Compound Word Recombination for German LVCSR

Markus Nußbaum-Thom, Amr El-Desoky Mousa, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6 - Computer Science Dept.
RWTH Aachen University, Aachen, Germany
{nussbaum,desoky,schlueter,ney}@cs.rwth-aachen.de

Abstract

Compound words are a difficulty for German speech recognition systems since they cause high out-of-vocabulary and word error rates. State of the art approaches augment the language model by the fragments of compounds in order to increase lexical coverage, lower the perplexity and out-of-vocabulary rate. The fragments are tagged in order to concatenate subsequent equally tagged fragments in the recognition result, but this does not guarantee the recombination of proper words. Such recombination techniques neglect the large vocabulary of the language model training data for recombination although most compounds are covered by it. In this paper, we investigate the use of this vocabulary for the recombination of compound words from the recognition result. The approach is tested on two large vocabulary tasks on top of full-word and fragment based language models and achieves good improvements of 3–7% relative over the baseline compound-sensitive word error rate.

Index Terms: speech recognition, compound words, German LVCSR, sub-lexical

1. Introduction

Compound words are part of the complex German morphology and make up most of the German words even though their frequency is low. A simple example for a compound in English is the word "darkroom" which is composed of the word "dark" and "room", but German allows more complex compositions like "Donaudampfschifffahrtshauptbetriebsgesellschaftsbeamter" which is a not too realistic concatenation of eight words. In general there is probably an infinite number of compound words in German. Hence, the language model training data for German large vocabulary tasks often exceeds several million distinct words. Due to data sparseness, it is not possible to estimate higher order language models as needed for large vocabulary continuous speech recognition systems (LVCSRs) for such a large number of words. The vocabulary for state of the art LVCSR is chosen by the cut-off on the most frequent words in order to reduce the expected out-of-vocabulary (OOV) rate and as well as the data sparseness to enable the estimation of higher order language models. On the other hand, this reduction changes most German compounds into out-of-vocabulary words. Therefore, approaches dealing with the OOV problem also need to handle the compound word problem.

Most approaches dealing with the compound word or more general the OOV problem decompose compounds into smaller fragments in order to increase lexical coverage, lower the language model perplexity and OOV rate: For example the methods reported in [1, 2, 3, 4] decompose compounds into sub-words by supervised word-splitting, while [5, 6, 7, 8, 9, 14] use unsupervised word splitting. In [10] words are decomposed into graphones, which are short sequences of characters, using the statistical grapheme-to-phoneme conversion toolkit [13]. Other approaches convert words into even smaller units like multi-phoneme sequences as in [12] or single phonemes as in [8, 11].

When using fragment based language models, the fragments have to be recombined in the recognition result in order to obtain words, but the properness of the words can not be guaranteed. A number of publications [2, 3, 6, 14] show that a WER improvement can be achieved by not decomposing the most frequent compounds. In [7] the reconstruction into compounds from sub-words is obtained by recombining the most frequent bigrams and trigrams into compounds. While, most approaches like [3, 4, 8, 14] add tags to the fragments in order to recombine subsequent fragments with the same tag in the recognition result. The tagged fragments are integrated as regular words into the language model and the decision about subsequent tags is left over to the language model.

The recent sub-lexical approach [14] based on unsupervised word splitting was chosen as representative for the fragment based language models in our recombination experiments.

The complete vocabulary of the language model training data is a knowledge source which is neglected by most approaches in literature so far for the recombination of fragments into proper words, although it covers several million compounds for a large German vocabulary task. In this work, an approach is investigated which recombines compounds from a given recognition result by using this large vocabulary. In literature [2] recombines compounds in lattices with a manual generated list of compounds, but this leads to an increase in word error rate (WER). In contrast to [2] the recombination process of our approach will be restricted to the recognition result, no explicit list of hand selected compound words is used for recombination, but all words seen in the language model training data. In addition the recombination of certain compounds will be forbidden if one of the words being recombined is very frequent. Later, we try to carry over the insights from experiments on single best recognition result to lattices.

In short, our approach can be summarized as follows: A lattice is generated by recombining compounds from the recognition result using a large vocabulary. Then the lattice is rescored with a unigram language model and the most likely compound sentence in the lattice is chosen as recognition result. The proposed approach was tested on the QUAERO large vocabulary task for different sized full-word and sub-lexical language models and showed good improvements of 3–7% relative over the baseline compound-sensitive WER.

The remainder of this paper is organized as follows: In Section 2 we introduce the details of our approach. In Section 3 the experimental setup is described, while in Section 4 the experimental results are presented and discussed. The paper concludes with Section 4.

2. Compound Word Recombination

Statistical speech recognition chooses the most likely word sequence $w_1^N := (w_n)_{n=1}^N$ as recognition result given an acoustic observation. In case an OOV compound word is spoken an automatic speech recognition system tends to recognize the more frequent sub-words instead. Therefore, the compound word can not be recognized without further processing. Here, the compounds will be recombined in a three step procedure: A lattice $\mathcal{L}(w_1^N)$ is generated from the recognition result. Then the lattice $\mathcal{L}(w_1^N)$ is rescored with a unigram language model and finally the most likely compound sentence is decoded as recognition result.

In the first step, the lattice generation, the recognition result can be considered as a single path lattice. A new word arc labeled with a compound is added to the lattice if a number of subsequent words can be recombined into a compound like shown in Figure 1. Notice German written numbers are also compounds and can be identified via a regular expression match for numbers up to a certain degree, for example up to trillions.
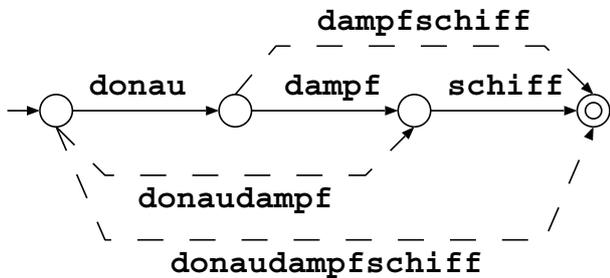

Figure 1: An example for a compound lattice generation from the recognition result "donau dampf schiff" (indicated with solid lines) in which the compounds "donaudampf", "dampfschiff" and "donaudampfschiff" (indicated with dashed lines) can be recombined.

For this recombination the compound vocabulary V and the discard vocabulary D (both to be defined later) are needed. The recombination of the word sequence $w_i^j := (w_i, \ldots, w_j)$ into the compound $w_{i,j} := w_i \circ \ldots \circ w_j$ ($\circ$: concatenation) is accepted iff:

1. $w_{i,j}$ is a number (matchable by regular expression) or,
2. $w_{i,j} \in V$ (seen in language model training data) and $w_i, \ldots, w_j \notin D$ (reject compound if it has a very frequent sub-word).

In a nutshell the condition 2 means that the new compound is covered by the language model training data and none of its sub-words is a very frequent word. The maximal number of words M being recombined into a compound is limited for efficiency reasons.

The compound vocabulary V covers all words seen in the language model training data. Some words like "beiden" (English: both) can be composed of very frequent sub-words like "bei" (English: at) and "den" (English: the). The unigram language model rescoring produces, in case that a word like "beiden"' is recombined, relations like:

$$p(\text{"beiden"}) > p(\text{"bei"}) \cdot p(\text{"den"})$$

Thus it is very likely that "beiden" will be decoded instead of "bei den" which is an incorrect decision in most cases. There-

fore, the discard vocabulary consists of the r most frequent words in V in order to refuse word recombinations which have very frequent sub-words. From now on we will refer to r as the frequency cut-off. Notice that r has to be tuned on a development set.

In the second step, the constructed lattice is rescored with a unigram language model which is estimated from the complete language model training data for the vocabulary V. It is not reasonable to favor some numbers for recombination over others, because the cardinality of numbers is infinite and they carry no unique semantic information about material things like other words. Therefore, the probability mass of numbers is redistributed equally among all numbers seen in the language model training corpus. The maximum likelihood estimate is used for non-numbers. Let be $C(w)$ the count of a word, C the total count of all words, C_n the total count of all numbers and W_n the number of all numbers seen in the language model training data. Then the lattice is rescored with the following unigram probability distribution:

$$p(w) = \begin{cases} \frac{C(w)}{C} & , w \text{ is not a number} \\ \frac{C_n}{C} \cdot \frac{1}{W_n} & , w \text{ is a number} \end{cases} \quad (1)$$

In the third step, Bayes decision rule is applied to the compound search space in the form of the lattice $\mathcal{L}(w_1^N)$ by choosing the most likely compound sentence as recognition result:

$$\text{opt}(w_1^N) = \underset{v_1^I \in \mathcal{L}(w_1^N)}{\text{argmax}} \left\{ \prod_{i=1}^{I} p(v_i) \right\} \quad (2)$$

The scheme presented above can easily be extended from recombination on a single recognition result to recombination on recognition lattices. This extension is straightforward since the recombination process presented so far, has just to be applied to each path in the lattice. The extension is implemented via a depth first search on the lattice, where for each lattice node the histories up to a certain length M are generated consecutively (also via depth first search on the transposed lattice) and used for recombination. When recombining a compound word the acoustic and language model scores of the words being concatenated are added up and assigned to the new compound word arc. In the lattice rescoring the language model probabilities in the lattice are linearly interpolated with the unigram language model probability. In the following all recombination experiments will use recombination on the recognition result only (unless stated explicitly). Later, we try to carry over the insight from recombination experiments on single recognition results to recombination on recognition lattices.

3. Experimental Setup

The proposed compound word recombination approach was tested on the German 2009 and 2010 QUAERO web, broadcast news and broadcast conversation task. For each task the baseline recognition results were obtained with a task dependent acoustic model while the lexicon and language model was altered.

The baseline systems have a two pass architecture. In the first pass a fast-VTLN method is applied followed by a speaker adaptation in the second pass using Constrained Maximum Likelihood Linear Regression (CMLLR) based on the first pass output.

All language models, as well as the compound vocabulary and the unigram language models used for recombination on

the QUAERO 2009 (QUAERO 2010) task were estimated on 306M (QUAERO 2010: 500M) running words covering 2.5M (QUAERO 2010: 3.8M) distinct words. The recombination approach was evaluated on the QUAERO 2009 and 2010 development and evaluation sets (dev09: 7.5h, eval09: 3.8h, dev10: 3.5h, eval10: 3.5h). A more detailed description of the language model training and test data is given in [14, 15, 16] since a complete description is beyond the scope of this paper.

The baseline recognition results for QUAERO 2009 were obtained using $100\,\text{k}$, $200\,\text{k}$, $300\,\text{k}$ full-word language models and a sub-lexical language model with $5\,\text{k}$ full-words and $95\,\text{k}$ fragments. The full-word and sub-lexical recognition results used here are the baseline systems reported in [14].

The official QUAERO 2009 scoring is compound-insensitive, since compound words are mapped to their sub-words for scoring by manually created mapping rules. In order to investigate the real potential of compound recombination the scoring script was changed to compound-sensitive by deleting the mapping rules. Due to this change in the scoring script WERs reported here differ from the WERs reported in [14, 15].

The baseline recognition results for the QUAERO 2010 task were produced with a $300\,\text{k}$ full-word and a sub-lexical language model comprised of $300\,\text{k}$ full-words and $95\,\text{k}$ fragments. The acoustic model of the baseline recognition results is the second pass of the best German sub-system described in [16].

4. Experimental Results

In the first series of experiments numbers were recombined only in order to distinguish between the contribution to the WER of recombining numbers and non-numbers, followed by recombination experiments for non-numbers.

We first investigated the relationship of the WER to the maximal number of words M being recombined. All experiments of this type showed WER curves similar to Figure 2 where a stabilizing behavior after recombining at most $M = 6$ subsequent words into a number could be observed.
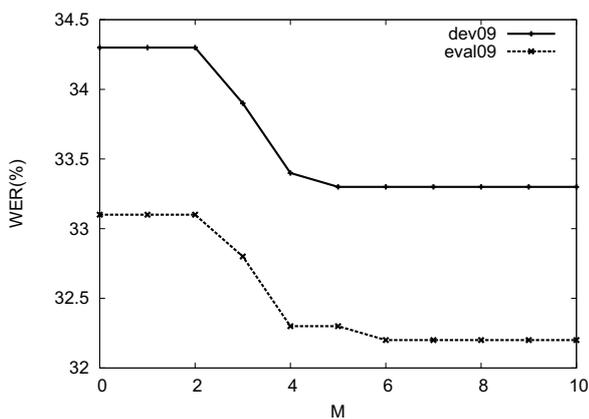

Figure 2: Progress of the WER vs. the maximal number M of recombined words (recombination of numbers only) obtained with a $100\,\text{k}$ full-word lexicon on the QUAERO 2009.

The same saturation effect was also observed for the recombination of non-numbers. In the conducted experiments, the value $M = 10$ has shown to be sufficient in order to achieve the maximal WER gain for all recombination experiments and was kept constant at this value for all later experiments. In sum-
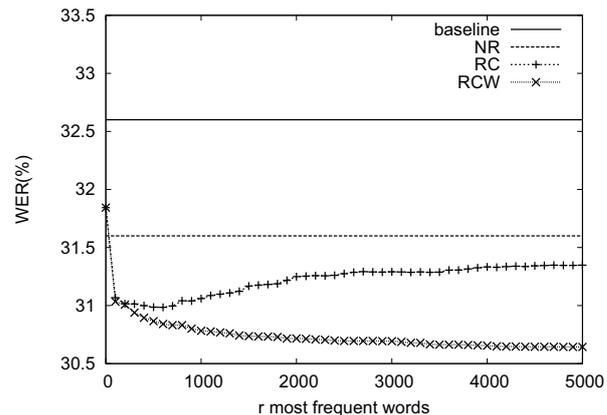

Figure 3: Progress of the WER vs. the discard vocabulary size which consist of the r most frequent words. The baseline WER is incrementally enhanced by using number recombination (NR), compound word recombination (RC), and compound word recombination using the improved discard vocabulary (RCW). The baseline was produced on the QUAERO dev09 development corpus using the $300\,\text{k}$ the full-word language model.

mary, the recombination of numbers only already achieved a consistent WER improvement of 2–5% relative over the baseline WER as shown in Table 1 in the column using the abbreviation NR for number recombination.

In the next series of experiments non-number compounds were recombined in addition to the number recombination. We used the following recombination techniques and abbreviations:

NR: Recombination of numbers only.

RC: NR + recombination of non-numbers using a discard vocabulary which consists of the r most frequent words.

RCW: RC + using a discard vocabulary which consist of the r most frequent words which have a contribution in lowering the WER.

The size of the discard vocabulary, which consists of the r most frequent words, was optimized in the range of 0 and 5000 with step sizes of 100 on the dev09 and dev10 development corpora with the WER as objective function. The method RCW uses a more sophisticated construction of the discard vocabulary in which for each of the 5000 most frequent words the contribution to the WER is measured by using only this single word as discard vocabulary. The final discard vocabulary was built then only of the r most frequent words which lead to a WER decrease in comparison to the WER for using no discard vocabulary. For all recombination experiments, WER curves were observed similar to Figure 3.

Table 1 shows the WER results for all recombination experiments. In summary a relative WER reduction of about 3–7% depending on the vocabulary size is achieved for all compound methods recombining both numbers and non-numbers. In general, number recombination reduces the WERs in all experiments. Further, the WER gain for the non-number recombination experiments diminishes with increasing lexicon size. Surprisingly, the $300\,\text{k}$ full-word language model ends up in the same WER as the sub-lexical model, composed of $5\,\text{k}$ full-words and $95\,\text{k}$ fragments after compound word recombination, although the sub-lexical model starts with a lower WER.

Table 1: WERs for number recombination only (NR), compound recombination using (RC), and compound word recombination using the improved discard vocabulary (RCW). The baseline recognition results were generated on the QUAERO test sets for different sized full-word (*full*) and sub-lexical (*frag*) lexica.

corpus	lexicon		WER			
	full	+frag	baseline	NR	RC	RCW
dev09	100 k	-	34.3	33.3	32.2	31.8
	200 k	-	33.1	32.1	31.4	31.1
	300 k	-	32.6	31.6	31.1	30.8
	5 k	95 k	32.1	31.1	31.0	30.7
eval09	100 k	-	32.7	31.8	30.9	30.7
	200 k	-	31.4	30.5	29.9	29.8
	300 k	-	30.8	30.0	**29.4**	**29.4**
	5 k	95 k	30.5	29.5	**29.4**	**29.4**
dev10	300 k	-	21.7	20.5	19.9	19.4
	300 k	95 k	20.8	19.8	19.8	19.5
eval10	300 k	-	19.7	19.0	**18.9**	**18.9**
	300 k	95 k	19.1	18.4	**18.4**	**18.4**

Next, the generalization ability of the compound word approach was tested. The requirements for a generalization ability are met, when the parameters, in terms of the discard vocabularies and the cut-off r which have been optimized on a specific development corpus for a specific lexicon, carry over to other test corporas and lexica. In order to confirm the generalization ability, the compound vocabulary and the tuned discard vocabulary which were tuned on a specific task were used on all other test corpora in combination with all remaining lexica for compound word recombination. In these recombination experiments the WER fluctuated only negligible by 0.1% absolute. Therefore, the generalization ability can be approved.

Finally, compound word recombination on lattices was tested using a discard vocabulary based on frequency cut-off. In addition to the frequency cut-off the language model interpolation factor had to be tuned in addition. Although an improvement in WERs over the non-compound baseline was measured, the WERs in all experiments turned out to be slightly worse than for single best path and a constant increase in WER was observed for increasing lattice density. In summary the compound word recombination on the recognition result should be preferred over the recombination on lattices.

5. Conclusions and Outlook

An approach was investigated for German LVCSR for the recombination of compound words from a given recognition result by using the vocabulary of the complete language model training data for recombination. The approach was tested on the QUAERO 2009 and 2010 large vocabulary tasks with good improvements of 3–7% relative WER reduction over the baseline recognition result on top of a full-word and fragment based language models. We found evidence that the compound word problem can be handled in a relative simple and efficient postprocessing and that compound words can also be recovered from the recognition result produced with a regular full-word language model, in contrast to the very laborious generation of fragment based language models.

Future work includes the generalization the proposed compound word recombination approach to other inflective languages like Finish.

7. References

[1] M. Adda-Decker and G. Adda, "Morphological decomposition for ASR in German", in Workshop on Phonetics and Phonology in ASR, Saarbrcken, Germany, Mar. 2000, pp. 129-143.

[2] A. Berton, P. Fetter and P. Regal-Brietzmann, "Compound words in large-vocabulary German speech recognition systems", in Proc. Int. Conf on Spoken Language Processing, Philadelphia, PA, USA, Oct. 1996, vol. 2, pp. 1165-1168.

[3] L. Lamel, A. Messoudi and J.L. Gauvain, "Investigating morphological decomposition for transcription of Arabic broadcast news and broadcast conversation data", in Interspeech, Brisbane, Australia, Sept. 2008, vol 1, pp 1429-1432.

[4] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units", in Proc. European Conf. on Speech Communication and Technology, Aalborg, Denmark, Sept. 2001, vol 1, pp 69-72.

[5] M. A. Adda-Decker, "A corpus-based decompounding algorithm for German lexical modeling in LVCSR", in Proc. European Conf. on Speech Communcation and Technology, Geneva, Switzerland, Sept. 2003, pp 257-260.

[6] R. Ordelman, A. V. Hassen and F.D.Jong, "Compound decomposition in Dutch large vocabulary speech recognition", in Proc. European Conf. on Speech Communication and Technology, Geneva, Switzerland, Sept. 2003, pp. 225-228.

[7] M. Larson, D. Willett, J. Köhler and R. Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches", in Proc. Int. Conf. on Spoken Language Processing, Bejing, China, Oct. 2000.

[8] M. Creutz, et al., "Morph-based speech recognition and modeling of out-off-vocabulary words across languages", ACM Transactions on Speech and Language Processing, vol. 5, no.1, Dec. 2007.

[9] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models", in Proc. European Conf. on Speech Communication and Technology, Geneva, Switzerland, Sept. 2003, pp. 434-451s.

[10] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models", in Interspeech, Lisbon, Portugal, Sept. 2005, pp. 725-728.

[11] I. Bazzi and J. R. Glass, "Modeling out-of-vocabulary words for robust speech recognition", in Proc. Int. Conf. on Spoken Language Processing, Bejing, China, Oct. 2000.

[12] D. Klakow, G. Rose and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers", in Proc. European Conf. on Speech Communication and TEchnology, Budapest, Hungary, Sept. 1999, vol. 1, pp. 49-52.

[13] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", Speech Communication, vol. 50, no. 5, May 2008, pp. 434-451.

[14] A. El-Desoky Mousa, H. Ney et. al., "Sub-Lexical Language Models for German LVCSR", in IEEE workshop on Spoken Language Translation, Dec. 2010, pp. 803-806.

[15] M. Nußbaum-Thom, H. Ney et. al., "The RWTH 2009 Quaero ASR Evaluation System for English and German", in Interspeech, Brighton, UK, Sept. 2010, pp. 1517-1520

[16] M. Sundermeyer, H. Ney et. al., "The RWTH 2010 QUAERO ASR Evaluation System for English, French, and German" accepted for publication at ICASSP, Prague, Czech, May 2011.