Speech recognition for huge vocabularies by using optimized sub-word units

Jan Kneissler, Dietrich Klakow

Philips GmbH Forschungslaboratorien
Weißhausstr.2, D-52066 Aachen, Germany
{jan.kneissler|dietrich.klakow}@philips.com

Abstract

This paper describes approaches for decomposing words of huge vocabularies (up to 2 million) into smaller particles that are suitable for a recognition lexicon. Results on a Finnish dictation task and a flat list of German street names are given.

1. Introduction

No vocabulary can cover all possible words of a language and just extending the size of the vocabulary of the speech recognizer to push the rate of unknown words beyond a reasonable limit is not a very elegant approach. Moreover, as the topic of the text to be recognized shifts, new words not considered before may become very important.

In addition for dialogue systems deployed in the Internet it is very desirable to have a domain independent first recognition stage that can recognize in principle any word from any domain and only a domain specific second stage extracts the relevant key words and concepts and ignores all other parts of the utterance not essential for the domain. This idea has been very much advocated by the MIT group [1, 2].

In [3] we have already discussed this issue from a purely language model point of view by clustering letters together to longer units (so-called “letter-phrases”) to create infinite-coverage vocabularies. Later, one aspect of the OOV-problem was addressed: how to reduce misrecognitions of words before or after an OOV by creating phoneme-based word fragments trained on “typical” unknown words [4]. No attempt was made to reconstruct words from the recognized sequences of phoneme-fragments.

This paper describes a novel approach to split low-frequency words into fragments that allow a reconstruction of proper words from a recognized sequence of fragments. This approach is applied to a Finnish dictation task which is particularly interesting because a vocabulary of 2 million words has the same OOV-rate as an English vocabulary of 64 thousand words. Also flat list of German street names where resource restrictions require a limited number of recognition units are briefly investigated.

2. Approaches

2.1. General scheme for recognizing with fragment lexica

In order to run a LVCSR system with a fragment lexicon, a few new items on the list of preparation and postprocessing steps have to be considered.

At first, there is the obvious question, of how a suitable fragment inventory should be constructed.

Then, for training a fragment language model, the words of the text corpus have to be cut accordingly (we will call this ”segmentation”). Here the question is how to choose one of the many possible segmentations.

Finally, for correctly gluing fragments together after recognition, the positions of word boundaries within the sequence of recognized fragments has to identified.

Obviously, for a fragment based approach, the size and coverage − the essential features of full-word lexica − become less significant (infinite coverages may be produced by tiny fragment lexica). The upper considerations make clear that there is a whole new set of criteria that presumably influence the recognition performance, due to the following sources for recognition errors:

- loss of phonetic context (if cross-fragment decoding is too expensive)
- loss of semantic context (the effective LM ranges are reduced),
- ambiguity (alternative word segmentations may appear as rivals in search space),
- wrongly positioned word boundaries.

2.2. Optimizing the segmentations

For all four tested fragment inventories, we used the same procedure for disambiguating the corpus segmentation. Its idea is to take the most probable word segmentation with respect to the frequencies of fragments. Since the fragment frequency distribution is not known in advance, we iteratively approximate it, starting with a zerogram:

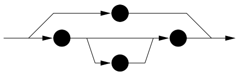
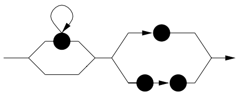
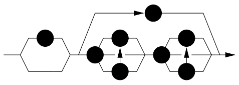
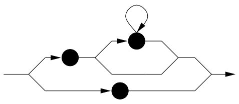
1. Assume an uniform fragment distribution.
2. Find a segmentation of the corpus with the maximal likelihood, based on the previously assumed / measured fragment distribution.
3. Take the unigram of this segmentation as new fragment distribution.
4. Repeat step 2 and 3 until the overall likelihood of the segmentation converges.

2.3. Reconstructing words

In order to recover full words, one has to classify all transitions between recognized fragments into those corresponding to within word transitions and those at word boundaries. This can be effectively done by either adding a tag to the first (or alternatively the last) fragment of each segmented word, or by grouping fragments into a set of classes and specifying the transition type for each possible class-to-class transition.

In both approaches, having multiple copies of some fragments in the lexicon (differing in their tags / belonging to different fragment classes) and letting the language model decide between these homophones appears to be inevitable.

Table 1: *Considered segmentation strategies*

Approach	Philosophy
max2cut 	Splitting into prefixes, optional infixes and suffixes.
pre+max1cut 	Words are split into stems and suffixes and can be preceded by an arbitrary number of prefixes.
morphological 	Words are split into stems and suffixes (by morph. knowledge) which are then cut further into one to three fragments.
phrases 	Some fragments carry the information of word beginnings, all other fragments are treated equally.

From there it is a natural step to consider more general systems for either tagging or class-based transition rules, that on one side restrict which fragment transitions are allowed and on the other side specify which of them represent word boundaries. Such systems will be called ”segmentation strategies”.

In this generality, there is no difference any more between the two approaches: one can show that every class based segmentation strategy is equivalent to a tagging system and vice versa.

2.4. Segmentation strategies

Segmentation strategies can be visualized by oriented graphs, where to some of the vertices (called states) a subset of the fragments is attached. A word segmentation is allowed, if the sequence of states corresponding to its fragments can be reproduced by an oriented path through the graph.

The graph must have exactly one source and one drain. States that can be directly (i.e. without passing other states) reached from the source are called ”initial” states; ”terminal” states are defined similarly. The obvious way to insure that word boundaries can be reconstructed, is to require that there exists no (non empty) oriented path from a terminal to an initial state.

Since there should be a way to represent short words and frequent words by a single fragment, it is also advisable to have one state that is both terminal and initial.

We have experimented with four segmentation strategies representing different splitting ideologies (table 1).

2.5. Variants to generate candidate fragments

While generally speaking, one has to take much care to insure that the fragments allow consistent transcriptions, the Finnish phonetics is quite cooperative in this regard. In fact a few graphemic constraints to the segmentations (e.g. ”thou shalt not split double letters!”) suffice to guarantee phonetic consistency. We even observed that ignoring these constraints did not lead to large performance losses.

For simple segmentation strategies there are straightforward ways of optimizing (at least locally) the fragment inventory at fixed size with respect to coverage.

The third strategy exploits morphological knowledge to split words at a first step into stems and suffix-tails, paying tribute to a special feature of the Finnish language (typically a combination of several grammatical units is appended to words). The sets of possible stems and suffix-tails are still quite large and thus treated in a way similar to the above described approaches.

The phrase approach is an extension of [3]. It starts with taking the most frequent 32 thousand words as initial fragments and splits all other words into letters. On this modified corpus standard phrase building is performed and all letter-phrases which are at least three letters long are added as fragments. The two-letter-phrases are inspected by our Finnish-language expert for being useful and pronounceable. Most two-letter-phrases do not pass this test and are not added to the fragment list. Now the complete vocabulary is tested for being splitable and on the non-splitable part of the corpus we start again the procedure just described. This iteration is done until the desired number of fragments is reached.

3. Experimental Setup

The basis for the Fragment generation and language modeling is a 50 million word corpus of Finnish texts (newspaper archives and internet sources). Interestingly enough, there are 2.6 million different word forms in this corpus, which illustrates the necessity for a fragment based approach in this case. Perplexities have been evaluated on an independent corpus of 1.4 million words.

The acoustic model consists of CART-clustered triphones (40 phonemes) with 70k densities, trained on 100 hours material spoken by several female speakers. The recognition tests are evaluated on a 2-3h test set (444 sentences), after 15 minutes MAP+MLLR adaptation.

4. General properties of the different variants

A first idea of the advantages and disadvantages of the different approaches can be obtained when looking at elementary statistics of the corpus segmented by a given method. For technical reasons the number of fragments of the pre+max1cut approach is significantly smaller as compared to the other approaches. The average length of the fragments shows only a very small fluctuation. In contrast we observe very strong variation in OOV-rate and the very restrictive max2cut and pre+max1cut approaches show rather high OOV-rates. This is eased by the more flexible morphological approach and the purely data driven phrase construction which during training aimed at creating fragments that can split words unsplitable so far indeed shows the lowest OOV-rate. The column with the average number of fragments per word indicates where the different approaches spent their effort. A smaller number shows that the more frequent words are treated more accurately. Similarly the number of one-fragment words in the corpus is higher for such approaches. Finally the fraction of uniquely segmentable words confirms that, as expected, less restrictive segmentation strate-

Table 2: *Comparison of segmentations according to different strategies. The second column gives the total number of fragments considered, the third column the average number of phonemes per fragment (weighted by the frequencies in the corpus). The last but one column gives the fraction of words that don't need to be split into fragments and the last column the fraction of uniquely segmentable words.*

Method	#fragments	av. Length	OOV-rate	av. #frag./word	1-frag. words	uniq. segm. words
max2cut	55393	4.42	7.0%	1.50	51%	70%
pre+max1cut	40932	4.48	7.7%	1.46	56%	54%
morphological	56303	4.54	4.7%	1.47	59%	37%
phrases	51668	5.14	2.4%	1.34	82%	43%

gies exhibit higher segmentation ambiguity.

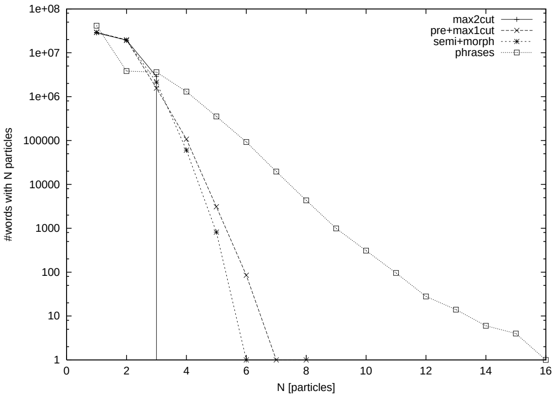


Figure 1: *Logarithmic distribution of segmentation lengths.*

In Fig. 1 the distribution of the words that need N fragments to be split is given. In the max2cut approach each word is split by one, two or three fragments by construction. pre+max1cut and the morphological approach behave similarly and split words into up to seven units. Phrases are different in two respects. A much higher fraction of words (50% more) is not split. This is also a consequence of the construction algorithm which started with taking the most frequent 32 thousand words as fragments. The distribution shows also a very long tail which explains the low OOV-rate of this approach. If it encounters a difficult word it can still be split at the price of using many fragments. This may cause problem as the LM-context on word-level completely breaks down.

5. Perplexities and recognition experiments

The recognition performance of the different approaches has been compared for different acoustic models and speakers. The typical results are summarized in table 3.

We have decided that the individual transition structure should be learned by the language model itself, instead of putting hard LM-constraints according to the segmentation strategies. As a result, the sequence of recognized fragments may contain forbidden transitions between fragments. As the third column of table 3 shows, this effect is quite negligible for bi- and trigram models.

The data of the fourth column indicates that our disambiguating algorithm has been quite successful. Even for the strategies for which segmentations of most of the words are not unique, less than 2% of the words are recognized correctly, but using an alternative segmentation. It should be noted, that only for the phrases approach, these percentages increase with language model range M, which indicates that in this case a bigram-criterion for the disambiguating algorithm of section 2.2 might be useful.

The letter error rates are surprisingly low. For a typical US-English dictation task, we measured a letter error rates of 6.4% at a word error rate of 10.6%. This strengthens our impression that for a language that packs whole phrases into single words, the word error rate is not the appropriate measure.

Perplexities have been measured on a segmentation of an independent corpus, and were then normalized w.r. to the differing numbers of fragments by the formula:

$$\mathrm{Perplexity}_{\mathrm{words}} \quad := \quad (\mathrm{Perplexity}_{\mathrm{fragments}})^{\frac{\#\mathrm{fragments}}{\#\mathrm{words}}}$$

In order to eliminate the effect of the differing OOV-rates, we also excluded all words (10.7%) that were not segmentable by all four strategies from perplexity measurement. The normalized trigram perplexities on the commonly segmentable corpus agree almost perfectly.

Only for the phrases approach, the average segmentation lengths differ significantly on the full test set and the common in-vocabulary part. The explanation is that uncommon words are split using high numbers of fragments (see section 4).

For all approaches the word error rates are much lower on the common in-vocabulary part than on the full test set. In fact, the error rates on the removed part of the corpus are close to 100% for all approaches, no matter what the actual coverage (ranging from 28% to 77%) on this part was (note that since long words may be misrecognized as several shorter words, we also removed up to four insertions around excluded words from the error statistics).

6. Outlook

In spite of the fact that not many languages are as generative as Finnish, and do not necessarily require more than 100k words for a typical dictation task, the subword unit approach may be useful in some other circumstances.

For instance for long item list recognition applications might not be run on a relatively small platform (or in some cases might not be run at all) unless a very limited lexicon is used.

To explore the usefulness of our approach to this type of applications, we experienced with it on a German street name recognition task. The segmentation characteristics are compared to Finnish in table 4.

As described in section 2.5, the Finnish segmentations are done at the level of graphemes; the atomic units correspond to

Table 3: *Comparison of recognition results (using M-gram LMs with $M = 1, 2, 3$). Columns 3 to 5 show the percentages of forbidden transitions, the rate of segmentation deviations sequences and the letter error rates. The last six columns list the average splitting lengths, perplexities and word error rates on the full test set and restricted to words that are segmentable by all strategies.*

		evaluated on full test set						splitable by all (89.3%)		
Method	M	wr. trans.	dev. seg.	LER	#frag./w.	w-Perp.	WER	#frag./w.	w-Perp.	WER
max2cut	1	57.9	2.0	10.5	1.46	76.87k	63.6	1.42	41.38k	58.8
	2	14.7	0.5	6.3		1.98k	35.4		1.28k	27.4
	3	13.6	0.6	5.9		1.28k	32.3		0.83k	24.1
p+max1cut	1	48.0	1.9	9.6	1.44	62.99k	56.3	1.41	38.47k	49.0
	2	9.6	1.4	6.3		1.84k	32.5		1.28k	20.8
	3	9.0	1.5	5.9		1.16k	30.2		0.81k	18.2
morph.	1	44.7	2.4	9.4	1.44	80,71k	55.9	1.38	32.34k	49.4
	2	11.2	1.5	5.8		2.40k	32.2		1.27k	22.6
	3	10.9	1.5	5.5		1.51k	29.8		0.82k	20.5
phrases	1	6.2	0.4	7.9	1.30	55.73k	47.3	1.15	10.55k	32.7
	2	0.5	1.4	5.9		2.91k	33.9		1.07k	21.5
	3	0.5	1.8	5.4		1.89k	30.6		0.80k	19.1

Table 4: *Comparison of segmentations based on phrases for Finnish dictation texts and a German street name list.*

	Finnish dictation	German street names
words for fragm. gener.	3M	11k
fragments	52k	1k
#atomic units	42	125
av. #phonemes / fragment	5.1	2.8
av. #fragments / word	1.3	4.2
OOV rate (train mat.)	2.4%	0.0%
OOV rate (test mat.)	3.4%	2.4%
OOV rate (full word list)	3.0%	57%

the 28 letters, a few text formatting characters, and combinations of letters due to phonetic constraints (mostly double letters).

For the German street names task, we used an alignment of graphemes and transcriptions to produce the set of atomic units. Each atom is a combination of sequences of graphemes and phonemes. There are more atomic units, but the numbers of phonemes/letters per unit (weighted by frequencies) are almost identical to those in Finnish (namely 1.0/1.1).

The specifications for the street name segmentation were rather hard; we insisted on full coverage and 100% correct transcription, at a reduction of lexicon size by a factor of 10. Accordingly the constructed fragments were shorter and the majority of words are split into 3-5 fragments.

The last three rows of table 4 indicate that the fragment inventories generalize quite well for both tasks, compared to the full word approach (that simply puts all seen words into a lexicon).

7. Conclusion

Four different approaches to segment words into shorter fragments are demonstrated. Depending on the desired target function (OOV-rate, WER, LER) one or another segmentation strategy comes off as winner. Nevertheless, the differences between the variants are rather small, which lets us conjecture that, in general, speech recognition using lexica of subword units is feasible − even without spending too much linguistic expertise for the concrete realization. This notion is further supported by successfully applying the phrases based approach to a German street name task.

8. Acknowledgment

The authors would like to thank Henrik Botterweck, Hans Dolfing and Jochen Peters for many stimulating discussions.

9. References

[1] Bazzi, I. and Glass, J., "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigation", Proc. ICASSP (2000), Vol. 3, 1257-1261.

[2] Seneff, S., "The Use of Linguistic Hierarchies in Speech Understanding", Keynote at ICSLP (1998).

[3] Klakow, D., "Language-Model Optimization by Mapping of Corpora", Proc. ICASSP (1998), Vol. 2, 701-704.

[4] Klakow, D., Rose, G. and Aubert,X., "OOV-Detection in a Large Vocabulary System Using Automatically Defined Word-Fragments as Fillers", Proc. EUROSPEECH (1999), Vol. 1, 49-53.