Unsupervised Vocabulary Expansion with Whole Word Morphology

Maciej Janicki
macjan@o2.pl

Abstract

Unsupervised learning of morphology has recently been applied to predict unseen, but morphologically possible words. We propose to approach this task using a model of morphology based on string transformations on whole words, rather than segmentation. In this way, overgeneration caused by inaccurate modeling of morphotactics is avoided, as the transformation-based model generates words very similar to the ones seen in the training set (mostly missing inflections). The evaluation on several inflecting languages shows that the whole-word vocabulary expansion is much more successful in reducing OOV word rates than a recently suggested approach based on state-of-the-art unsupervised morphological segmentation and language modeling.

1 Introduction

Natural Language Processing methods frequently make use of fixed lists of words from a given language, called *vocabularies*. The need for listing all words of a language arises typically in cases where the words need to be recognized in a noisy representation, like an image (Optical Character Recognition, OCR) or acoustic signal (Automatic Speech Recognition, ASR). Such lists are usually constructed automatically based on corpora and thus incomplete.

The purpose of *vocabulary expansion* is to suggest further possible words given a vocabulary. In the speech recognition context, syllable-based generation of new words has previously been suggested (Lei et al., 2009; Trmal et al., 2014), as well as using directly a lexicon of morphs instead of words (Kurimo et al., 2006).

Recently, unsupervised vocabulary expansion based on morphological criteria has been attempted (Rasooli et al., 2014; Varjokallio and Klakow, 2016). Both these papers use Morfessor (Creutz and Lagus, 2005b,a; Virpioja et al., 2013) to segment the training vocabulary and subsequently generate new words by combining the induced morphs into new sequences. In order to account for morphotactics and avoid overgeneration, the authors use additional language models to score the proposed words. (Rasooli et al., 2014) use scoring based on character trigrams, while (Varjokallio and Klakow, 2016) use language models trained on morph sequences.

An alternative method for unsupervised learning of morphology and its application in vocabulary expansion was proposed by (Neuvel and Fulop, 2002). It consists of learning transformational rules operating on whole words from tagged corpora. The rules learned this way were subsequently used for lexicon expansion. No statistical inference was involved: all discovered rules exceeding a frequency threshold were used.

The approach of (Neuvel and Fulop, 2002) is based on a linguistic theory called *Whole Word Morphology* (Ford et al., 1997), which rejects the notion of morpheme as means of describing structural similarities between words. Instead, it proposes to treat words as atomic units of language and express morphology in terms of patterns operating on whole words. Although this view is uncommon in theoretical linguistics, the criticism of morpheme and a description on morphology based on rewrite rules have been voiced by multiple other authors (Aronoff, 1976; Anderson, 1992; Starosta, 1988; Singh and Starosta, 2003).

In the remainder of this paper, we propose a novel method of unsupervised vocabulary expansion. Our approach utilizes a probabilistic model designed for unsupervised learning of transformational rules in the spirit of Whole Word Morphology, which was recently presented by (Janicki, 2015; Sumalvico, 2017). In Section 2, we briefly describe the model and show how it can be used to

suggest and score new words. Section 3 describes experiments, in which the proposed model is evaluated on the task of reducing OOV rates and compared to one of the previous approaches. Section 4 contains a brief conclusion.

2 The Method

2.1 Unsupervised Learning of Whole Word Morphology

The probabilistic model proposed by (Janicki, 2015; Sumalvico, 2017) describes the morphological structure of a lexicon as a graph, in which words are vertices and directed edges denote a derivation of a word by a productive morphological rule. The rules are formulated as patterns, e.g. the rule responsible for the pair (*related*, *relation*) might have the following form (cf. Ford et al., 1997):

$$/X\text{ed}/ \rightarrow /X\text{ion}/ \tag{1}$$

In this notation, X is a wildcard that can be instantiated with any string and a pattern between slashes refers to a whole word in its surface form. The rules may also include context: $/X\text{ted}/ \rightarrow /X\text{tion}/$ would be a possible alternative to (1).

The model defines a probability $P(V, E|R, \theta)$ of vocabulary V and edges E given rules R and numeric parameters θ. The latter is a vector of rule application probabilities: for each rule r, θ_r denotes the probability that r is applied to generate a new word provided that the conditions for applying it are met. For example, the rule (1) can only apply to words ending in *-ed*. The graph probability is defined as follows:

$$Pr(V, E|R, \theta) \propto \prod_{v \in V_0} \rho(v) \times \prod_{v \in V} \prod_{r \in R} \prod_{v' \in r(v)} \begin{cases} \theta_r & \text{if } \langle v, v', r\rangle \in E, \\ 1 - \theta_r & \text{if } \langle v, v', r\rangle \notin E \end{cases} \tag{2}$$

with V_0 being the set of root nodes and $\rho(\cdot)$ being a probability distribution over arbitrary strings (for example, based on unigram frequencies). $r(\cdot)$ is a function applying the rule r to a word.[1]

The training of the model consists of a preprocessing step, in which candidates for rules and edges are extracted from pairs of string-similar words, and subsequent statistical inference using Monte Carlo Expectation Maximization (Wei and Tanner, 1990). In the latter step, Metropolis-Hastings sampling is used to obtain large samples of graphs within the space defined by candidate edges.

2.2 Vocabulary Expansion

In the following, we propose a cost function $c(\cdot)$ for extending a vocabulary with new words given a trained morphology model. For a single new word $v \notin V$, we consider the following log-likelihood ratio:

$$c(v) = -\log \frac{P(V \cup \{v\}|R, \theta)}{P(V|R, \theta)} \tag{3}$$

$$= -\log \frac{\sum_E P(V \cup \{v\}, E|R, \theta)}{\sum_E P(V, E|R, \theta)} \tag{4}$$

The sums are taken over all forests[2] possible within the set of candidate edges generated from V. In the numerator, the edges involving v are additionally considered. In order to make the computation tractable, we will restrict this sum to graphs, in which v is either a root or a leaf node.

Let $p_0 = \rho(v)$ and $p_1, \ldots, p_n$ denote the probabilities of possible edges deriving v. Observe that the probability of each graph considered in the denominator contains a multiplicative term $\prod_{i=1}^{n}(1 - p_i)$: none of these edges is present in such graphs (as v is not present as a node), but each of them is possible (i.e. considered in (2)), because the corresponding source nodes and rules are present. By drawing this term in front of the sum and calling the rest of the sum Z, we can write the denominator as $\prod_{i=1}^{n}(1 - p_i)Z$.

Now, let us consider graphs, in which v is a root node. Each of them can be thought of as one of the graphs from the denominator with the addition of v as a root. The probability of such graphs can thus be summarized as $p_0 \prod_{i=1}^{n}(1-p_i)Z$. Finally, let us consider graphs, in which v is derived by the j-th edge. Their probability can be similarly summarized as $\frac{p_j}{1-p_j} \prod_{i=1}^{n}(1 - p_i)Z$. Shortening the common therm $\prod_{i=1}^{n}(1 - p_i)Z$ from the numerator and the denominator, we obtain the final formula for the cost of adding v to the vocabulary:

$$c(v) = -\log \left[p_0 + \sum_{j=1}^{n} \frac{p_j}{1 - p_j} \right] \tag{5}$$

[1] Note that the result of applying a rule to a word is in general a *set* of words, as there might be multiple resulting words. The set is empty if the context on the left-hand side of the rule is not matched.

[2] Graphs without cycles, in which every node has at most one incoming edge.

The cost is thus dependent on all possible ways of deriving v: the more possible source words and rules 'support' it, the cheaper it becomes. Note that in Whole Word Morphology there is no concept of 'lemma': any inflected form can be derived from any other, provided that the pattern is sufficiently productive. The summing of probabilities of individual derivations enables the model to take into account whole morphological paradigms and reduces overgeneration caused by applying a productive rule to a wrong base word (like e.g. adding a verb suffix to a noun).

3 Experiments

3.1 Datasets and Setup

As a source of unannotated corpora, we utilize the Leipzig Corpora Collection[3] (Biemann et al., 2007; Goldhahn et al., 2012), which provides freely available corpora in standardized sizes for a large variety of languages. The experiments described here were carried out on corpora of 100k sentences for seven different languages. Each corpus was randomly split on sentence basis into training and evaluation set in proportion 1:9. Both parts were converted into lists of word types. Table 1 provides some statistics about the resulting datasets. No preprocessing was applied – the wordlists were taken as they are, including punctuation, numbers, uppercase etc.

We compare the following methods:

SEG is a simplified[4] version of the approach proposed by (Varjokallio and Klakow, 2016). It consists of segmenting the training vocabulary with Morfessor 2.0 (Virpioja et al., 2013) and training the RNNLM language model (Mikolov et al., 2010; Mikolov, 2012) on the segmentations. The language model treats morphs as words and segmented words as sentences. We subsequently sample 100 million sequences from the language model, which results in around 4 to 5 million new word types, depending on the dataset. The training parameters for both Morfessor and RNNLM were set exactly as described by (Varjokallio and Klakow, 2016) (Morfessor: $\alpha = 0.8$; RNNLM: 50 classes, hidden layer size 20). For RNNLM training, one fifth of the segmentations was used as a validation set.

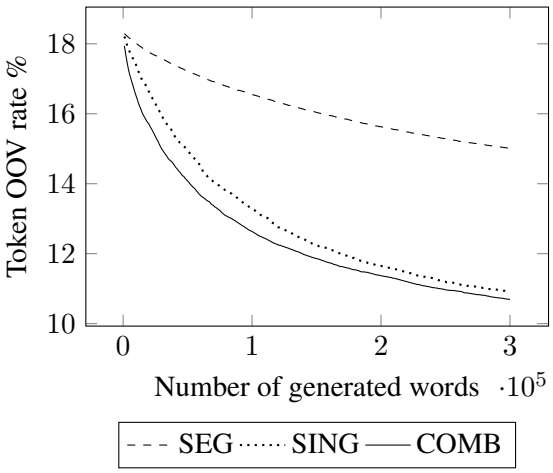

Figure 1: Token OOV rate for Latin.

SING is the Whole Word Morphology model described in Sec. 2.1. However, instead of using (5) to derive word costs, only the minus log-likelihood of the best edge is used. This approach roughly corresponds to the previous evaluations of (Janicki, 2015; Sumalvico, 2017). The model is fitted using 5 iterations of the MCEM algorithm with 10 million iterations of sampling each and the number of rules discovered in the preprocessing step is limited to 5,000 most frequent ones.

COMB is the Whole Word Morphology model utilizing (5) to compute the word costs via combining the probabilities of all possible derivations. The training parameters are the same as in the SING setup.

3.2 Results

We compare the reduction of OOV rates for different sizes of the generated vocabulary. The results presented in Tables 2-3 and Figures 1-2 indicate a clear advantage of the word-based model over the segmentation approach. The workings of both methods are fundamentally different: while the word-based model explores the nearest neighborhood of the training vocabulary, applying small and well-motivated changes to known words, the 'disassemble and reassemble' approach of the morph-based model frequently produces words very much unlike anything seen in the training data.

The difference is especially prominent for languages with extensive fusional inflection (the bot-

[3] http://corpora.uni-leipzig.de

[4] (Varjokallio and Klakow, 2016) obtained their best results for an interpolation of n-gram and RNN language models. However, as evidenced by Figure 1 in that paper, the difference between the interpolated model and the individual models is not critical, so we decided to reproduce a simpler setup for the sake of comparison.

Language	Training corpus		Evaluation corpus			
	Types	Tokens	Types	Tokens	Token OOV rate	Type OOV rate
Finnish	50,570	142,741	256,465	1,286,794	27.6 %	88.0 %
German	41,877	194,267	203,645	1,744,533	16.2 %	87.1 %
Latin	41,566	167,169	183,033	1,482,448	18.3 %	84.3 %
Latvian	42,488	165,135	173,683	1,479,769	18.1 %	82.0 %
Polish	46,977	173,159	200,939	1,562,446	19.6 %	83.2 %
Romanian	40,803	216,335	161,858	1,950,586	12.7 %	81.2 %
Russian	50,420	173,233	227,231	1,557,690	21.2 %	85.0 %

Table 1: The datasets used for evaluation.

Language	SEG	SING	COMB
Finnish	7.78 %	14.71 %	17.70 %
German	7.23 %	14.82 %	16.51 %
Latin	9.65 %	27.43 %	31.03 %
Latvian	9.64 %	29.80 %	33.23 %
Polish	10.59 %	28.91 %	32.76 %
Romanian	8.35 %	29.86 %	31.98 %
Russian	9.57 %	26.60 %	30.90 %

Table 2: Token-based OOV rate reduction for vocabulary expansion with 100k new words.

Language	SEG	SING	COMB
Finnish	3.58 %	7.73 %	9.29 %
German	3.57 %	7.36 %	8.33 %
Latin	5.51 %	15.72 %	17.80 %
Latvian	5.40 %	17.39 %	19.54 %
Polish	5.58 %	16.32 %	18.56 %
Romanian	4.57 %	16.03 %	17.47 %
Russian	4.79 %	14.06 %	16.63 %

Table 3: Type-based OOV rate reduction for vocabulary expansion with 100k new words.

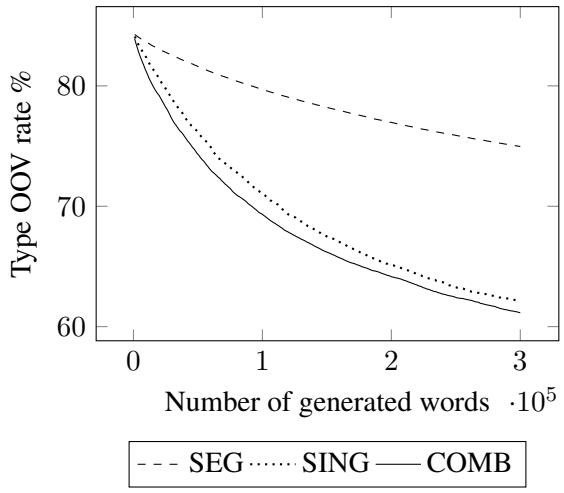

Figure 2: Type OOV rate for Latin.

tom five in the tables), which are most affected by the inadequacies of the segmentational view on morphology. Furthermore, missing inflected forms account for most OOV words in those languages. On the other hand, Finnish exhibits agglutinative morphology, which fits much better to the segmentational view. In German, the fusional inflection is comparatively small and the concatenative mechanism of compounding accounts for most OOV words.

The difference between the COMB and SING setups, measuring the impact of summing all possible derivation alternatives with (5), is small compared to the difference between both and SEG. The most plausible reason for that seems to be that the probabilities $p_1, \ldots, p_n$ of different derivations might differ in the order of magnitude, so their sum is often close to the single largest value. Nevertheless, the COMB setup provides a significant and consistent improvement for all languages and extended vocabulary sizes.

4 Conclusion

We presented a method for unsupervised vocabulary expansion based on a probabilistic model of morphology as whole-word string transformations. In the task of vocabulary expansion, this approach turned out to be much more effective than previous methods relying on unsupervised morph segmentation and scoring of newly generated morph sequences. Its advantage lies in generating mostly missing inflected forms through small changes in the known vocabulary. The main contribution of the present paper – taking into account multiple possible ways of deriving a word to compute a combined score – provided a consistent improvement in the OOV rate reduction.

References

Stephen R. Anderson. 1992. *A-Morphous Morphology*.

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. MIT Press.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection: Monolingual corpora of standard size. In *Proceedings of Corpus Linguistics 2007*, Birmingham, UK.

Mathias Creutz and Krista Lagus. 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*.

Mathias Creutz and Krista Lagus. 2005b. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Technical report, University of Helsinki.

Alan Ford, Rajendra Singh, and Gita Martohardjono. 1997. *Pace Pāṇini: Towards a word-based theory of morphology*. American University Studies. Series XIII, Linguistics, Vol. 34. Peter Lang Publishing, Incorporated.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Language Resources and Evaluation (LREC'12)*.

Maciej Janicki. 2015. A multi-purpose bayesian model for word-based morphology. In *Systems and Frameworks for Computational Morphology – Fourth International Workshop, SFCM 2015*. Springer.

Mikko Kurimo, Antti Puurula, Ebru Arisoy, Vesa Siivola, Teemu Hirsimäki, Janne Pylkkönen, Tanel Alumäe, and Murat Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 487–494, New York.

Xin Lei, Wen Wang, and Andreas Stolcke. 2009. Data-driven lexicon expansion for mandarin broadcast news and conversation speech recognition. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan.

Tomáš Mikolov. 2012. *Statistical Language Models Based on Neural Networks*. Ph.D. thesis, Brno University of Technology.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *11th Annual Conference of the International Speech Communication Association 2010 (INTERSPEECH 2010)*, pages 1045–1048.

Sylvain Neuvel and Sean A Fulop. 2002. Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*, MPL '02, pages 31–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Thomas Lippincott, Nizar Habash, and Owen Rambow. 2014. Unsupervised morphology-based vocabulary expansion. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 1349–1359.

Rajendra Singh and Stanley Starosta, editors. 2003. *Explorations in Seamless Morphology*. SAGE Publications, New Delhi.

Stanley Starosta. 1988. *The Case for Lexicase: An Outline of Lexicase Grammatical Theory*. Open linguistics series. Pinter Publishers.

Maciej Sumalvico. 2017. Unsupervised learning of morphology with graph sampling. In *Proceedings to RANLP 2017*, Varna, Bulgaria.

Jan Trmal, Guoguo Chen, Dan Povey, Sanjeev Khudanpur, Pegah Ghahremani, Xiaohui Zhang, Vimal Manohar, Chunxi Liu, Aren Jansen, Dietrich Klakow, David Yarowsky, and Florian Metze. 2014. A keyword search system using open source software. In *Proceedings of the IEEE 2014 Workshop on Spoken Language Technology*, South Lake Tahoe, NV, USA.

Matti Varjokallio and Dietrich Klakow. 2016. Unsupervised morph segmentation and statistical language models for vocabulary expansion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 175–180.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor baseline. Technical report, Aalto University, Helsinki.

Greg C. G. Wei and Martin A. Tanner. 1990. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704.