Unsupervised segmentation of words into morphemes – Morpho Challenge 2005 Application to Automatic Speech Recognition

Mikko Kurimo, Mathias Creutz, Matti Varjokallio

Adaptive Informatics Research Centre
Helsinki University of Technology
P.O.Box 5400, FIN-02015 HUT, Finland
Mikko.Kurimo@tkk.fi

Ebru Arisoy and Murat Saraclar

Bogazici University
Electrical and Electronics Eng. Dept.
34342 Bebek, Istanbul, Turkey
{arisoyeb,murat.saraclar}@boun.edu.tr

Abstract

Within the EU Network of Excellence PASCAL, a challenge was organized to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. Ideally, these are basic vocabulary units suitable for different tasks, such as speech and text understanding, machine translation, information retrieval, and statistical language modeling. Twelve research groups participated in the challenge and had submitted segmentation results obtained by their algorithms. In this paper, we evaluate the application of these segmentation algorithms to large vocabulary speech recognition using statistical n-gram language models based on the proposed word segments instead of entire words. Experiments were done for two agglutinative and morphologically rich languages: Finnish and Turkish. We also investigate combining various segmentations to improve the performance of the recognizer.

Index Terms: speech recognition, language modelling, morphemes, unsupervised learning.

1. Introduction

Segmentation is a common problem in the analysis of data from many modalities such as gene sequences, image analysis, time series, and segmentation of text into words. The task proposed here was to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. The purpose is to obtain a set of basic vocabulary units for different tasks, such as speech and text understanding, machine translation, information retrieval, and statistical language modeling [1, 2].

In many European languages this task is both difficult and necessary, due to the large number of different word forms found in text. In highly-inflecting agglutinative languages, such as Finnish and Turkish there may be thousands of different word forms of the same root, which makes the construction of a fixed lexicon for any reasonable coverage hardly feasible. Also in compounding languages, such as German, Swedish, Greek and Finnish, complex concepts can be expressed in one single word, which considerably increases the number of possible word forms and calls for the use of sub-word segments as vocabulary units.

The discovery of meaningful word segments has already shown to be relevant for language modeling for speech recognition in Finnish, Turkish and Estonian [2, 3], where language models based on statistically discovered sub-word units have rivaled language models that utilize words.

A good segmentation algorithm should be able to find units that are meaningful (that is, usable for representing text for many different tasks), that cover as much of the naturally occurring language as possible (including unseen words), and that can be used to generate the totality of the language. The field of linguistics has attempted to capture these properties by the concept of "morpheme", the difference being that a morpheme may not correspond directly to a particular word segment but to an abstract class. However, in this challenge the task was to uncover concrete word segments.

In obtaining such a segmentation, the use of linguistic analysis and manual coding may be an option for some languages, but not all, due to being very labor-intensive. Furthermore, statistical machine learning methods might eventually discover models that rival even the most carefully linguistically designed morphologies.

2. The Challenge

The task in *Morpho Challenge*[1] was the unsupervised segmentation of word forms into sub-word units (segments) given a data set that consists of a long list of words and their frequencies of occurrence in a corpus.

Data sets were provided for three languages: Finnish, English, and Turkish. Participants were encouraged to apply their algorithm to all of these test languages. Separately "tweaked" solutions for each test language were discouraged, since the aim of the challenge was the unsupervised (or very minimally supervised) segmentation of words into morphemes.

The segmentations were evaluated in two complementary ways: *Competition 1:* The proposed morpheme segmentation were compared to a linguistic morpheme segmentation gold standard. *Competition 2:* Speech recognition experiments were performed, where statistical n-gram language models utilized the proposed word segments instead of entire words. Competition 1 included all three test languages. Winners were selected separately for each language. As a performance measure, the F-measure of accuracy of discovered morpheme boundaries was utilized. Competition 2 included speech recognition tasks in Finnish and Turkish. The organizers trained a statistical language model based on the segmentations and performed the required speech recognition experiments. As a performance measure, the letter error rate in speech recognition was utilized.

[1] http://www.cis.hut.fi/morphochallenge2005

2.1. Data sets

The data sets provided by the organizers consisted of word lists. Each word in the list was preceded by its frequency in the corpora used. The participants' task was to return exactly the same list(s) of words, with spaces inserted at the locations of proposed morpheme boundaries.

The Finnish word list was extracted from newspaper text and books stored at the Language Bank of CSC[2]. Additionally, newswires from the Finnish National News Agency were used. The Turkish word list was based on prose and publications collected from the web, newspaper text, and sports news.

The desired segmentations, according to the gold standard, for a small sample of words (500 – 700 words) in each language were provided for download and inspection by the participants. For some words there were multiple correct segmentations.

The Finnish gold standard is based on the two-level morphology analyzer FINTWOL from Lingsoft, Inc. The Turkish linguistic segmentations were obtained from a morphological parser developed at Bogazici University [4, 5]. The Turkish parser is based on Oflazer's finite-state machines, with a number of changes.

2.2. The segmentations

By the deadline of the Challenge, 12 research groups had submitted the segmentation results obtained by their algorithms. Totally 14 different algorithms were submitted and 10 of them ran experiments on all three test languages (A1, A2a, A2b, A4a, A4b, A5, A7, A9. A11, A12a). More information about the submitted algorithms can be found in [6].

In addition to the competitors' algorithms, we evaluated a public baseline method called Morfessor (M1) [7] by the organizers[3] as well as its two more recent versions "Categories-ML" (M2) [8] and "Categories-MAP" (M3) [9]. Together with one of the challenge participants, Eric Atwell (Univ. of Leeds), the organizers also extended Atwell's original committee classifier algorithm "Cheat" [10] to utilize the segmentations of all (A12b) and the best five (A12c) of the other submissions in addition to only the segmentations from Univ. of Leeds (A12a).

Furthermore, for comparison, we used the gold standard segmentations and word transcriptions in our experiments. In both cases, the OOV words were dealt with in two ways: by splitting them into letters (G1 and W1) or by considering them as unknown words (G2 and W2).

3. Application to Speech Recognition

3.1. Evaluation Metrics

The segmentations provided by the participants were utilized to segment the words in large corpora of Finnish and Turkish text. An n-gram language model was trained for this segmentation and this language model used in speech recognition experiments.

Letter error rate (LER) was the main metric for the Competition 2 of the Challenge. We also used traditional Word Error Rate (WER) metric for the recognition results in this study.

One way to directly evaluate the accuracy of a language model is to compute the average probability of an independent test text. To obtain a useful comparison measure, this probability is normalized by the number of words in the text. Typical comparison measures derived from this normalized probability are *perplexity* and *cross-entropy*. For the competition we chose cross-entropy, which is the logarithmic version (log2) of perplexity.

Given the held-out text data T consisting of W_T words and a language model M, the *cross-entropy* $H_M(T)$ was computed as: $H_M(T) = -\frac{1}{W_T} \log_2 P(T|M)$. Here it is important that it is normalized by the number of *words*, not morphs, because a different morph lexicon was used for each model and, thus, the number of morphs in the test text varied.

In addition to accuracy, memory and time consumption are also important for real systems. We also report lexicon sizes and the recognition speed measured by the real-time factor (RTF).

3.2. Large-vocabulary continuous speech recognition systems

The objective of the competition was to evaluate the word splits in an application that would be as realistic as possible. Large-vocabulary continuous speech recognition was chosen, because of our own interests and previous experience in building morph-based recognition systems for Finnish, Estonian and Turkish [2, 3].

The speech recognizer consists of four main components: Acoustic phoneme models, language models, a lexicon and a decoder. The systems used in the experiments differ only in the vocabulary and the language model which were created from the word splits of each competition participant. The language models were trained by using exactly the same text corpus which was previously used for extracting the original word list that each competitor had processed as the competition entry.

Finnish. The Finnish speech data utilized for recognizer training and evaluation was exactly the same book reading corpus as in [2, 3]. The speaker-dependent reading recognition is not the most difficult large-vocabulary recognition task as can be seen from the rather low error rates obtained, but it suits well to the scope of the Finnish language model training data and has several interesting previous benchmark results.

For the acoustic models we chose the same speaker and context-dependent cross-word triphones with explicit phone duration models as for the Finnish models in [3] and also the same decoder. The real time factors were measured on 2.2 GHz CPU.

The Finnish newspaper, book and newswire training corpus included 40 M words and 1.6 M different word forms. After splitting the whole corpus into subwords and adding the word break symbols to assist the language model, n-gram language models were trained as if the units were word sequences. The language model used resembled the traditional n-gram model as used in [2], but instead of a fixed maximum value for n, the n was allowed to be optimized for each sequence context using the growing n-gram algorithm [11]. The idea in this approach is to start from unigrams and gradually add those n-grams that maximize the training set likelihood with respect to the increase of the model size. In addition to controlling the memory consumption for training and recognition, restricting the model complexity is important also to avoid over-learning, because natural language corpora are always very sparse, even if morph units are utilized.

In a complete speech recognizer there is an almost endless amount of parameter "tweaking" in order to tune the performance, speed, memory consumption, hypothesis pruning etc., not to mention the various parameters tuned for training the models. To save effort we adopted as much as possible the same parameters as in the previous works [2, 3, 11] even if they were perhaps not exactly optimal for the new models. The only parameter that we optimized

[2] http://www.csc.fi/kielipankki/.

[3] http://www.cis.hut.fi/projects/morpho/

Table 1: The obtained LM performance for Finnish and Turkish. CE is the average cross-entropy in the test text. OOV is the average out-of-vocabulary rate in the test text. OOV rates for Turkish were 0 except for G2 (0.19) and W2 (5.52). Size is the size of the lexicon.

	Finnish			Turkish	
	CE	OOV	Size	CE	Size
A1	13.65	0.36	297 981	16.80	121 942
A2a	13.54	0.03	73 178	15.32	48 619
A2b	13.63	0.04	65 557	15.99	37 253
A4a	13.55	2.70	609 458	16.51	204 555
A4b	*12.93*	0.99	1 559 199	16.39	561 905
A5	13.50	1.24	650 154	16.63	195 487
A7	13.81	0.85	530 543	16.24	189 239
A9	13.78	0.95	615 809	17.77	218 320
A11	13.59	0.58	690 601	*15.14*	264 502
A12a	13.66	0.40	317 870	16.55	148 650
M1	13.59	0.02	121 862	15.22	51 542
M2	13.53	0.08	155 065	16.27	96 182
M3	13.53	0.16	164 311	15.91	88 429
A12b	13.45	0.47	355 145	15.43	169 703
A12c	13.58	0.14	171 663	15.67	93 128
G1	13.62	0.03	69 929	15.20	23 680
G2	13.31	0.61	368 412	14.17	23 666
W1	13.95	0.00	394 266	12.53	120 001
W2	*12.04*	5.47	410 001	*10.02*	120 000

individually for each competitor was the weighting factor between the acoustic and language model. In order to achieve comparable models, the size of the language models was set to approximately 10 million n-grams.

Turkish. The main differences between the Finnish system and our Turkish large-vocabulary continuous speech recognizer were the speaker-independent acoustic models, the HTK[4] frontend and that no explicit phone duration models were applied. The acoustic training data contained 40 hours of speech from 550 different speakers. The Turkish evaluation was performed using the AT&T decoder[5] on a 2.4GHz CPU. The recognition task consisted of approximately one hour of newspaper sentences read by one female speaker.

In Turkish language model training corpus, there are totally 16.6 M words and 583 K different word forms. For language modeling and perplexity experiments, we used SRILM[6] to build 4-gram language models with interpolated modified Kneser-Ney smoothing. Entropy based pruning with a pruning constant of 10^{-8} was applied to each model to reduce the model size. The pruned model was used in the first pass to generate lattices which were then rescored using the full language models.

3.3. Language model evaluation

Table 1 shows the obtained cross-entropies on Finnish and Turkish test texts. For Finnish, a test text of 50,000 sentences was randomly selected from our text corpus and held-out from the training. Although the unsupervised morph lexicons were designed to process all words, there was a small OOV (out-of-vocabulary rate) in the test text. The OOV is shown in the table, because the higher it is, the more it affects the perplexity and cross-entropy by making it look smaller than it actually would be, if the OOV was zero.

For Turkish, the text of the test corpus consisting of 553 newspaper sentences (6989 words) was used. If the segmentation of a test word was available in the segmentation list, we split that word into the corresponding subwords. Otherwise, the test word was split into letters. In all of the submissions, the lexicon contained the individual letters of the Turkish alphabet as morphs. Therefore, the OOV rates were zero except for G2 and W2.

3.4. Speech recognition performance

The results of the speech recognition evaluation are shown in Table 2 (Finnish) and Table 3 (Turkish).

Table 2: The obtained speech recognition performance for Finnish.

Finnish	LER %	WER %	RTF
A1	1.42	10.58	17.67
A2a	1.39	9.53	12.88
A2b	*1.32*	*9.47*	15.92
A4a	*1.32*	9.81	15.59
A4b	1.64	13.54	10.89
A5	1.88	13.10	13.55
A7	1.55	11.33	13.97
A9	1.59	11.71	16.31
A11	1.45	11.17	*10.10*
A12a	1.40	10.72	15.65
M1	1.31	9.84	12.34
M2	1.32	10.18	14.38
M3	*1.30*	10.05	15.64
A12b	1.31	10.12	12.01
A12c	*1.25*	9.80	13.60
G1	1.33	9.60	10.58
G2	1.34	9.88	11.74
W1	1.37	10.83	11.84
W2	2.07	17.86	*7.42*

Table 3: The obtained speech recognition performance for Turkish.

Turkish	First Pass			Rescored	
	LER %	WER %	RTF	LER %	WER %
A1	15.0	43.0	2.68	11.6	33.1
A2a	13.6	38.9	2.15	*11.1*	*31.0*
A2b	*13.4*	*37.5*	2.19	11.8	32.7
A4a	15.7	46.3	2.43	11.3	32.0
A4b	16.7	50.2	*1.75*	12.2	36.3
A5	13.5	38.9	2.46	12.1	34.0
A7	13.8	40.3	2.33	11.3	32.5
A9	16.9	47.7	3.03	12.3	34.2
A11	14.6	41.4	1.85	11.8	33.6
A12a	14.5	41.9	2.56	11.4	32.4
M1	13.7	39.4	1.98	11.1	31.4
M2	14.3	41.2	2.10	11.6	32.8
M3	*13.2*	*37.2*	1.89	11.3	31.6
A12b	14.2	40.3	2.30	11.1	31.9
A12c	13.4	38.2	2.44	11.2	31.5
G1	12.0	33.1	2.03	11.4	31.4
G2	12.0	33.0	1.60	11.4	31.5
W1	12.3	33.9	1.98	11.8	32.8
W2	12.3	33.4	*1.40*	11.9	32.6

[4]http://htk.eng.cam.ac.uk/

[5]http://www.research.att.com/sw/tools/dcd/

[6]http://www.speech.sri.com/projects/srilm/

In the Finnish task, the winners were the models obtained from algorithm A2b and A4a. A2b was also the best system for the first pass in the Turkish task, but A2a was better in the final pass. The Morfessors M1, M2 and M3 were all very close to the winner in both tasks.

Since the best speech recognition error rates were not far apart, we performed pairwise statistical significance tests between every algorithm pairs. For the Finnish data we used the Wilcoxon's Signed-Rank test as in [2] and found that best Morfessor M3 was significantly better than M1, A9, A5 and A4b. The winners of the competition A2b and A4a were both significantly better than A12a, A11, A9, A7, A5, A4b and A1. For Turkish, we used the NIST MAPSSWE test on the final outputs and found A2a and M1 to be significantly better than A1, A2b, A4b, A5, A9, A11, and M2. For the case of WER, A2a is the clear winner.

3.5. Comparisons to previous methods

For Finnish, the gold-standard morphs (G1) and the word lexicon (W1) [2] seem to be very close in performance to the M1. However, if the OOVs (the words that cannot be segmented by the lexicon) are skipped as we did for other algorithms for Finnish, the error rates grow and cross-entropies shrink, especially for the word lexicon (W2) because of the much higher OOV rate than for any other model.

For Turkish, the performance of the gold-standard segmentations (G1) is close to the Morfessors and top segmentations in the competition. Although the word lexicon is slightly behind, this might be due to OOVs caused by the restricted vocabulary size.

3.6. Combination techniques

Inspired by one of the submissions (A12a) that combined different segmentations by voting, we used the segmentations of all (A12b) and the best five (A12c) of the submissions in order to get a better segmentation. This strategy did not improve the speech recognition performance significantly.

However, when we combined the recognition outputs based on the systems trained on different segmentations using ROVER [12], we obtained significant reductions in error rate. When used to combine word level outputs for Turkish, this technique yields 29.1% WER without an improvement in LER. Furthermore, using the combination at the level of letters while keeping the word boundary symbols to recover the words results in a LER of 10.1% which corresponds to a WER of 30.0% for Turkish.

For Finnish, the results follow the same trend. Using ROVER to combine the top five systems at the word level yields 8.0% WER with no improvement in LER, whereas combining the letter hypotheses gives a LER of 1.18% and a WER of 9.1%. Excluding the word boundaries results in a LER of 1.03%.

4. Conclusions

The objective of the *Morpho Challenge* was to design statistical machine learning algorithms for unsupervised segmentation of words into morphemes. In this paper, we evaluated the application of these segmentation algorithms to large vocabulary speech recognition using statistical n-gram language models based on the proposed word segments instead of entire words. We also investigated making use of multiple segmentation algorithms. For speech recognition purposes, combining recognition outputs is more effective than combining segmentations to get a better segmentation.

5. Acknowledgments

We thank all the participants for their submissions and enthusiasm. We owe great thanks as well to the organizers of the PASCAL Challenge Program who helped us organize this challenge and the challenge workshop. Our work was supported by the Academy of Finland. Funding was also provided by the Graduate School of Language Technology in Finland. We thank the Finnish Federation of the Visually Impaired for providing the speech data and the Finnish news agency (STT) and the Finnish IT center for science (CSC) for the text data. The authors would like to thank Sabanci and ODTU universities for the Turkish acoustic and text data and AT&T Labs – Research for the software. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

6. References

[1] P. Geutner. 1995. Using morphology towards better large-vocabulary speech recognition systems. In *Proc. ICASSP*, pages 445–448.

[2] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*. (In press).

[3] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar. 2006. Unlimited vocabulary speech recognition for agglutinative languages. In *Proc. HLT-NAACL*.

[4] Ozlem Cetinoglu 2000. Prolog based natural language processing infrastructure for Turkish. M.Sc. Thesis. Bogazici University, Istanbul, Turkey.

[5] Helin Dutagaci 2002. Statistical Language Models for Large Vocabulary Continuous Speech Recognition of Turkish. M.Sc. Thesis. Bogazici University, Istanbul, Turkey.

[6] M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy and M. Saraclar. 2006. Unsupervised segmentation of words into morphemes – Challenge 2005: An Introduction and Evaluation Report. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*. http://www.cis.hut.fi/morphochallenge2005/results.shtml

[7] M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. ACL/SIGPHON'02*, pages 21–30.

[8] M. Creutz and K. Lagus. 2004. Induction of a simple morphology for highly-inflecting languages. In *Proc. ACL/SIGPHON'04*, pages 43–51.

[9] M. Creutz and K. Lagus. 2005a. Inducing the morphological lexicon of a natural language from unannotated text. In *Proc. AKRR'05*, pages 106–113.

[10] E. Atwell and A. Roberts. 2006. Combinatory Hybrid Elementary Analysis of Text. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised segmentation of words into morphemes*.

[11] V. Siivola and B. Pellom. 2005. Growing an n-gram language model. In *Proc. Eurospeech*.

[12] J. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proc. ASRU*.