Combined Machine-Learning Approach to PoS-Tagging of Middle English and Old Norse Texts

Raoul Karimov[1[0000-0003-0313-0309]], Andrei Akinin[1[0000-0001-5214-6819]], Dmytro Yakymets[2[0000-0002-4908-3797]]

[1] Chelyabinsk State University, Chelyabinsk, 454001, Russia
[2] Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, 03056, Ukraine
raoul.karimov@hotmail.com, akinin96@gmail.com, fayanzar@gmail.com

Abstract. This paper considers the problem of part-of-speech tagging in Middle English and Old Norse corpora (as well as historical corpora in general). Whereas PoS-tagging generally performs well with journalistic and canonical Modern English texts, the approaches used to solve the problem are not always applicable to older Germanic languages due to various morphology- and syntax-related factors. As such, we believe that Middle English or Old Norse could be handled by a morphographemic encoding and machine learning algorithms like SVM, random forests, kNN, etc. Using a moving-average method to generate multidimensional vectors giving a reliable numeric representation of character composition and sequences, we have achieved a precision and recall of 87.5% in classifying Middle English words by their part of speech while using a simplistic combined voting-based binary classifier; a multinomial classifier was used for a bigger Old Norse sample and performed much worse with an average precision of 64%. This result, however, does encourage further research in the area to solve language-specific methodological problems, while indicating the infancy of the approach proposed.

Keywords: Machine Learning, Corpus, Middle English, Old Norse, PoS-Tagging, Moving Average.

1 Introduction

Part of speech tagging is one of the central issues in the discipline known as natural language processing; it is quite frequently approached by means of hidden Markov models [1]. As Modern English follows a very strict word order, HMMs can be efficiently used to correctly classify words by their part of speech. Furthermore, PoS-tagging is assisted by finite-state transducers, which derive a given word's morphological properties by identifying grammatically-significant character sequences, or morphemes [2]. That, however, might not be applicable to older Germanic languages, which feature less regular word order as well as rich morphology and very inconsistent orthography, still presenting a challenge for linguists working in the field of corpus linguistics and natural language processing.

When speaking of the state of the art in this research area, Moon and Baldridge [3] have found an efficient solution to the PoS-tagging of Middle English using a parallel corpus where two diachronically separated versions of the Bible were aligned to train the algorithm. However, such a resource may not always be available for a historical Germanic language (or any historical language in general). Rögnvaldsson and Helgadóttir [4] applied a TnT tagger to an Old Norse corpus, also achieving a very high accuracy of ~91%; their model was based on an existing tagger for Modern Icelandic, retrained on a manually corrected 95,000-word Old Norse sample. While this is an impressive result, it did require extensive manual work. Neural networks have been made for Slavic and other morphologically complex languages [5, 6], but utilized better-codified and larger-in-volume language data than we could ever afford in our Middle English / Old Norse effort. Nevertheless, PoS-tagging and other functions provided by NLP applications could be of great use for philologists studying language history who are currently restricted to corpora with limited annotation and often have to perform annotation manually. With that in mind, we decided to find a way to automate the process of PoS-tagging by applying existing machine learning methodology.

For this research, the following was hypothesized: there should exist a simple instance-based machine learning method that would enable efficient PoS-classification of orthographically volatile Middle English or Old Norse words while trained on a relatively small set of data. We believed that support vector machines (SVM), random forest models (RFM), k nearest neighbors (kNN), and multilayer perceptron (MLP) could all be used for such learning. The hypothesis is to be verified by means of 10-fold cross-validation.

2 Theory and Methodology

2.1 Research Data

This research derives data for analysis from two sources: the Helsinki Corpus for Middle English and the Menota Archive for Old Norse.

The Helsinki Corpus of English Texts [7] contains about 450 texts and a total 1.5 million words. Preliminary analysis of the corpus data and the preparation of training and test samples were done by consulting Mayhew and Skeat's A Concise Dictionary of Middle English [8]. For the goal of this research, we limited ourselves to only one of the texts from the corpus: Vespasian Homilies, ca. 1167, which partially reduced the overall orthographic and grammatical inconsistency that could be observed across the dialects of that time; from this text, a small 200-word (110 verbs and 90 adjectives) was drawn for the initial machine learning effort, in which we were to perform simple binary classification to see what results could be reasonably achieved on a smaller training set. Another reason to restrict the research to such a small set was the fact that we could not obtain restricted-access parsed Middle English corpora like the Penn-Helsinki Corpus or the Corpus of Early English Correspondence, which confined us to making a manually annotated set.

The Medieval Nordic Text Archive, or Menota, is a 1.6-million fully-parsed open-access collection of texts from Old Icelandic, Old Swedish, and Old Norwegian,

commonly referred to in some sources as Old Norse (albeit the definitions of this language vary); Menota is freely available via the Clarino platform hosted and maintained by the University of Bergen [9]. From this corpus, we derived four subsets of data: 21,464 nouns (common only), 17,068 verbs, 10,585 adjectives, and 2649 adverbs, for a total 51,766 words. Numerals and pronouns were excluded for the initial experiment due to being very limited in number, while prepositions and conjunctions were excluded due to their frequent homography. Having a bigger set of data would enable us to compare the performance of machine learning in two different settings: small sample, binary-opposition, single-text vs large sample, quaternary-opposition collection of texts, although the small-sample result was expected to be more important, as the final goal of this research was (and is) to develop a technique applicable to small user-made corpora.

2.2 Algorithms and Methodology

In this on-going research effort, we investigated the capacities of several known machine-learning algorithms (for space considerations, we are not providing any detailed descriptions of those algorithms): SVM, kNN, RFM, and MLP.

Both Middle English and Old Norse, despite being rather inconsistent both grammatically and orthographically, did have regular morphs that are still referred by historical linguistics as the primary categorial markers. Hypothetically, if we were able to generate word-vectors such that similar character sequences occurring in similar intra-lexeme positions would produce closely-positioned vectors, then a vector-based machine learning algorithm such as SVM or an RFM should be able to correctly link together words that have similar initial and/or final grapheme clusters, which in many cases would suffice for part-of-speech classification.

Therefore, the method to use had to focus on the recurring sequences of symbols observed in words and signifying its PoS-category. As such, we had to find a simplistic yet efficient method that would enable us to represent words in a vector form that would be shaped by both the character composition of, and character positioning in, a given word. Takala [10] cites several methods of vector-word embedding, of which we decided to choose the moving-average method that uses relatively small dimensionality to collect information from all parts of a word.

The moving-average representation is essentially a vector of n dimensions, where n = number of characters in the alphabet, with each dimension being assigned to a single character. A word representation $w = (w_a, w_b \ldots w_z)^T$:

$$W_\alpha = \sum \frac{(1-\alpha)^c \alpha}{Z} \quad (1)$$

where c is the character index (1 for the first symbol in the word, 2 for the second one, etc.), α is a hyper-parameter to control the decay, and Z is a normalizer proportional to the word length (which we decided to be the word length itself, i.e. 4 for word). Thus, each word-vector contains a weighted sum in each dimension representing any character that found in the word, and 0 in the rest of dimensions. The operation is repeated backwards, and the new vector is concatenated to the previous one so

that *word* is represented as *word* ⁀ *drow*. Takala also suggests concatenating a third vector which only contains character counts; for now, we decided not to use that option.

Before the experiment was conducted, we had done a limited normalization of spelling for the Middle English sample: both thorn and eth had been replaced with the cluster *th*, whereas ash, *æ*, had been replaced with *ae*, and yogh, *ȝ* had been replaced with *g*. We also removed diacritics and decapitalized all the words in the text. Thus, we came to an alphabet of 26 characters, which resulted in word-vectors in a 52-dimensional space over the field of real numbers (26x2). This set thus contained 200 instances of 52 numeric attributes + one binary nominal attribute POS {VERB,ADJ}.

For Old Norse, no normalization was done due to the use of a very large alphabet in Menota (we identified 110 different letters after setting the entire sample to the lower case), which could not be reasonably reduced to a Standard Latin 26-character alphabet. The set thus contained 51,766 instances of 220 numeric attributes + one quaternary nominal attribute POS {NOUN,VERB,ADJ,ADV}. All machine learning algorithms were run in the Weka data-mining environment [11].

3 Experimentation and Discussion

As mentioned above, training and verification by 10-fold cross-validation were performed on a small 200-word sample containing 110 verbs and 90 adjectives from a single Middle English texts, then on a relatively large 52k-word 4-PoS Old Norse sample not restricted to any particular text, dialect, or period. All the four models discussed in Section 2.2 above were combined in a single voting-based meta-classifier.

Table 1. Class-specific and weighted average precision (P) and recall (R) values for the combined 4-algorithm classifier: small Middle English sample.

VB P	VB R	ADJ P	ADJ R	Wgt. P	Wgt. R
0.870	0.909	0.882	0.833	0.875	0.875

Apparently, a combined classifier achieved a weighted-average precision and recall of 0.875, which we believe indicates that the combined model showcases a sufficient capability of predicting the part of speech of a given word when trained on 52-dimensional word-vectors generated by the moving-average method. However, a few problems should be highlighted.

First it would be useful to note that verbs generally demonstrate better results than adjectives, which we think is due to the sampling method: as we did not lemmatize or otherwise normalize the form of words we tested the approach on, some adjectives in both sets were given in the superlative form, the suffix of which coincided with the verbal 2SG suffix [12]. Second, it should be borne in mind that the experiment was oversimplified and reduced to two parts of speech, one of which (the verb) is known to be very morphologically complex and rich, featuring better and more indicative character-string markers. On the other hand, ME nouns and adjectives did share many

of their case-specific suffixes, which would probably result in multiple confusions of these two parts of speech should both be included in the experiment. This means that the included algorithms might not make a sufficient PoS-tagging tool despite the morphological richness of Old and Middle English, necessitating further refinement.

To evaluate how using a larger, multi-PoS data set with a considerable dialectal and diachronic span would affect the performance of the classifier, we ran 10-fold cross-validation on the Old Norse sample, and the results turned out to be much worse.

Table 2. Class-specific and weighted average precision (P) and recall ® values for the combined 4-algorithm classifier: large Old Norse Sample.

N P	N R	VB P	VB R	ADJ P	ADJ R	AV P	AV R	Wgt. P	Wgt. R
0.69	0.68	0.66	0.69	0.58	0.56	0.35	0.31	0.64	0.64

In the context of such worse performance, it would also be useful to analyze the confusion matrix for the algorithm.

Table 3. Old Norse PoS confusion matrix.

NOUN	VERB	ADJECTIVE	ADVERB	Classified as
14608	4219	2187	450	NOUN
3722	**11800**	1269	277	VERB
2305	1568	**2881**	831	ADJECTIVE
549	404	872	**824**	ADVERB

As was expected for Middle English, the adjective appears to be a very problematic part of speech for the classification on the basis of character vectors alone; in both Middle English and Old Norse [13], the adjective is morphologically similar to the noun, as it follows a similar declensional paradigm and bears similar suffixes; the adjective is also morphologically homographic to the verb, as in Old Norse, the comparative ending is similar to rhotacized verb endings; finally, the adjective is barely distinguishable from the adverb, as many adverbs are essentially derived from the adjectival neuter gender [Ibid.] As such, character vectors as obtained by the moving-average method prove to be extremely insufficient for handling multi-PoS classification of a large, dialectally and diachronically discrepant text classification. The over-extensive alphabet used in Menota as well as the aforementioned dialectal discrepancy (the corpus contains multiple dialects on the verge of becoming separate languages) might have impeded the performance of the algorithm, necessitating further research into its improvement. Computational performance became an issue for the large sample as well: the algorithm runtime on the Old Norse set exceeded 5,500 seconds **per fold** on a home-PC, which raised the issue of resource intensity for potential at-home application to user-made corpora.

4 Conclusions

This paper analyzes a combination of classifiers that use multidimensional word-vectors generated by means of a moving-average formula applied to every word in a set in direct and reverse order to create a vector reflecting both the character composition of, and the weighted character-specific position in, a given word. The model is cross-validated on a small binary Middle English sample, returning a seemingly good result; however, cross-validation on a relatively large quaternary Old Norse sample indicates a poor performance, which might necessitate further investigation into the algorithm improvements, potentially including the use of complementary techniques such as HMM or multigram-based approaches such as TnT, which has been proven very efficient for Old Norse. Computational performance is currently deemed an issue as well, since the main idea behind the research is to create a simplistic machine that could be easily applied to user-made corpora, a context where the computing power is often limited. Future research will be driven by the need to combine further algorithms while also seeking ways to optimize both data sampling and the algorithm performance as well. Another potential improvement may lie with the use of complementary vectorization methods.

References

1. Jurafsky, D., Martin, J.H.: Speech and Language Processing. New Jersey. Prentice Hall (2008).
2. Beesley, K.R., Karttunen, L.: Finite-State Morphology. Journal of Computational Linguistics, 30-2, 237–249 (2004).
3. Moon, T., Baldridge, J.: Part-of-speech Tagging for Middle English through Alignment and Projection of Parallel Diachronic Texts. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 390–399 (2007).
4. Rögnvaldsson, E., Helgadóttir, S.: Morphological Tagging of Old Norse Texts and Its Use in Studying Syntactic Variation and Change. In: Sporleder C., van den Bosch A., Zervanou K. (eds) Language Technology for Cultural Heritage. Theory and Applications of Natural Language Processing, 63–72. Springer, Berlin, Heidelberg (2011).
5. Jędrzejowicz P., Strychowski J. A.: Neural Network Based Morphological Analyser of the Natural Language. In: Intelligent Information Processing and Web Mining. Advances in Soft Computing, vol 31, 199–208. Springer, Berlin, Heidelberg (2005).
6. Malouf, R.: Generating morphological paradigms with a recurrent neural network. San Diego Linguistic Papers 6, 122–129 (2016).
7. Helsinki Corpus of English Texts, www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus, last accessed 2018/04/03.
8. Mayhew, M.A., Skeat, W.: A Concise Dictionary of Middle English From A.D. 1150 to 1580. Oxford, Clarendon Press (1888).
9. Medieval Nordic Text Archive, www.menota.org, last accessed 2018.06.15
10. Takala, P.: Word Embeddings for Morphologically Rich Languages. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium (2016).

11. Frank, E, Witten, I.H.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2016).
12. Ilyish, B.A.: History of the English Language. Vysshaya Shkola, Moscow (1968).
13. Haugen, O.E. Handbok i Norrøn Filologi. 2. utgave. Bergen. Fagbokforlaget (2013).