basque was used (Alegria et al., 2009). From the remaining candidates, the most probable transduction according to *Phonetisaurus*'s weight model is selected.

We measured the quality of the different approaches using the usual parameters: precision, recall and the harmonic combination of the two, the F_1-score, and we analyzed how the different options in each approach affect the results.

System	Precision	Recall	F-score
Baseline (memory based)	0.9487	0.3922	0.5550
Phonetisaurus1, using only variants for training	0.9153	0.7827	0.8438
Phonetisaurus2, using also identical pairs for training	0.9184	0.7951	0.8523

Table 2: Results for Basque. Precision, Recall and F-score.

The baseline of our experiments has been a simple method based on a dictionary of equivalent words learned from the training data. This entails simply memorizing all the distinct word pairs detected among the historical and standard forms and subsequently applying this knowledge during the evaluation task. As expected, the precision is high (94.87%): when the baseline gives an answer it is usually the correct one, as it is the same it has seen before. But the recall of the baseline is low (39.22%) and consequently the F-score too (55.50%), as is expected: less than half of the words in the evaluation corpus have been encountered before.

After the tuning process using cross-validation in the development corpus (we asked *Phonetisaurus* to increase the number of retrieved answers ranging from: 5, 10, 20 or 30, where 5 produced the strongest result), the system was evaluated against the test corpus.

The main results are show in Table 2. In addition to results using the baseline approach two systems were trained using *Phonetisaurus*:

- learning only from pairs corresponding to OOVs (pairs where the historical forms and normalized forms are different)
- learning from all the pairs, including pairs where historical form and standard one are the same

[2] A set including only pairs labeled as “Variation” and “Correct” from list of OOVs (5th column in Table 1).

The results are similar to those we achieve in previous works on dialectal corpora.

5. Corpus and evaluation for Spanish and Slovene

We deemed the results of this language-independent system to be strong enough to warrant further experiments with other languages and corpora.

5.1. Spanish

Our first comparison is with the results reported by Porta et al. (2013) for the normalization of Old Spanish. For this, we used the FL-EM DATASET used in that paper. As the authors of the paper note “FL-EM basically corresponds to the lexicon found within the FreeLing distribution for analysing Old Spanish”.

The corpus was kindly provided by the cited authors and, after preprocessing, a list of 31,046 word pairs (old word – standard word) was obtained. Half of these were stored for testing and the other half for training (no new tuning was carried out). The *FreeLing* suite (Carreras et al., 2004) was used for filtering the proposals.

System	Precision	Recall	F-score
Phonet., 200 examples	0.9648	0.7972	0.8730
Phonet., 500 examples	0.9650	0.8672	0.9135
Phonet., 1000 examples	0.9664	0.8883	0.9257
Phonet., 5000 examples	0.9661	0.9225	0.9438
Phonet., all (15,523)	0.9662	0.9458	0.9559
Porta et al. (2013)	0.6975	0.8902	0.7822

Table 3: Results for Spanish. Precision, Recall and F-score. The three first figures are the averages obtained using different samples of the relevant size. All the results pertain to the same test set.

Due to the relatively big size of the corpus we were able to test the results using increasingly larger slices of the corpus for training (200, 500, 1000...), documented in the results below.

The accuracy results (given in Table 3) are comparatively high, even with a small training set; the method outperforms those reported in Porta et al. (2013), particularly precision.

Due to the fact that almost all the words in the corpora are OOVs these results are quite comparable to those obtained for the Basque. In the Spanish case, the results are even better despite the use of a smaller corpus for training and without new tuning, leading us to conclude that the task is easier.

5.2. Slovene

For Slovene, a similar process was carried out. The dataset from Scherrer and Erjavec (2015) consists of a training (*goo* corpus) and a testing lexicon (*foo* corpus) of historical Slovene as well as a frequency-annotated reference word list of modern Slovene, kindly provided to us by the authors.

Both corpora (training and test) are split into three parts (Scherrer and Erjavec, 2015):

- *18B* Texts from the second half of the 18th century, all written in the Bohorič alphabet;
- *19A* Texts from the first half of the 19th century, written in the Bohorič alphabet;
- *19B* Texts from the second half of the 19th century, written in the Gaj alphabet.

The sizes of the corpora can be consulted in Table 4.

Corpus	Unique words	Identical pairs
Training		
Goo 18B	6,494	1,181 (17.8%)
Goo 19A	11,352	2,755 (23.8%)
Goo 19B	27,252	19,635 (70.1%)
Test		
Foo 18B	4,641	340 (7.1%)
Foo 19A	5,801	890 (15.1%)
Foo 19B	10,470	8,120 (76.1%)

Table 4: Corpus for Slovene.

Using these subsets for training and testing and the same experimental setup as used in the original experiments, we obtained the results shown below in Table 5. It is worth pointing out that no new tuning was carried out and that all the pairs in the training corpus, including identical pairs, were used during training.

In contrast to the test corpus for Basque and Spanish this corpus contained far more identical pairs.

System	Accuracy
18B Phonetisaurus, always responding	0.674
18B Scherrer and Erjavec (2015), without filtering	0.614
18B Scherrer and Erjavec (2015), filtered	**0.678**
19A Phonetisaurus, always responding	**0.794**
19A Scherrer and Erjavec (2015), without filtering	0.747
19A Scherrer and Erjavec (2015), filtered	0.784
19B Phonetisaurus, always responding	**0.868**
19B Scherrer and Erjavec (2015), without filtering	0.866
19B Scherrer and Erjavec (2015), filtered	0.846

Table 5: Comparing the results for Slovene. Accuracy for the three subsets.

To compare the results, we added an extra step to our system: when none of the proposals from the WFST was found in the list of correct words, the first proposal was used. In this way, some normalization candidate is always obtained and accuracy can be calculated.

Table 5 compares the results: our method, without tuning, improves (in the case of the 19A subset) or equals (in the case of the 18B and 19B subsets) the performance of the rest of the methods.

For filtering the proposals we have used the same word-list used in Scherrer and Erjavec (2015).[3] We note that the coverage of this filter is more limited than that of the analyzers used in conjunction with the Basque and Spanish experiments; this may partially explain the comparative weakness of these results.

6. Conclusions and future work

We have extended our previous work on the normalization of dialectal texts and tested the method against a 17th century literary work in Basque. This preprocessed corpus is made available in the LREC repository.[4]

Our phonological induction inspired method has been evaluated and produces F-scores above 80% in the task. It is a solution that uses a limited amount of supervision in order to achieve adequate performance without the need of significant manual annotation efforts or experts for writing rules.

To assess the performance and language-independence of the method, training and evaluation was carried out using Spanish and Slovene historical text corpora. The results are similar or better to those reported in the bibliography even though no new tuning process was performed.

In the near future, our goal is to try to improve upon the performance of the system by taking advantage of additional morphological information (morphemes and partial paradigms) that can be inferred from the corpora.

7. Acknowledgments

We are indebted to Josef Novak for his useful help concerning Phonetisaurus, and to Jordi Porta, Yves Scherrer and Tomaž Erjavec for providing the corpora of Spanish and Slovene and for addressing our inquiries about their experiments. We also want to thank the anonymous referees for helpful comments.

The research leading to these results was carried out as part of the TACARDI and TADEEP projects (Spanish Ministry of Economy and Competitiveness, TIN2012-38523-C02-011 and TIN2015-70214-P, with FEDER funding) and the ELKAROLA project (Basque Government funding).

References

Alegria, I., Etxeberria, I., Hulden, M., and Maritxalar, M. (2009). Porting Basque morphological grammars to foma, an open-source tool. In *Finite-State Methods and Natural Language Processing*, pages 105–113. Springer.

Alegria, I., Etxeberria, I., and Labaka, G. (2013). Una cascada de transductores simples para normalizar tweets. In *Tweet-Norm@ SEPLN*, pages 15–19.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., and Mohri, M. (2007). OpenFST: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23. Springer.

Almeida, J. J., Santos, A., and Simoes, A. (2010). Bigorna–a toolkit for orthography migration challenges. In *Seventh International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta.*

Beesley, K. R. and Karttunen, L. (2003). Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford.*

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC.*

Etxeberria, I., Alegria, I., Hulden, M., and Uria, L. (2014). Learning to map variation-standard forms using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural*, 52:13–20.

Hulden, M. (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. Association for Computational Linguistics.

Hulden, M., Alegria, I. n., Etxeberria, I., and Maritxalar, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, DIALECTS '11, pages 39–48, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiampojamarn, S., Kondrak, G., and Sherif, T. (2007). Applying many-to-many alignments and hidden Markov models to letter-to-phoneme conversion. In *HLT-NAACL*, volume 7, pages 372–379.

Jurish, B. (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.

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

Mann, G. S. and Yarowsky, D. (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.

[3] http://eng.slovenscina.eu/sloleks/opis

[4] http://ixa.eus/Ixa/Produktuak/1456746818

Muggleton, S. and De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming*, 19:629–679.

Novak, J. R., Minematsu, N., and Hirose, K. (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. Association for Computational Linguistics.

Pettersson, E. (2016). Spelling normalisation and linguistic analysis of historical text for information extraction.

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

Porta, J., Sancho, J.-L., and Gómez, J. (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.

Scherrer, Y. (2007). Adaptive string distance measures for bilingual dialect lexicon induction. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, ACL '07, pages 55–60. Association for Computational Linguistics.

Scherrer, Y. and Erjavec, T. (2015). Modernising historical Slovene words. *Natural Language Engineering*, pages 1–25.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (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.