Compound decomposition in Dutch large vocabulary speech recognition

Roeland Ordelman, Arjan van Hessen and Franciska de Jong

Department of Computer Science, University of Twente, The Netherlands
{ordelman,hessen,fdejong}@cs.utwente.nl

Abstract

This paper addresses compound splitting for Dutch in the context of broadcast news transcription. Language models were created using original text versions and text versions that were decomposed using a data-driven compound splitting algorithm. Language model performances were compared in terms of out-of-vocabulary rates and word error rates in a real-world broadcast news transcription task. It was concluded that compound splitting does improve ASR performance. Best results were obtained when frequent compounds were not decomposed.

1. Introduction

Decomposition of compound words in compounding languages such as Finnish, German and Dutch was addressed earlier in a number of studies in the context of spelling checking, information retrieval (IR) and automatic speech recognition (ASR). For spelling checking, compound splitting can be useful when the compound is not in the spelling lexicon ([1]). In IR, decomposing compounds can be used for query expansion as to improve retrieval recall ([2],[3]). The main purpose of compound splitting in the ASR domain is reducing lexical variability and thereby the number of out-of-vocabulary (OOV) words[1] ([4],[5],[6]). In this study, compound splitting is addressed once more in a speech recognition context. First of all, because applying compound splitting in a *Dutch* large vocabulary continuous speech recognition (LVCSR) framework has as yet not been thoroughly investigated. A second motivation was the small amount of evidence in the literature warranting a rejection or acceptance of the hypothesis that compound splitting indeed improves ASR *performance* in terms of word error rates (WER). That compound splitting substantially reduces the number of OOV words has been reported frequently, but exactly whether the application of the improved vocabularies in a real-world, large vocabulary, speech recognition task improves ASR performance, is often not directed. Therefore, this study compares Dutch LVCSR performance using language models trained on original text data on the one hand and decomposed text data on the other. Finally, as most of the splitting algorithms found in the literature rely on decomposition rules or lexicons with morpho-syntactic information that are costly to develop and usually have a limited coverage, a novel, data-driven compound splitting procedure is proposed that only needs a large development corpus, such as language model training data.

In section 2 the motivation for applying compound splitting in Dutch ASR is addressed shortly, followed by a brief discussion of possible disturbing factors of compound splitting applied in a (Dutch) ASR framework (section 3). In section 4, the compound splitting procedure will be described that was used to generate decomposed text versions for language model training. The language models based on the original text data and the decomposed text data were compared in an ASR evaluation, described in section 5. Some final conclusions and remarks can be found in section 6.

2. Reduction of lexical variability

The rationale underlying the decomposition of compound words in the context of speech recognition, is that OOV rates can drastically be reduced. The number of distinct words in compounding languages is relatively high compared to languages such as English or Italian. This is illustrated in Table 1 that was borrowed from [4], except for the statistics of Dutch. Given the high lexical variability of Dutch and German, a larger vocabulary is needed to achieve the same lexical coverage (or its inverse, the same OOV rate) as for Italian, English or French. As vocabulary size in LVCSR is typically limited to 65 K words —so practically invariable— vocabulary space for these languages may be regarded as particularly sparse. Word compounding is, besides stemming (and for German, case declension for articles, adjectives and nouns) the major cause of the rich lexical variability for German and Dutch. As compound words may as well be represented by means of the compound *parts* or *constituents*, decomposing compounds could be a way to free some of the available vocabulary mass for other words to be contained in the vocabulary, so that eventually more words in the application domain can be covered.

	EN	IT	FR	DU	GE
#wrds (M words)	37.2	25.7	37.7	37	36
#dst. (K words)	165	200	280	462	650
5K-cov (%)	90.6	88.3	85.2	84.02	82.9
20K-cov (%)	97.5	96.3	94.7	92.64	90.0
65K-cov (%)	99.6	99.0	98.3	97.15	95.1
20K-oov (%)	2.5	3.7	5.3	7.36	10.0
65K-oov (%)	0.4	1.0	1.7	2.85	4.9

Table 1: Comparison of languages in terms of number of distinct words, lexical coverage and OOV rates for different vocabulary sizes (borrowed from [4] except for the statistics of Dutch).

3. Disturbing side-effects of compound splitting in ASR

However, although compound splitting may improve lexical coverage and reduce OOV words, it is uncertain whether it improves overall ASR *performance*. There are a number of side-effects of compound splitting that may undue a possible performance gain thanks to an improved OOV rate. Such disturbing side-effects can be distinguished according to the different

[1] Reducing lexical variability can also be advantageous for grapheme-to-phoneme (G2P) conversion, as described in [4]

stages in the recognition (development) process in which they occur:

- Acoustic modeling
  From an acoustic modeling point of view it is easier to recognize longer words than shorter words as longer words bear more acoustic information. Some evidence for the reduced speech recognition accuracy caused by the introduction of short compound constituents was found in [5].
- Dictionary generation
  The phonetic transcriptions of former compound parts may depart from the actual pronunciation within a compound when co-articulation effects occurred at constituent boundaries. Consequently, there will be a mismatch between the actual pronunciation of a compound and the phonetic representation in the phonetic dictionary of the recognizer.
- Language modeling
  It can not directly be foreseen what the effect of compound splitting is on n-gram estimation. The n-gram information is practically reduced to the $(n-1)$-gram information as the decomposed compound pushes one or more context words out of the n-gram.

One could anticipate on some of these side-effects by applying restrictions to the compound splitting procedure aiming at ASR performance optimization, for example, by restricting compound splitting to low frequent compounds. Highly frequent compounds normally have a high chance of being recognized correctly as the n-gram estimates may be expected to be reasonably well trained. It may therefore be preferable not to decompose such compounds, in spite of the fact that it would improve lexical coverage.

A complicating factor regarding compound splitting in a Dutch ASR context, is the binding morpheme "s". For Dutch, it is allowed to insert this binding morpheme between specific constituents[2], as in "*regering-s-leider* (*Eng.: leader of the government*)". There are three possible approaches for dealing with the binding morpheme in compound splitting: regard it as a single constituent (*regering s leider*), attach it to the preceding word (*regerings leider*) or delete it (*regering leider*). Each approach has its own drawbacks. Introducing the monophone word "s" is not desirable given acoustic modeling considerations, introducing new words by attaching the "s" will decrease the compound splitting effect on lexical coverage, and completely removing the "s" complicates the re-composition of compound parts in a post-recognition step. As the last approach seemed too far-fetched also from a linguistic point of view, in this study, only the first two approaches were implemented for evaluation.

4. Splitting method

A number of compound splitting methods have been discussed in the literature. In [7] a (empirically developed) set of decomposition rules is used for compound splitting. A drawback of using rules is that they are costly to develop, are language specific and may not guarantee a broad coverage. Moreover, evaluating the precision of the rules over a large set of words is difficult. In [2] and [3] morpho-syntactic information derived from a background lexicon (e.g., CELEX[3]) is used for compound detection and decomposition. A disadvantage of such methods is that they rely on both the availability and quality of the background lexicon. As for instance the Dutch CELEX lexicon was released in 1990, it does not contain words that were introduced recently, such as "*poederbrief* (*Eng.: a letter containing, possibly poisonous powder*)". Consequently, compounds that contain such words can not be detected. The coverage of lexicon based approaches may therefore be sub-optimal –especially for recency-sensitive domains as the news domain– although decomposition precision will generally be high. Data-driven approaches of compound splitting, such as applied by [6], may be expected to have a better coverage but as no linguistic information is used precision of such methods may be lower.

For achieving a high compound coverage, a data-driven method seemed to be the best alternative. A compound-search algorithm was developed that uses sorting, word length information and word frequency information to detect and split compounds.

4.1. Search algorithm

First, a compound was defined as a word that can be split into at least two separate words (α and β constituent) that both occur as single words with a minimum frequency of 10 in a large text collection. The minimum frequency was introduced to avoid that words that normally in Dutch do not occur as single items but by accident[4] appear in the text data, produce incorrect compounds. Furthermore, both constituents of a compound must have a minimum length of six characters, firstly to reduce possible disturbing effects on speech recognition performance of shorter words in advance, and secondly to eliminate false compound detections in cases such as "*per-vers* (*Eng.: perverse*)". Finally, a compound was allowed to have a binding morpheme "s" that at this stage was interpreted as a stand-alone constituent (e.g., *regering-s-leider*). The decision to attach the binding morpheme to the preceding word or delete it completely, as suggested earlier, was postponed to later processing stages.

To collect the largest possible amount of compounds, a word list of more then 1.5 M unique words collected from the text data, was alphabetically sorted. In this way, the first part of a compound, referred to as the α-constituent, always directly precedes the compound: "*voetbal* (*Eng.: football*)" for instance, precedes compounds such as "*voetbalschoen* (*Eng.: football shoe*)", "*voetbalstadion* (*Eng.: football stadium*)" and so on. By descending the word list and checking if the current word is used as an α-constituent in the next entries, compound words could be detected. Note that words with an initial uppercase were discarded to avoid false compound detections for named entities. This method was repeated using a list of words printed in reversed order so that words became search key for final constituents: the word "stadion (reverse: noidats)" could for instance be found as final constituent in "voetbalstadion (reverse: noidatslabteov)". A third compound-search detected words with constituents ending in the suffixes such as "*ing(s)*"," "*heid(s)*", and "*schap(s)*". The compound search algorithm found 323 K ($\backsim$ 20%) with at least two constituents in the first run. These compounds were put in a compound conversion table with the compound in one column and a compound splitting solution in the other.

[2] Whether the insertion of a binding morpheme is correct or incorrect is specified in "*Het Groene Boekje*" containing Dutch spelling rules

[3] CELEX is available from the Linguistic Data Consortium

[4] Due to normalization procedures or because they are foreign words

For some 2 % of the detected compounds the algorithm produced two or more possible decomposition alternatives. This happened for example when a compound could be split into three or more constituents, such as in "*wassen-beelden-gallerij* (*Eng.: waxwork gallery*)". This is usually not problematic, as the compound that is left after a first decomposition step, is decomposed in successive steps of the iterative compound splitting procedure. However, other examples, such as those in Table 2 also produce implausible compound splitting alternatives[5].

compound	plausible	implausible
reactiestappen	*reactie-stappen*	*reacties-tappen*
koningspaarden	*koning-s-paarden*	*koning-spaarden*
meubelsmokkel	*meubel-smokkel*	*meubels-mokkel*
zeeroverschatten	*zeerover-schatten*	*zeerovers-chatten*
politiekringen	*politie-kringen*	*politiek-ringen*

Table 2: Multiple splitting alternatives

Although a majority of the decomposition alternatives produced by the compound search algorithm are *possible*, not every alternative is *plausible*. A native speaker of Dutch will in general be able to pick the most plausible one out of the multiple compound splitting alternatives although in cases context information is required to make this decision. In order to select the most plausible compound splitting solution automatically, overall constituent frequencies (A) and within-compound constituent frequencies (B) were deployed to obtain a splitting plausibility measure $\mathcal{Q}_{split}$:

$$\mathcal{Q}_{split} = \overbrace{\left(\frac{f(\alpha)}{N_{const}} \cdot \frac{f(\beta)}{N_{const}}\right)}^{A:overall} \cdot \overbrace{\left(\frac{f_C(\alpha)}{N_{comp}} \cdot \frac{f_C(\beta)}{N_{comp}}\right)}^{B:within-compound} \qquad (1)$$

The overall frequencies $f(\alpha)$ and $f(\beta)$ were normalized over all words that are also compound constituents (N_{const}). The within-compound frequencies $f_C(\alpha)$ and $f_C(\beta)$ were normalized over all compounds (N_{comp}). As the space for this paper is limited we refer to [8] for a detailed explanation of this method. The performance of the decompound probability measure was evaluated using the "*Grote Van Dale*[6]" (*GVD*), a large background lexicon with 1.3 M words containing compound boundary information. Of all compounds with multiple decomposition alternatives, 1 2% existed in the *GVD*. According to the *GVD* lexicon, 86 % of the decompositions proposed by the selection algorithm (Equation 1) were correct.

The compound conversion table discussed so far only provided a pairwise compound splitting solution with only an α and a β constituent. To enable a further decomposition into three or more constituents, the compound search procedure was applied iteratively on the conversion table.

4.2. Evaluation of splitting accuracy

To evaluate the complete compound table, again the *GVD* lexicon was consulted. If a compound was found as an entry in the *GVD* lexicon, the compound splitting solutions produced by the splitting algorithm and those provided by the *GVD* lexicon were compared. If the *GVD* lexicon did not provide a compound splitting solution, it was assumed that the compound search algorithm had produced a false alarm: the compound apparently was not a compound. Furthermore, as the intention was to develop a compound splitting table with an optimal performance, the *GVD* corrections were directly used to improve the compound table. So, after the evaluation of the initial compound splitting iteration, detected incorrect splitting solutions were replaced by the correct solutions from the *GVD* lexicon. The corrected table was then used in the next iterations. The precision score after the initial compound splitting iteration is 96.69 %.

Given the high precision score after the first iteration, the final conversion table is assumed to have a high precision score as well. With this table, newspaper text data can appropriately be decomposed so that the effect on ASR performance can be investigated. A comparison of the original data set (well over 300 M words of Dutch newspaper data from 1999-2001) with the decomposed data set revealed that compound splitting had resulted in a 20% reduction in distinct words. The effect on the (self) lexical coverage of vocabularies is visually depicted in Figure 1 that shows the differences (deltas) in lexical coverage of lexicons based on decomposed data relative to the ones based on the original data. Only vocabulary sizes up to 100 K are shown. The figure shows that compound splitting has a negative effect on lexical coverage of very small vocabularies, a large positive effect on lexical coverage up to a vocabulary size up to some 20 K words and this positive effect slowly decreases when larger vocabularies are used. In general, this result can be interpreted as a confirmation of the hypothesis that lexical coverage of a vocabulary improves when compound words are decomposed, with an exception for very small vocabularies.

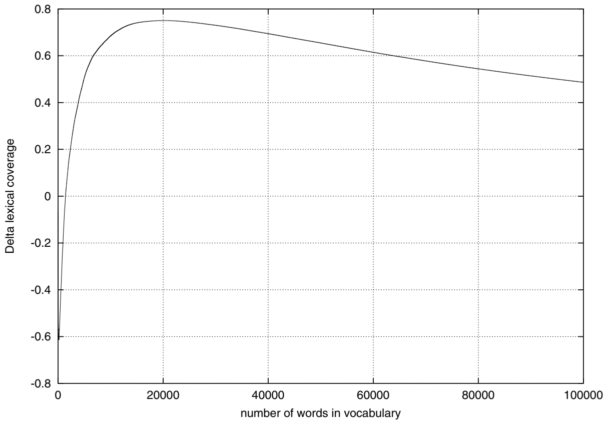

Figure 1: Differences in lexical coverage (y-axis) before and after compound splitting for vocabularies of different sizes (x-axis).

5. ASR experiment

5.1. Method

To investigate the effect of compound splitting on ASR performance, language models with a 65 K vocabulary (top 65 K most frequent words) were created based on different text versions of a Dutch newspaper collection of more then 300 M words. The original data set served as training data for the baseline language model, a number of differently decomposed text versions were used for the test language models. Compound splitting was done:

1. using an unrestricted compound splitting procedure:

[5] Translations of the compounds in Table 2 from above: *reaction steps, horses of the king, smuggling of furniture, pirate treasures and police circles*

[6] The *Grote Van Dale* lexicon was provided by the Dutch dictionary publisher Van Dale Lexicografie

   (a) treating the binding morpheme as a separate constituent.

   (b) attaching the binding morpheme to the preceding constituent (Glue-S).

2. Using restricted compound splitting procedures. Compounds were only decomposed if their frequency of occurrence was too small to be included in top N most frequent words in the original data were N was chosen to be 5 K, 20 K and 65 K.

The restricted procedure was created to investigate whether excluding frequent and probably well-modeled words could improve ASR performance over a unrestricted compound splitting procedure. Tri-gram back-off language models were created using a 65 K vocabulary of the most frequent words in the subsequent text data sets. Word pronunciations were obtained using a background pronunciation lexicon of 1.3 M words and a grapheme-to-phone conversion tool ([8]). As the grapheme-to-phoneme conversion may produce incorrect transcriptions, the pronunciation of words that were not included in all vocabularies (e.g., only occurred in one vocabulary), were manually checked to avoid that language model versions are put at a disadvantage as more words have to be produced by the grapheme-to-phoneme conversion tool, hence may have more incorrect word pronunciations. The *ABBOT* hybrid RNN/HMM speech recognition system ([9]) was used for the speech recognition evaluation. Acoustic models were trained forward and backward in time on broadcast news training data of 2000. The speech data consisted of a collection of 18 Dutch broadcast news shows (*NOS Acht uur journaal*) recorded January–March 2002. These were transcribed and segmented manually. Segments containing non-speech or speech of a foreign language were excluded from the test data, resulting in approximately 6.5 hours of Dutch speech (70 K words). For the scoring of the hypotheses based on decomposed data, the reference transcripts were decomposed accordingly.

5.2. Results

In Table 3 the OOV rates and WER's of all language models are listed. It shows that all language models created using decomposed text versions performed better then the baseline. The model based on the text version in which all compounds were decomposed except for those that were included in the 5 K most frequent words, achieved the best performance.

method	OOV	WER
baseline 65 K	2.59	39.8%
unrestricted 65 K	2.18	39.2%
unrestricted+glueS 65 K	2.22	39.2%
restricted 65 K	2.25	39.6%
restricted 20 K	2.19	39.1%
restricted 5 K	2.18	39.1%

Table 3: Comparison of speech recognition performances.

6. Discussion and conclusion

The results demonstrate that using decomposed text data for language model training improves the coverage of ASR vocabularies and as a result of that, ASR performance, regardless possible disturbing side-effects of compound splitting as mentioned in the introductory section. No effects could be observed caused by a different treatment of the binding morpheme "s". The hypothesis that one should not alter compounds that are highly frequent as they will probably have robust n-gram probability estimates, was confirmed by the experiment. The results suggest that omitting compounds in the 0-20 K word frequency range is sufficient for optimal performance. Noteworthy is the fact that the WER of the 20 K restricted model equals the one of the 5 K restricted model, although its OOV rate is slightly worse. This may indicate that the negative effect of having more OOV words is neutralized by more robust n-gram models. We intend therefore to address further research to compound splitting on the n-gram level. It may be worthwhile to include the opposite of compound splitting in this investigation: the combination of frequent orthographic word tuples, referred to as multi-words, into single items in the recognition lexicon (see e.g., [10]) .

7. References

[1] T. G. Vosse, "The Word Connection," Ph.D. dissertation, University of Leiden, The Netherlands. Neslia Paniculata Uitgeverij, Enschede, 1994.

[2] C. Monz and M. de Rijke, "Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian," in *Proceedings CLEF 2001*. Springer, 2002.

[3] R. Pohlmann and W. Kraaij, "Improving the precision of a text retrieval system with compound analysis," in *Proceedings of the 7th Computational Linguistics in the Netherlands (CLIN 1996)*, J. Landsbergen, J. Odijk, K. van Deemter, and G. V. van Zanten, Eds., 1996, pp. 115–129.

[4] M. Adda-Decker and L. Lamel, "The Use of Lexica in Automatic Speech Recognition," in *Lexicon Development for Speech and Language Processing*, F. v. Eynde and D. Gibbon, Eds. Kluwer Academic, 2000.

[5] A. Berton, P. Fetter, and P. Regel-Brietzmann, "Compound words in large-vocabulary german speech recognition systems," in *Proc. ICSLP '96*, vol. 2, Philadelphia, PA, 1996, pp. 1165–1168.

[6] M. Larson, D. Willett, J. Köhler, and G. Rigoll, "Compound splitting and lexical unit recombination for improved performance of a speech recognition system for German parlianmentary speeches," in *6th Int. Conference on Spoken Language Processing (ICSLP)*, Beijing, China, 2000.

[7] M. Adda-Decker and G. Adda, "Morphological decomposition for ASR in German," in *Workshop on Phonetics and Phonology in ASR*, Saarbrücken, Germany, March 1-3 2000.

[8] R. Ordelman, "Using dutch speech recognition for the creation of document representations in multimedia information retrieval (provisional title)," Ph.D. dissertation, University of Twente, The Netherlands, October 2003, to appear.

[9] A. Robinson, G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and D. Williams, "Connectionist Speech Recognition of Broadcast News," *Speech Communication*, vol. 37, pp. 27–45, 2002.

[10] J. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," in *Proc. ARPA Speech Recognition Workshop, Chantilly, Virginia*, Februari 1997, pp. 56–63.