Unsupervised morph segmentation and statistical language models for vocabulary expansion

Matti Varjokallio *
Dept. of Signal Processing and Acoustics
Aalto University
Espoo, Finland
matti.varjokallio@aalto.fi

Dietrich Klakow
Spoken Language Systems
Saarland University
Saarbrücken, Germany
dietrich.klakow@lsv.uni-saarland.de

Abstract

This work explores the use of unsupervised morph segmentation along with statistical language models for the task of vocabulary expansion. Unsupervised vocabulary expansion has large potential for improving vocabulary coverage and performance in different natural language processing tasks, especially in less-resourced settings on morphologically rich languages. We propose a combination of unsupervised morph segmentation and statistical language models and evaluate on languages from the Babel corpus. The method is shown to perform well for all the evaluated languages when compared to the previous work on the task.

1 Introduction

Language modelling for different natural language processing tasks like speech recognition, machine translation or optical character recognition require large training corpora to achieve good language model estimates and high enough vocabulary coverage. Sometimes such resources are not readily available or easily acquirable. This is especially the case for the many less-resourced languages. In the case of morphologically rich languages, these issues are emphasized, as words appear in many forms, thus increasing the required vocabulary size and the data sparsity. Automatic speech recognition of spontaneous speech is a task with some special characteristics, as speech transcriptions are expensive to acquire. Taking all these factors into account, the importance of making the most out of the available resources becomes evident.

This work was done while the author was visiting the Saarland University Spoken Language Systems group

Previous work on handling out-of-vocabulary (OOV) words in automatic speech recognition have included explicit OOV word modelling and confidence measures (Hazen and Bazzi, 2001) and hybrid word-subword language modelling for OOV word detection (Yazgan and Saraçlar, 2004). Speech recognition by directly using optimized subword units has also (Kneissler and Klakow, 2001) proven a good approach for speech recognition of a morphologically rich language.

In this work, we study unsupervised vocabulary expansion for conversational speech recognition of morphologically rich languages in a less-resourced setting. We expand the recognition vocabulary, and thus lower the OOV rate, by generating new word forms. Two recent works also target the unsupervised vocabulary expansion.

In (Rasooli et al., 2014), an unsupervised morphological segmentation was inferred from the training corpus using the Morfessor Categories-MAP (Creutz and Lagus, 2007) method. The prefix-stem-suffix structure estimated by the model was then represented as a finite-state-transducer for sampling new word forms. Different reranking schemes using a bigram language model and a letter trigraph language model were evaluated.

The Kaldi speech recognition package (Povey et al., 2011) includes an approach (Trmal et al., 2014) for vocabulary expansion. In this approach, the provided syllable segmented pronunciation lexicon is used as the basis for the expansion. An n-gram model is trained over the syllable segmentation and syllabic words are generated from the model. Finally a phoneme-to-grapheme mapping is performed to obtain the grapheme form for the words.

In our approach, statistical language models are trained over a morph segmentation, which is learned unsupervisedly from the data. Words are

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 175–180,
Berlin, Germany, August 7-12, 2016.

sampled from the language models and ordered according to the probabilities given by the language models. We evaluate the method on seven morphologically rich languages from the Babel (Harper, 2013) corpus and compare to the previously suggested approaches.

2 Suggested method

We present a combination of unsupervised morph segmentation and statistical language models for unsupervised vocabulary expansion. The suggested approach operates in four steps: unsupervised morph segmentation, statistical language model training, sampling of new word types and reranking of the sampled words. The phases are described in more detail in the corresponding subsections.

2.1 Unsupervised morph segmentation

Morfessor Baseline (Creutz and Lagus, 2002) is a method for unsupervised morphological segmentation. The algorithm optimizes a two-part minimum description length code, finding a balance between the cost of encoding the training corpus and the lexicon, as in Formula 1.

$$\arg\min_{\theta} L(x, \theta) = \arg\min L(x|\theta) + L(\theta) \quad (1)$$

The corpus encoding is based on a unigram model. A so-called α-term may be used for fine-tuning the corpus encoding cost. For the experiments in this work, a recent Python implementation Morfessor 2.0 (Smit et al., 2014) was used.

2.2 Statistical language models over morphs

As statistical language models, two state-of-the-art models were selected. These language models were trained on a corpus, where one segmented word was treated as what would in normal language model training be a sentence. The training was done using log-weighted word frequencies, thus some words appearing multiple times in the training corpus. The rationale of the log-weighting was to slightly emphasize the most common words. As a last step, the order of the training words was randomized.

The first model was a trigram model trained with the modified Kneser-Ney smoothing (Kneser and Ney, 1995) using three discounts per order. The discount parameters could normally be optimized on a held-out-set, but here leave-one-out estimates were used, as it is not clear what would in this case constitute a reasonable held-out set. The model was trained using the VariKN software package (Siivola et al., 2007).

It has recently been shown, that the recurrent neural network language models may efficiently be trained using the backpropagation algorithm (Mikolov et al., 2010), making it also an appealing choice for language modelling. As the second language model, a recurrent neural network language model was trained using the RNNLM toolkit. The words were treated as independent of the preceeding words in both the model training and the word sampling phases.

2.3 Sampling and reranking

The initial set of candidate words was obtained by sampling separately from both the n-gram model and the recurrent neural network language model. These word lists were then merged. It is very important to rerank the obtained word list, as the goal is to improve the OOV rate as much as possible with introducing as little incorrect words as possible to the vocabulary. As the final estimate on the word likelihood, the linear interpolation of these two model scores was used. The linear interpolation was applied morph-wise. The list of the sampled words was sorted in descending order with the linearly interpolated likelihood as the score.

3 Experiments

3.1 Training corpus

The vocabulary expansion experiments were conducted on the Babel corpus (DARPA, 2013). The experiments were run on the following set of languages: Assamese, Bengali, Pashto, Tagalog, Tamil, Turkish and Zulu. The training corpora consist mainly of conversation transcriptions, but also additional scripted data is provided. Including the scripted training data in general helps to lower the OOV rate. The OOV reduction rate reachable by the vocabulary expansion then becomes, with some exceptions, slightly slower. Statistics of the datasets are in the Table 1.

As preprocessing, all special symbols were removed from the texts. Asterisk symbols are used to denote misspellings in cases where the real word was identifiable. Asterisk symbols were removed and the words included in the training corpus. Dash symbols in the beginning and end of

Language	Training data		Development data			
	Types	Tokens	Types	Tokens	Type OOV%	Token OOV%
Assamese	8738	73284	7309	66357	49.75	8.36
Bengali	9507	81564	7844	70724	50.90	8.56
Pashto	7027	115225	6174	108273	44.91	4.26
Tagalog	6370	69791	5614	64506	55.61	8.13
Tamil	16284	76916	14279	70429	65.08	16.89
Turkish	12147	77310	9944	67171	57.25	12.53
Zulu	16008	65821	13848	57217	68.88	21.91

Table 1: Statistics of the datasets used in the experiments. The scripted training corpus is included.

a word are used to indicate hesitations. These words were removed from the training corpus. Only proper names were written in uppercase in the transcriptions, so these words were kept intact.

3.2 Expansion model

As statistical language models, we evaluated a trigram language model, a recurrent neural network language model, and the linear interpolation of these models. 10 million new distinct word types were sampled from both the models separately. These lists were then merged and reranked as explained in the Section 2.3.

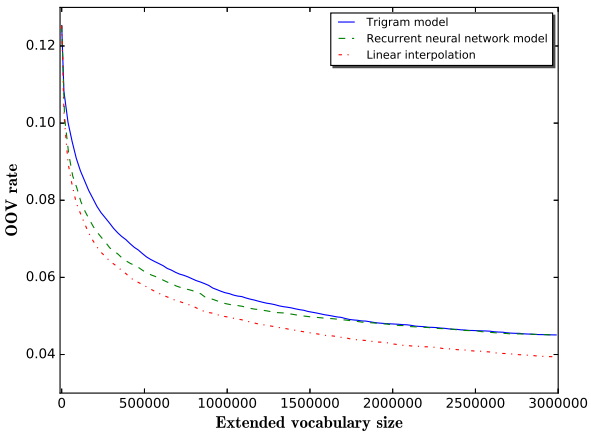
The model parameters were optimized on selected languages and these parameters were used in all the experiments. For the recurrent neural network language model, the number of classes was set to 50 and the hidden layer size to 20. These values were reasonably close to optimum for all the languages.

The suitable α-value for the Morfessor Baseline segmentation was studied. With the default value of 1.0, the method seemed to suffer from a slight undersegmentation. To encourage the method to segment more, the α value was set to 0.8. This setting was equal or better for all the evaluated languages.

When evaluating the language models as standalone models, the trigram model provided better generation accuracy for 4 of the in total 7 languages and the recurrent neural network language model for 3 of the languages. Linear interpolation of the models was without exceptions the most accurate model. The linear interpolation weight was set to 0.5.

Figure 1 shows an example of the OOV rate development as a function of the extended vocabulary size for Turkish. The rapid improvement of the OOV rate for small extensions and the superiority of the linearly interpolated model are characteristics shared by all the languages.

Figure 1: Token-based OOV rate as a function of the extended vocabulary size for Turkish



3.3 Comparison to the previous work

We compared the approach to the previous results in (Rasooli et al., 2014). They reported the results for a vocabulary expansion of 50k best words. Table 2 compares the type-based expansion results and Table 3 the token-based expansion results. Models for these comparisons were trained with the scripted data included.

Language	Rasooli et al.	Suggested method
Assamese	28.46	31.93
Bengali	24.75	33.20
Pashto	19.43	32.95
Tagalog	16.81	21.27
Tamil	-	16.27
Turkish	14.79	28.32
Zulu	13.87	21.18

Table 2: Type-based OOV reduction rates for the 50k best words

Language	Rasooli et al.	Suggested method
Assamese	29.43	35.17
Bengali	25.61	35.16
Pashto	21.27	35.55
Tagalog	16.88	23.75
Tamil	-	19.24
Turkish	17.82	31.89
Zulu	15.67	23.62

Table 3: Token-based OOV reduction rates for the 50k best words

Language	Vocabulary size	Kaldi	Suggested
Assamese	845k	26.4	21.2
Bengali	834k	27.4	22.0
Pashto	494k	26.7	20.3
Tagalog	581k	37.5	33.2
Tamil	896k	45.2	38.0
Turkish	704k	37.1	28.4
Zulu	818k	40.7	37.0

Table 4: Type-based OOV rate comparison to Kaldi

Language	Vocabulary size	Kaldi	Suggested
Assamese	845k	4.3	3.5
Bengali	834k	4.6	3.6
Pashto	494k	2.4	1.9
Tagalog	581k	5.3	4.6
Tamil	896k	11.2	9.1
Turkish	704k	7.9	6.0
Zulu	818k	12.5	11.4

Table 5: Token-based OOV rate comparison to Kaldi

We ran the Kaldi vocabulary expansion in the limited language pack setting as in (Trmal et al., 2014). In the default setting, around 1M distinct syllabic words are generated and converted by a phoneme-to-grapheme mapping to obtain the graphemic word form. Table 4 compares the type-based expansion results and Table 5 the token-based expansion results for a vocabulary expansion of similar size (in graphemic words). The scripted data was not used in training the models for these comparisons.

3.4 OOV reduction and type to token ratio

The OOV reduction was evaluated as a function of the type/token ratio. This analysis may provide information about the properties of the evaluated languages. The token-based analysis is in the Figure 2 and the type-based analysis in the Figure 3. As the type/token ratio is dependent on the number of tokens, these values are computed on a matched number of tokens (65821) from the training corpus. The plots show that there are similarities, but also big differences between the languages. Most notable exceptions seem to be Tamil and Tagalog. For Tamil, the number of the most frequent words was lower with a slightly more even tail of less frequent words. For Tagalog, the average number of morphs per word as estimated by the Morfessor Baseline algorithm was 2.8, which was the highest value among all the languages. Still, the number of distinct word types in the training set was the lowest. These properties seem to play a role in the different vocabulary expansion characteristics.

Figure 2: Token-based OOV reduction rate for 50k word expansion as a function of type/token ratio

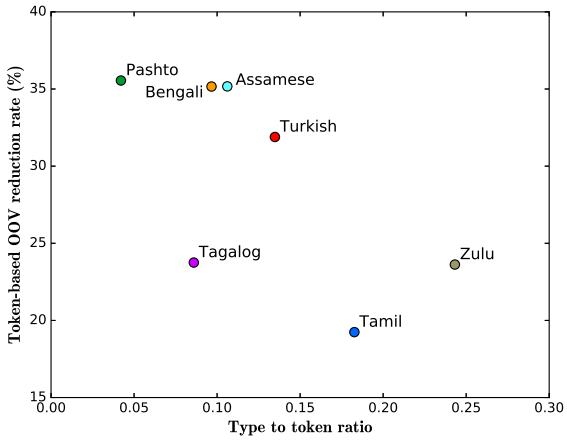
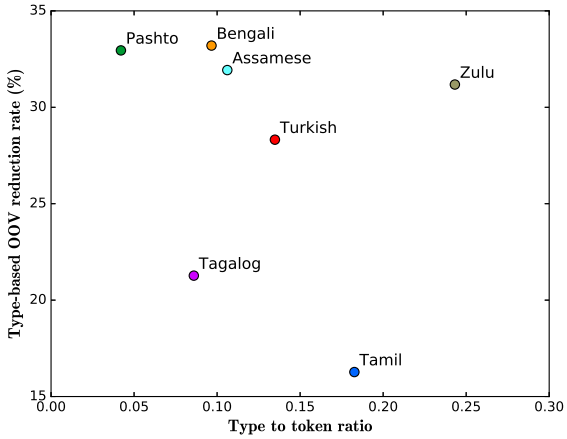


Figure 3: Type-based OOV reduction rate for 50k word expansion as a function of type/token ratio



4 Discussion

This work concerned the use of unsupervised morphological segmentation and statistical language models for the task of vocabulary expansion. Unsupervised vocabulary expansion has large potential for reducing OOV-rates and improving results in NLP tasks especially in less-resourced settings for morphologically rich languages.

The suggested method was evaluated on some of the morphologically rich languages of the Babel corpus in the limited language pack condition. The performance of the method was evaluated in terms of the improvement of the OOV-rate on the development set. The suggested combination of segmentation and interpolation of statistical language models provided to our understanding the best results on the task so far. Compared to (Rasooli et al., 2014), our approach differed in that the statistical language models were used directly in the word generation phase. As opposed to (Trmal et al., 2014), our approach operated purely on the grapheme level.

It is perhaps noteworthy, that the methods are not that different from what one would use in a normal language modelling scenario for automatic speech recognition. Morfessor Baseline (Creutz and Lagus, 2002) has been seen to give good results in morph-based speech recognition (Creutz et al., 2007) when used along with standard n-gram models. If a larger training corpus is available, optimizing unigram likelihood more directly may be a good choice (Varjokallio et al., 2013).

Morph segmentations provided by the Morfessor Flatcat (Grönroos and Virpioja, 2014) -method were also evaluated for this work, but Morfessor Baseline was found to perform better. It is possible, that the tradeoff between the lexicon cost and the corpus encoding cost, as given by the Minimum Description Length -principle, is important for the modelling accuracy in this type of a less-resourced scenario. Morfessor Flatcat will in most cases segment more accurately according to the grammatical morph boundaries. This is likely a more valuable property for statistical machine translation than for the present task.

The linear interpolation of an n-gram model and a recurrent neural network language model provides at the moment state-of-the-art modelling accuracy in many statistical language modelling tasks. Some forms of class n-grams were also evaluated for this work. Sampling from a class n-gram provided many complementary word forms, not easily generated by the other models. However, it became successively harder to improve the OOV reduction rates by a combination of three models.

This work concentrated only on methods for expanding the vocabulary. Naturally some language modelling methods are required to utilize these generated words in speech recognition or some other task. One possibility is to extend the unknown symbol and improve the obtained estimates via class n-gram models (Trmal et al., 2014). Morph-based language models may be utilized using a constrained vocabulary as suggested in (Varjokallio and Kurimo, 2014). In this case word-level pronunciation variants may be applied. Performing the vocabulary expansion may also provide insights into unlimited vocabulary speech recognition (Kneissler and Klakow, 2001; Hirsimäki et al., 2006) with morph language models. Finding units with consistent grapheme-to-phoneme mapping may, however, be challenging for some of the Babel languages.

Regarding the type of approaches considered in this work, it is possible that advances in either unsupervised morph segmentation or statistical language models could bring about further improvements in the expansion accuracy. Unsupervised learning of morphological paradigms is also a potential direction when seeking for improvements in the task.

5 Conclusion

Unsupervised vocabulary expansion has great potential for reducing out-of-vocabulary rates and improving results in different natural language processing tasks, including ASR. In this work, an approach comprising of unsupervised morph segmentation and statistical language models was suggested. The model was evaluated on the Babel languages and was shown to give large improvements compared to the previous work on the task.

Acknowledgments

This research was conducted while the first author was visiting the Saarland University Spoken Language Systems group. The work was partially funded by the Saarland University SFB1102 Collaborative Research Center for Information Density and Linguistic Encoding.

References

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.

Mathias Creutz and Krista Lagus. 2007. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based Speech Recognition and Modeling of Out-of-vocabulary Words Across Languages. *ACM Transactions on Speech and Language Processing*, 5(1):3:1–3:29, December.

DARPA, 2013. *IARPA Babel Data Specifications for Performers*.

Stig-Arne Grönroos and Sami Virpioja. 2014. Morfessor FlatCat: An HMM-Based Method for Unsupervised and Semi-Supervised Learning of Morphology. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 1177–1185, Dublin, Ireland, August.

Mary Harper. 2013. The Babel Program and Low Resource Speech Technology. Automatic Speech Recognition and Understanding Workshop (ASRU), Invited talk.

Timothy J. Hazen and Issam Bazzi. 2001. A Comparison and Combination of Methods for OOV Word Detection and Word Confidence Scoring. In *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 397–400.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited Vocabulary Speech Recognition with Morph Language Models Applied to Finnish. *Computer Speech and Language*, 20(4):515–541.

Jan Kneissler and Dietrich Klakow. 2001. Speech Recognition for Huge Vocabularies by Using Optimized Sub-word Units. In *Proceedings of the 2nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2001)*, pages 69–72.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-off for M-gram Language Modeling. In *Proceedings of the 1995 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–184.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, December.

Mohammad Sadegh Rasooli, Thomas Lippincott, Nizar Habash, and Owen Rambow. 2014. Unsupervised Morphology-Based Vocabulary Expansion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1349–1359, Baltimore, Maryland, June.

Vesa Siivola, Teemu Hirsimäki, and Sami Virpioja. 2007. On Growing and Pruning Kneser-Ney Smoothed N-gram Models. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5):1617–1624, July.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for Statistical Morphological Segmentation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 21–24, Gothenburg, Sweden, April.

Jan Trmal, Guoguo Chen, Daniel 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, USA, December.

Matti Varjokallio and Mikko Kurimo. 2014. A Word-Level Token-Passing Decoder for Subword n-gram LVCSR. In *Proceedings of the IEEE 2014 Workshop on Spoken Language Technology*, South Lake Tahoe, USA, December.

Matti Varjokallio, Mikko Kurimo, and Sami Virpioja. 2013. Learning a Subword Vocabulary Based on Unigram Likelihood. In *Proceedings of the IEEE 2013 Workshop on Automatic Speech Recognition and Understanding*, Olomouc, Czech Republic, December.

Ali Yazgan and Murat Saraçlar. 2004. Hybrid Language Models for Out of Vocabulary Word Detection in Large Vocabulary Conversational Speech Recognition. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–745–8.