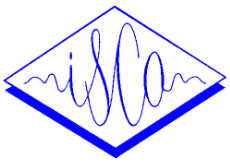
ISCA Archive
http://www.isca-speech.org/archive



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parliamentary speeches

Martha Larson[1,2], *Daniel Willett*[2,†], *Joachim Köhler*[1], *Gerhard Rigoll*[2]

[1]IMK: Institute for Media Communication
GMD German National Research Institute for Information Technology
Sankt Augustin, Germany
larson@gmd.de, joachim.koehler@gmd.de

[2]Department of Computer Science
Faculty of Electrical Engineering, Duisburg University
Duisburg, Germany
willett@fb9-ti.uni-duisburg.de, rigoll@fb9-ti.uni-duisburg.de

ABSTRACT

This paper proposes a novel combined compound splitting and phrase recombination method that optimizes the composition of the speech recognition lexicon for a given domain. Data-driven compound word splitting is followed by iterative recombination of high frequency combinations. Language model perplexity and size are the criteria used to identify a balance between compound decomposition, which reduces OOV, and lexical unit recombination, which packs additional context into a fixed-size vocabulary. The method provides a basis for lexicon design for a LVCSR system on the domain of German parliamentary speeches that is to be used as the foundation of a spoken document information retrieval system. The approach achieves a 35% reduction in OOV without a prohibitively large sacrifice in recognition performance.

1. INTRODUCTION

The convention of adopting the orthographic word as the basic unit in the LVCSR lexicon is not suited to handling so-called compounding languages like German, Dutch, Norwegian, Danish, Swedish and Greek, in which complex concepts are expressed as single words. In order to achieve the same OOV rates as non-compounding languages, compounding languages must use a much larger recognition lexicon [1][2]. Even if its size has been stepped up considerably, an orthographic word-based speech recognition lexicon lacks the flexibility to cover new compounds or compounds coined by the speaker on the fly.

The generally accepted remedy is that for compounding languages, orthographic words must be split into linguistically meaningful sub-units in the recognition lexicon [1][2][3]. Splitting lexicon entries, however, robs them of their inherent contextual content and exacerbates acoustic confusability. In fact, many authors have taken exactly the opposite tactic and have reported WER improvement gained from combining orthographic words into phrases in the recognition lexicon [4][5][6]. The work presented in this paper aims to provide a principled approach to optimizing the design of a speech recognition lexicon both with respect to compound decomposition and to phrase recombination.

The goal of our research is the design of a recognition lexicon to be used in the spoken document information retrieval system for German parliamentary speeches currently under construction at the GMD. Access to sub-orthographic semantic units is essential to effectively tackle the information retrieval task. Decomposing compounds in the speech recognition system not only controls OOV explosion, but also provides a possibility for a more tightly knit integration between the LVCSR and the IR systems. Decomposition in the extreme, however, can be detrimental to recognition accuracy. We offset the potential harm by combining compound splitting with lexical unit recombination.

Standard approaches to Indo-European languages have applied a morphology-based splitting procedure to identify compound constituents for use in the recognition lexicon. Many approaches have split compounds across the board [3][7], making the tacit assumption that all splittings are equally utilitarian. Our approach is grounded on a different premise, namely, that lexicon design is a domain-specific pursuit and that the best determinant of which elements should be used as lexical units in the recognition lexicon is the training data itself.

In this paper we first introduce our data-driven algorithm that splits compounds according to the statistical relevance of the resulting constituents. The result is our baseline split system, which uses a lexicon of maximally split units. Then we present the iterative method that recombines the lexical units of the maximally split system to arrive at three experimental systems. We are interested in determining the effects of lexicon redesign on the system, so are careful to hold all other factors constant. The experimental systems are all bigram language models with a lexicon size of 50,000 words and the same training corpus and represent different operating points between low OOV rates and high language model quality. As a measure of LM quality we use the perplexity of the model with respect to an unseen test set as well as the number of bigrams in the language model whose

† now with: NTT Communication Science Lab, Kyoto, Japan

probabilities can be estimated directly from the data. We present the recognition experiments and their results and finish with a discussion of the tradeoffs between coverage and recognition rates.

2. DATA-DRIVEN DECOMPOSTION

Our approach is founded on the assumption, also made in [8], that the lexicon should be tailored to the recognition task domain. Splitting of compounds according to morphological rules may guarantee that the resulting constituents are at some level linguistically real, but does not directly insure the relevance of these constituents for the speech recognition task. Morphology-based rules aim to capture semantic relationships as they exist in the language as a whole, many of which may not be relevant for the design of a domain-specific language model. Moreover, a morphology-based splitter may not be readily available, or may not be able to handle exceptions or coinages. Such considerations enhance the appeal of a data-driven approach to compound decomposition. This paper draws inspiration from statistically-based methods of word identification in Asian languages, where orthographic convention does not dictate word boundaries [8] [9][10][11].

By completely disassociating the lexical unit in the recognition lexicon from the orthographic word, we are also able to address two technical difficulties facing German LVCSR systems that opt for purely morphological decomposition. The first is the problem that compounding is derivationally often more complex than a simple juxtaposition of two independent word forms. In German, the initial compound sub-unit has a special compounding form, which defies synchronic analysis as either a plural or a genitive, despite surface similarities. The most prevalent way of deriving a compounding form from an independent form is the addition of a suffix, commonly *-s*, *-es*, *-n* or *-en*, but a system using morphological decomposition that handles compound words by adding a list of compounding suffixes to the recognition lexicon suffers in several areas. For one, compounding suffixes are short and therefore introduce added acoustic confusability. Additionally, some compounding forms are derived not by affixation, but rather by vowel changes or truncations. Thus, from the independent words *Sprache*, "speech" and *Erkennung*, "recognition", *Spracherkennung* and not *Spracheerkennung* is formed. Lastly, such a system fails to exploit the useful generalization that each independent word has a uniquely determined compounding form. These considerations have motivated us to model initial compound sub-units as separate entries in the recognition lexicon. If the splitting algorithm indicates that *Friedenspolitik*, "peace policy" is to be split, the resulting compound sub-units are `friedens_` and `politik`. `friedens_` is modeled as a compound subunit and is not conflated with the independent word `frieden`, "peace". The latter part of the compound, `politik`, "politics" or "policy" is, however, conflated with the independent word `politik`. Modeling initial compound sub-units, but not final compound sub-units, as independent lexical items is consistent with the linguistic generalization that in German the final compound subunit is the head of the compound, meaning that *Friedenspolitik* is a kind of *Politik* and not a kind of *Frieden*, and that *Politik* and *Friedenspolitik* would be more likely to share similar contexts than *Frieden* and *Friedenspolitik*. The second difficulty facing German LVCSR systems using morphological decomposition is how to recompose the words in order to end up with an orthographically correct recognizer output. By modeling non-final compound sub-units separately, the recognizer output can be uniquely assembled into orthographic words.

The data we use is a corpus of 15 million words transcribed in the German parliament. The text has been leveled to lower case ASCII and abbreviations and numbers have been replaced with the corresponding written forms. Our data-driven splitting algorithm iteratively generates splittings that are statistically relevant with respect to the training corpus. For each distinct word of the training set we generate an array, illustrated in Table 1, tabulating for each letter of that word the total number of words in the training set which began or ended with the same sequence of letters, i.e. with the same letter-based n-gram.

f	r	i	e	d	e	n	s	p	o	l	i	t	i	k
-	-	39	29	29	25	24	23	3	1	1	1	1	1	1
1	1	1	1	1	2	7	37	88	89	89	92	99	-	-

Table 1: Example compound word `friedenspolitik` "peace policy" with the counts of how many total words in the training corpus have the same beginning letters (middle row) and ending letters (bottom row).

We eliminate a priori splittings which would result in constituents three letters or shorter, drawing inspiration from [2], under the assumption that such short lexical units would introduce a detrimental level of acoustic confusability. For this reason, no counts are generated for the first two letters of the word from the forwards direction and for the last two letters of the word from the backwards direction. Moving from left to right (Table 2, middle line) and from right to left (bottom line) we calculate how much the counts fall off as the comparison n-gram becomes longer (the count ?'s).

f	r	i	e	d	e	n	s	p	o	l	i	t	i	k
-	-	-	10	0	4	1	1	20	2	0	0	0	0	0
0	0	0	0	0	5	30	51	1	0	3	7	-	-	-

Table 2: The deltas: How much the counts change moving from one letter to the next**.**

A potential splitting juncture is any point at which the ?'s reach a local maximum, marked with an asterisk in Table 3.

f	r	i	e	d	e	n	s	p	o	l	i	t	i	k
-	-	-	-		*			*						
							*				-	-	-	-

Table 3: The delta max's: asterisks indicate the points at which the deltas reach local maxima.

A word is split only when the splitting juncture is associated with a local change maximum both from the left-to-right (forward) and the (right-to-left) backward direction. In the example we see that decomposition of `friedenspolitik` into the string `friedens_` and `politik` is generated by the algo-

rithm. In this particular example, the constituents correspond with those which would be generated by a rule-based splitter. In fact, most splittings generated by our algorithm are consistent with the morphology. Lexical units do occasional arise, however, which are not linguistically self-sufficient enough to be independently pronounceable, and therefore cannot be assigned a phonetic transcription. We control this error in the system by filtering out those splittings that are rejected by the phonemization process. At the core of the phonemization process is an algorithm which uses combination and subtraction strategies of known phonemizations supplied by the comprehensive basic vocabulary in its input lexica. To break compounds which consist of more than two sub-constituents into their components, we apply the splitting algorithm iteratively. More than three iterations produces only a marginal number of additional splittings, however.

The splitting algorithm outputs a list of splittings which is used to split the orthographic words in the baseline training text, generating the baseline split text, in which compounds have been split into maximally small parts. The baseline split language model was generated using the 50K most frequently occurring "words" in the baseline split training text. As all of the language models presented in this paper it is a Katz-style back-off bigram language model with Good-Turing discounting and was trained using the CMU-Cambridge Language Modeling Toolkit [12].

3. PHRASE RECOMBINATION

The baseline split system that was a result of the compound splitting is the starting point for the iterative recombination that determines which segments are to be considered words in the optimal experimental system. We recombine lexical units in the maximally split system in order to arrive at a language model with a fixed 50K vocabulary that has been optimized with respect to perplexity measured with on an unseen data set of 1.6 million words. The perplexity of an experimental language model ($PP_{exp.LM}$) is measured by taking the reciprocal of the geometric mean of the probability that the language model assigns to the experimental test text corresponding to that model [13]. The number of words in the test text varies from one experimental system to the other as words are split and recombined, and it is therefore necessary to normalize the perplexity with respect to the number of words in the baseline unsplit test text.

$$PP_{exp.LM} = [P(w_1,w_2...w_{N\text{exp. test set}})]^{-1/N\text{baseline test set}}$$

The number of OOV words in the test text also varies from one experimental system to the next and for this reason we are careful to exclude OOV in the test set from the perplexity calculation. Since the perplexity calculation is resource intensive, we do not recalculate it for every possible word recombination, but rather establish classes of candidates based on co-occurrence frequency. Such an approach has been demonstrated to be effective in [6][8]. Infrequent events will have little impact on recognition performance and we want to also avoid over-learning which is caused by training on insufficient data. First, bi. and trigrams occurring more than 100 times were redefined to be words in their own right. We combine these bi- and trigrams in the training text and recalculate a lexicon containing the most frequently occurring 50K words in the training text. Then, bi- and trigrams occurring more than 150, 200, 300 etc. times in the text were also merged to words. The perplexities of each of the resulting language models with respect to the test set are depicted in Figure 1.

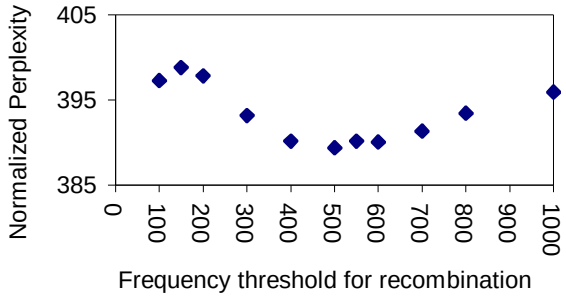


Figure 1: $PP_{exp.LM}$ for LM's of different thresholds

We discovered that if bi- and trigrams occurring 500 times or more are redefined to be words, the resulting LM built on a vocabulary of the most frequent 50K words of the new system has a minimum of perplexity with respect to the test data.

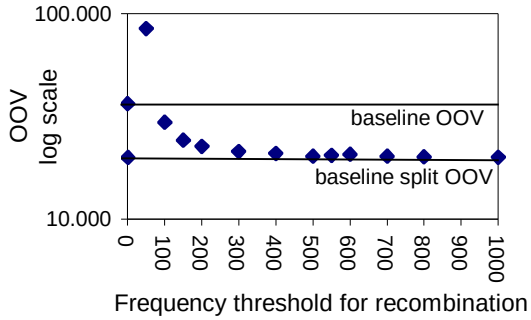


Figure 2: Effective number of OOV in baseline test

As can be observed in Figure 1, the experimental model in which bi- and trigrams occurring more frequently than 100 times were recombined to form words also achieved a local minimum with respect to perplexity. This model does not represent an ideal operating point, however, since it does not achieve a minimum OOV rate, as can be seen in Figure 2.

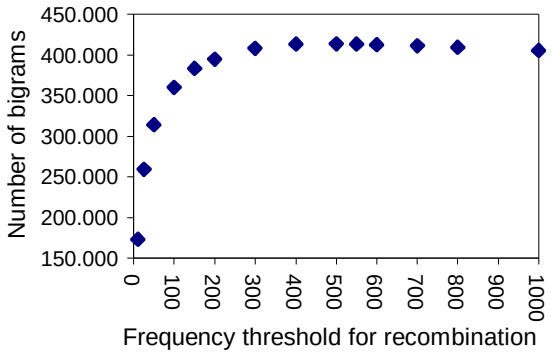


Figure 3: Number of bigrams in experimental LM's

To supplement perplexity we use a second measure of language model quality, namely the number of bigrams that the language model attains. As the frequency threshold for recombination increases, the language model built on a vocabulary of the most frequently occurring "words" contains more and more bigrams, as can be observed in Figure 3. A maximum is attained at the 400 threshold model. If the amount of training text is fixed, redesign of the lexicon makes it possible to estimate more bigram probabilities, presumably resulting in a higher quality language model.

4. EXPERIMENTS AND EVALUATION

The recognition experiments, whose results are presented in Table 4, were performed on the audio test set, a subset of the test set of 200 recorded sentences containing a total of 4.5K words. We ran the recognition experiments using the Duisburg University Decoder, [14], using tree-base clustered word-internal 3-state triphones and speaker-adapted acoustic models.

model	OOV rate	Accuracy
baseline unsplit system	2.35%	65.39%
baseline split system	1.35%	63.08%
100x's and more recombined	2.86%	58.53%
500x's and more recombined	1.53%	61.89%
1000x's and more recombined	1.42%	61.58%

Table 4: OOV rates vs. recognition results on audio test set

The language model in which bigrams and trigrams occurring 500 times and more were combined to words represents an optimal lexicon design. The 500x threshold language model outperformed the 100x and the 1000x threshold models and demonstrated a 35% reduction in OOV rate. The OOV rates and the accuracy rates given in Table 4 are expressed in terms of orthographic words and not in terms of the new lexical units of the experimental systems. Normalizing the results on the baseline unsplit test set in this way is necessary in order to permit a meaningful comparison between the systems. As mentioned in [2], reporting recognition rates with respect to the orthographic word may obscure the performance of the decoder, since the result no longer represents a one-to-one correspondence between mis-recognized lexical units and decoder error. The results do not demonstrate conclusively that recombination is able to compensate for the loss of context which leads to the deteriorated recognition rate in the split system.

5. CONCLUSION AND OUTLOOK

This paper has linked compound splitting with a related area of lexicon design, word recombination, and presented a data-driven approach to the principled composition of a recognition lexicon, optimized with respect to the perplexity criteria and size of the language model. Although the approach turns out to be well suited for increasing the text coverage of a fixed-size lexicon, in the real-world speech recognition experiments the recombined models trail the baseline system in performance with respect to recognition accuracy. A reexamination of our phonemization process holds promise of an explanation for this unexpected result, since automatic phonemization of split words and recombined word sequences poses some difficulties that were treated by our automatic phonemization algorithm only in a first approximation. Both the effectiveness of the data-driven splitting algorithm for OOV reduction and the potential of our optimization criteria have, however, been well established and will provide fertile ground for future investigation.

6. REFERENCES

1. Young, S.M., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J-L., Kershaw, D.J., Lamel, L., Leeuwen, D.A., Pye, D., Robinson, A.J., Steeneken, H.J.M. and Woodland, P.C. "Multilingual large vocabulary speech recognition: the European SQALE project", *Computer Speech and Language*, Vol. 11, pp. 73-89, 1998.
2. Berton, A., Fetter, P. and Regel-Brietzmann, P. "Compound Words in Large-Vocabulary German Speech Recognition Systems," *ICSLP*, pp. 1165-1168, 1996.
3. Spies, M. "A Language Model for Compound Words in Speech Recognition," *Eurospeech*, pp. 1767-1770, 1995.
4. Zitouni, I. Mari, J.F., Smaïli and K. Haton, J.P. "Variable-length Sequence Language Model for Large Vocabulary Continuous Dictation Machine," *Eurospeech*, pp. 1811-1814, 1999.
5. Kou, H-K. J. and Reichl, W. "Phrase-based Language Models for Speech Recognition," *Eurospeech*, pp. 1595-1599, 1999.
6. Klakow, D. "Language-Model Optimization by Mapping of Corpora," *ICASSP*, pp. 701-704. 1998.
7. Carter, D., Kaja, J., Neumeyer, L., Rayner, M., Weng, F. and Wirén, M. "Handling Compound Nouns in a Swedish Speech-Understanding System," *ICSLP*, pp. 26-29, 1996.
8. Hwang, K. "Vocabulary Optimization Based on Perplexity," *ICASSP*, pp. 1419-1422, 1997.
9. Ito, A. and Kohda M. "Language Modeling by String Pattern N-gram for Japanese Speech," *ICSLP*, pp. 490-493, 1996.
10. Kiecza, D., Schultz, T. and Waibel, A. "Data-Driven Determination of Appropriate Dictionary Units for Korean LVCSR," *Proceedings of the International Conference on Speech Processing (ICSP)*, pp. 323-327, 1999.
11. Ando, R.K. and Lee, L. "Mostly-Unsupervised Statistical Segmentation of Japanese: Applications to Kanji," *ANLP-First Conference of the NAACL*, pp. 241-248, 2000.
12. Clarkson P. and Rosenfeld R. "Statistical Language Modeling using the CMU-Cambridge Toolkit," *Eurospeech*, pp. 2707-2710, 1997.
13. Ueberla, J. "Analysing a Simple Language Model – Some General Conclusions for Language Models for Speech Recognition," *Computer Speech and Language*, Vol. 8, pp. 153 – 176, 1994.
14. Willett, D., Neukirchen, C. and Rigoll, G. "DUcoder – The Duisburg University LVCSR Stackdecoder," *ICASSP*, pp. 1555-1558, 2000.