Fitting German into N-Gram Language Models

Robert Hecht, Jürgen Riedler, and Gerhard Backfried

Speech, Artificial Intelligence, and Language Laboratories
Operngasse 20B, A-1040 Vienna, Austria
{robert,juergen,gerhard}@sail-technology.com

Abstract. We report on a series of experiments addressing the fact that German is less suited than English for word-based n-gram language models. Several systems were trained at different vocabulary sizes and various sets of lexical units. They were evaluated against a newly created corpus of German and Austrian broadcast news.

1 Introduction

The performance of an ASR system is to a large extent determined by how well the language model and vocabulary fit the speech to be recognized. The recognition lexicon should contain most of the words likely to appear, i.e. a minimum out-of-vocabulary (OOV) rate should be achieved, as low lexical coverage leads to high word error rates (WER). State-of-the-art ASR systems typically utilize full form word lexica and language models (LM). However, languages like German exhibit a large variety of distinct lexical forms. Thus, a morphological decomposition of orthographic words could improve OOV rates trading-off with larger WER due to neglection of coarticulatory effects and shorter span of language model contexts.

Amongst the main mechanisms for creating new words in German are such diverse elements as inflection, derivation, and compounding. Inflections and derivations are typically formed with a limited number of short affixes. Compounding is a very productive process in which the tendency to convey information in a compressed form leads to accumulation of more and more components [1]. New words are formed spontaneously and frequently occasionalisms emerge which will not be seen in any texts used to train language models. German orthography requires their constituents to be written as one word, thus resulting in a potentially unlimited set of compounds.

Remedies for strongly inflecting and compounding languages roughly divide into two branches: morphology-based [2, 3] and data-driven [4] compound word splitting, both often accompanied by grammar models other than simple n-grams (e.g. class-based or long-distance models).

In the broadcast news domain, [2] focussed on identifying and processing the statistically most relevant sources of lexical variety. A set of about 450 decomposition rules, derived using statistics from a large text corpus together with partial inflection stripping reduced OOV rates from 4.5% to 2.9%.

The study described in [3] is based on the Verbmobil database consisting of over 400 human-to-human dialogs. Decomposition along phonological lines reduced vocabulary size as well as perplexity, which in turn lead to a more robust language model. In a further step, suppression of decomposition of very frequent compounds yielded lowest WERs.

In [4] it is claimed that splitting of compounds according to morphological rules into linguistically meaningful sub-units does not ensure relevance of these constituents for the speech recognition task (in this case for German parliamentary speeches). Inspired from statistically-based methods of word identification in Asian languages, where orthographic convention does not dictate word boundaries, a data-driven approach to word splitting was examined. Splitting junctures within words were determined by counting the words in the corpus that start and end with the letter sequence under consideration. A 35% reduction in OOV rates was achieved. However, the recombined models trail the baseline system in performance.

2 Corpora

2.1 Acoustic Model

Our training (test) set see consists of 175 (9) TV news broadcasts from six German and one Austrian channel, recorded between November 1999 and March 2000. They comprise 60:54 (4:22) hours of raw data and 47:42 (3:36) hours of speech. Transcription was done using the Transcriber tool [5] whose output is a textual version of the transcripts with markup for speaker turns, speaker names, genders, and non-speech events.

2.2 Language Model

The language model (LM) consists of the audio transcripts (490k words) and texts from newspapers and journals from 1994 through 1999 (300M words). In the LM training, the audio transcripts were emphasized by a weighting factor.

The normalization performed on the LM texts included:

- Slang and dialect words were replaced by their Standard German correspondent (if it was possible to find one). Their pronunciation was kept.
- Cardinals were spelled out and split into "smallest possible" parts, e.g. 324 was transcribed as *drei hundert vier und zwanzig*.
- Ordinals: in German, ordinals are marked by a period after the number, e.g. 324. meaning 324th. This is problematic because German ordinals can be spelled out in six different forms, e.g. *dritter, dritte, drittes, drittem, dritten, drittens* (third). Since there is no way to find the correct ending without syntactical analysis, we chose to spell out the non-terminal parts, but to write the last component (which carries the ending) as a number followed by a period, e.g. 324. → *drei hundert vier und 20.* In the decoding dictionary, we listed all possible pronunciations for the entries 1., 2. etc.

- Words containing both numbers and letters were split, e.g. *16jähriger* $\rightarrow$ *sechzehn jähriger* (16 year-old).
- Case Normalization was applied, since nouns and nominalized words are capitalized in German.
- Abbreviations were expanded where possible.
- In 1998 the official German orthography was re-standardized. Part of our language model texts still followed the old German orthography and thus had to be converted to be consistent with the rest of the system.

3 Experiments

3.1 ASR system

A speaker-independent, large-vocabulary speech recognizer based on BBN's Rough'n'Ready suite of technologies [6–8] is used in our experiments. It employs continuous density Hidden Markov Models (HMM) for acoustic modeling and word-form n-grams for language modeling. The recognition process is implemented as follows: at first a cepstral analysis of the acoustic waveforms is performed to extract a feature vector for each frame of speech. These feature vectors serve as input for three decoder passes. The first (fast-match) pass uses a phonetically tied-mixture (PTM) model based on within-word triphones and a bigram language model. The second (backward) pass employs a state-clustered tied-mixture (SCTM) model depending on within-word quinphones and a trigram language model to produce a set of N-best hypotheses. Then a third, more detailed analysis using a cross-word SCTM model is used to rescore the N-best list to select the final answer.

3.2 Specifics of German

German has a number of peculiarities that make it less suited for n-gram modeling than e.g. English. Among these are:

- compound words (see Section 3.3)
- highly inflective nature: leads to bigger vocabularies and to difficulties to resolve word forms that sound similar
- flexible word order
- dialects: German has a large number of local varieties, making it difficult to build a system which can cope with speech from different regions.

3.3 Compound Words

German is a compounding language, capable of expressing complex concepts in a single word. A compound word consists of one terminal component which is qualified by one or several non-terminal components, e.g. in *Baumstamm* (tree trunk) the terminal component *Stamm* is qualified by the word *Baum*. The non-terminal components often end in suffixes, the most common ones being *-s*,

-es, *-n*, and *-en*. Sometimes the non-terminal components are truncated, as in *Sachkenntnis* (expertise), where the first part is derived from *Sache* (matter). Compounding can be applied to several classes of words: mostly nouns, but also verbs, adjectives and adverbs.

The abundance of compound words means that for any fixed dictionary size, the OOV rate is much higher than on comparable systems in English (typically for the BN domain and vocabularies of 65k words this number is less than a percent). On the other hand, it is not advisable to arbitrarily inflate the dictionary to improve the coverage, because too many entries will lead to higher WER because of confusion of similar sounding (rare) words.

We used the following simple approach to decompose them into their constituents: a decomposition of a word in the corpus was considered valid if

- all constituents were in the dictionary
- all constituents were at least four letters and four phonemes in length
- the case of the last constituent matched the case of the compound
- at least one combination of the pronunciations of the constituents matched a pronunciation of the compound.

The first rule was necessary to check if the fourth was fulfilled. The second rule is motivated by the fact that too short fragments lead to poor recognition results. The third rule has a grammatical background: since in German all nouns are capitalized, the rule essentially demands that if the compound is a noun, the terminal component also has to be a noun and vice versa. This is an attempt to filter out splittings which violate basic word-forming rules. The last rule was introduced to guard against parts with identical spelling but different pronuncations.

This procedure splits compound words mostly in accordance with morphological rules. However, the non-terminal components of identical spelling and pronunciation were sometimes "mixed-up", e.g. *Anfangspunkt* (starting point) was split into *anfangs Punkt*, using the word *anfangs* (initially) instead of the correct decomposition *Anfang* + suffix *-s*.

Because of their abundance, non-terminal constituents with suffix *-s* were treated separately. If a word could not be split, we checked if there was an allowed split of the form $(w_1 + s + w_2)$, where (w_1) and (w_2) were in the dictionary. If this split was successful (according to our rules), the word $(w_1 s)$ was added to the dictionary.

This procedure was able to decompose approximately 231k of the 559k entries of our dictionary. Fig. 1 shows a reduction of OOV rates for vocabulary sizes ranging from 50k to 300k entries. For a 50k-word vocabulary we obtain an improvement of almost 32%, for 100k words we still gain 28%. Of course, the improvement becomes smaller with increasing vocab size, so the effect is particularly noticable when the vocabulary size is limited.

In the audio and language model training texts, all compounds were replaced by the constituents into which the above procedure had split them. The language model was retrained and the vocabulary adjusted to the newly decomposed words. Following recognition, all sequences of words that could be the

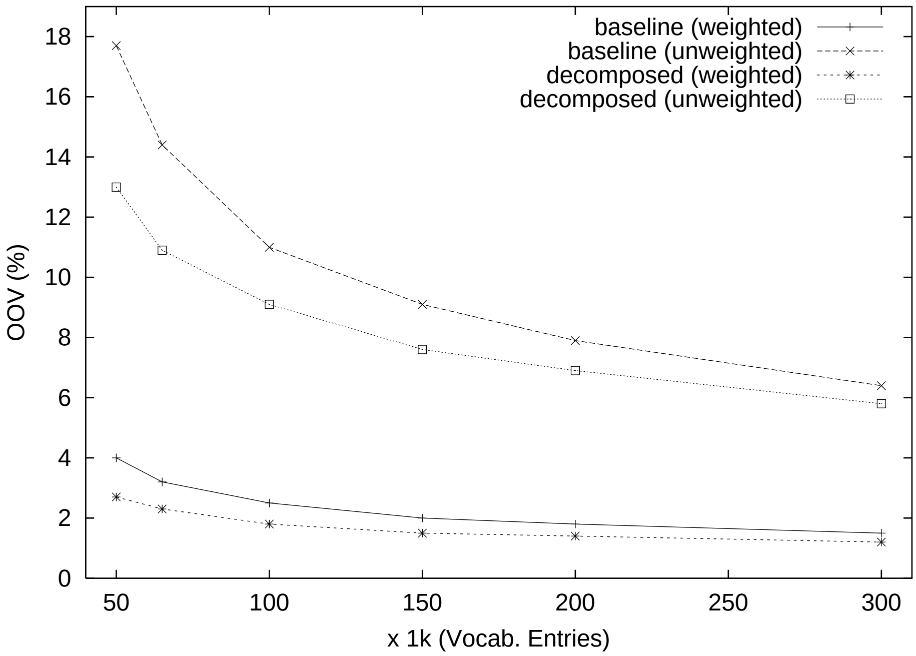

Fig. 1. OOV rates versus vocabulary size for the baseline and the decomposed system. Unweighted percentages take only unique words into account whereas the weighted numbers are calculated using total word counts.

result of a split compound word were recombined into one word. This procedure is somewhat simplistic, but it gave good results for vocabularies with approximately 50k words.

We also investigated an alternative approach in which all non-terminal components of a compound word were marked. Following the recognition step, the word following a marked word was appended to it and the casing adjusted. This procedure has the advantage that it is capable of generalization: it can recognize compound words that were never seen in the training material, if all components are in the recognition lexicon.

As Fig. 2 depicts, decompounding improves the recognition rate mainly at moderate vocab sizes. At 50k words, the simple (list-based) approach gave an improvement of almost 1%, while the approach with the marked non-terminal components leads to an improvement of 1.6%.

4 Summary

We created a corpus of German broadcast news of a variety and size which supercedes existing material in this domain [2, 9]. From this corpus several BN

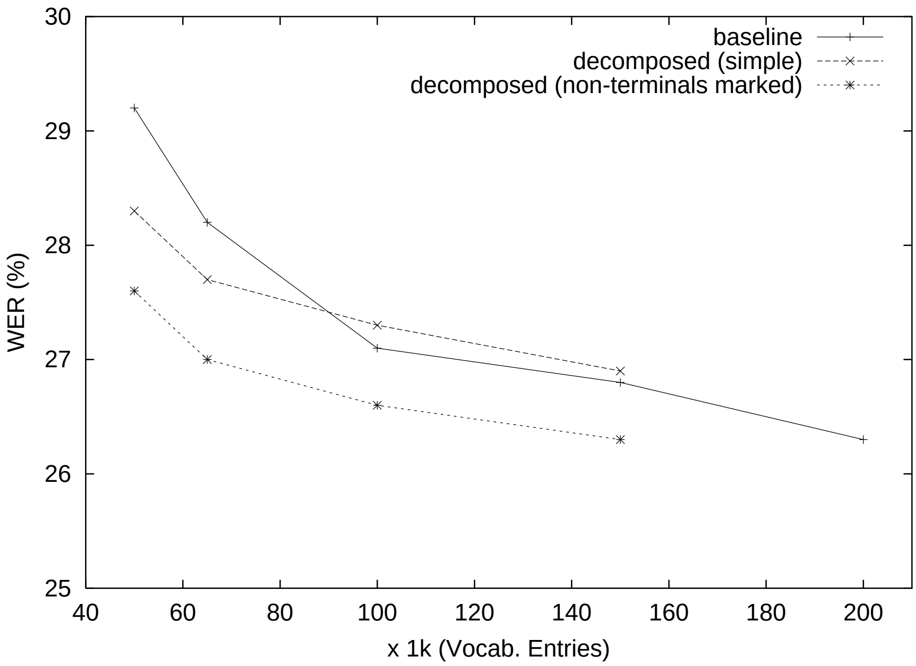

Fig. 2. WER rates for the baseline and the various decomposed systems.

transcription systems were trained and evaluated at different vocabulary sizes. Various schemes of decomposition were implemented to account for inflection and *ad hoc* formation of compound words in German. The benefit in word error rate we gained by decomposition of vocabularies of up to 65k entries seems to be annihilated at larger vocabulay sizes (200k). This may be due to lack of language model context and training and increased acoustic confusability. For the future we would like to examine the observed trends for even larger vocabulary sizes. We will measure the impact on run-time as the resulting models are meant to be used within the Sail Labs real-time BN indexing environment [10].

References

1. Stedje, A.: Deutsche Sprache gestern und heute. UTB für Wissenschaft, München (1999)
2. Adda-Decker, M., Adda, G., Lamel, L.: Investigating Text Normalization and Pronunciation Variants for German Broadcast Transcription. Proceedings of the 6th International Conference of Spoken Language Processing, ICSLP'2000 (Beijing, China)
3. Larson, M., Willett, D., Köhler, J., Rigoll, G.: Compound Splitting and Lexical Unit Recombination for Improved Performance of a Speech Recognition System for German Parliamentary Speeches. Proceedings of the 6th International Conference of Spoken Language Processing, ICSLP'2000 (Beijing, China)

4. Berton, A., P. Fetter, P., Regel-Brietzmann, P.: Compound Words in Large-Vocabulary German Speech Recognition Systems. Proceedings of the 4th International Conference of Spoken Language Processing, ICSLP'96 (Philadelphia, Pennsylvania, USA)
5. Barras, C., Geoffrois, E., Wu, Z., M. Liberman, M.: Transcriber: Free Tool for Segmenting, Labeling and Transcribing Speech. Proceedings of the 1st International Conference on LREC (1998)
6. Kubala, F., Jin, H., Nguyen, L., Schwartz, R., Matsoukas, S.: Broadcast News Transcription. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'97 (Munich, Germany)
7. Nguyen, L., Matsoukas, S., Davenport, J., Liu, D., Billa, J., Kubala, F., Makhoul, J.: Further Advances in Transcription of Broadcast News. Proceedings of the 6th European Conference on Speech Communication and Technology, EuroSpeech'99 (Budapest, Hungary)
8. Colthurst, T., Kimball, O., Richurdson, F., Han, S., Wooters, C., Iyer, R., Gish, H.: The 2000 BBN Byblos LVCSR System. Proceedings of the 6th International Conference of Spoken Language Processing, ICSLP'2000 (Beijing, China)
9. Kemp, T., Geutner, P., Schmidt, M., Tomaz, B., Weber, M., Westphal, M., Waibel, A.: The Interactive Systems Labs View4You Video Indexing System. Proceedings of the 5th International Conference of Spoken Language Processing, ICSLP'98 (Sydney, Australia)
10. Backfried, G., Hecht, R., Loots, S., Pfannerer, N., Riedler, J., Schiefer, C.: Creating a European English Broadcast News Transcription Corpus and System. Proceedings of the 7th European Conference on Speech Communication and Technology, EuroSpeech'01 (Aalborg, Denmark)