A HYBRID APPROACH TO COMPOUNDS IN LVCSR

Tom Laureys¹, Vincent Vandeghinste² and Jacques Duchateau¹

¹K.U.Leuven ESAT-PSI, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
²K.U.Leuven CCL, Maria-Theresiastraat 21, 3000 Leuven, Belgium
e-mail: {tom.laureys,jacques.duchateau}@esat.kuleuven.ac.be vincent@ccl.kuleuven.ac.be

ABSTRACT

In several languages compound words form orthographic units, which complicates the task of ensuring good lexical coverage for large vocabulary continuous speech recognition (LVCSR). A common approach to the problem consists of first recognizing the compound constituents, followed by an automatic recompounding process. We describe an accurate compound module, which combines a rule-based approach with statistical pruning. The module is incorporated in a broadcast news recognition task for Dutch and yields an 11% relative decrease in word error rate (WER).

1. INTRODUCTION

Compounding is an ubiquitous morphological process in most languages. Due to spelling conventions, compounds are orthographic units in some languages (e.g. Dutch, German, Swedish) while they are not in other languages (e.g. English). The orthographic representation of compounds as one word poses severe problems for automatic speech recognition systems, and more specifically for the design of the lexicon. One important task of the lexicon is specifying which words are known by the recognition system. This boils down to a selection of words so as to guarantee maximal lexical coverage for a given lexicon size (typically 40k to 65k nowadays). Obviously, the presence of orthographic compounds increases the number of distinct word tokens, leading to substantially higher Out-Of-Vocabulary (OOV) rates. In turn, high OOV rates have a severe impact on recognition performance as estimates of the number of recognition errors due to each OOV-word range from 1.6 to 2.2 [1].

Apart from compounding, inflection and derivation are two other mechanisms for creating new words on the basis of existing ones. Others have incorporated both morphological mechanisms in their approach to lexicon optimization, mostly with little success [2, 3]. We decided not to tackle inflection and derivation for two reasons. First, inflection and derivation only create a limited number of derived word forms, since affixes constitute a relatively small, closed set with a finite number of combination rules. Compounding, on the other hand, is an exponential mechanism as it allows for the combination of two or more constituents from a nearly infinite set. Second, most affixes form small and mostly unstressed phonemic units, which results in a worse recognition performance due to a higher degree of acoustic confusability [3].

The compound module described in this paper serves a double purpose as it is used for reducing the lexicon (by decompounding) as well as for recompounding the recognition output which consists of non-compounds only. The module takes a hybrid approach, meaning that it combines a rule-based approach with statistical pruning measures. We evaluated the system on Dutch LVCSR and report a 30% reduction in lexicon size and, most importantly, an 11% relative decrease in WER on a broadcast news task.

The paper is organized as follows. In section 2 we compare the lexical variety of Dutch to that of other languages and investigate how large a decrease in OOV rate can be expected from decompounding in Dutch. Section 3 describes the compound module and its integration in the LVCSR system. The speech recognition experiments and their results are reported on in section 4. We end with conclusions and suggestions for future research.

2. LEXICAL VARIETY IN DUTCH

As explained above, morphological processes and spelling conventions determine the amount of lexical variety in a language. Dutch is characterized by processes of inflection, derivation and mainly compounding. Moreover, compounds form orthographic units like in German. The influence of lexical variety on lexical coverage for different languages is summarized in table 1, adapted from [4]. Lexical coverage was measured on newspaper text, respectively on the Wall Street Journal (37.2M words) for English, Le Monde (37.7M

This publication was supported by the project AtraNoS (Automatic Transcription and Normalisation of Speech), which is funded by the Flemish government (IWT). More information about the project can be found on the following web site: http://atranos.esat.kuleuven.ac.be. We thank Kris Demuynck for fruitful discussions and suggestions.
Table 1. Lexical coverage as a function of vocabulary size.

<table>
<thead>
<tr>
<th>No. words</th>
<th>English</th>
<th>French</th>
<th>German</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>5k</td>
<td>90.6%</td>
<td>85.2%</td>
<td>82.9%</td>
<td>82.9%</td>
</tr>
<tr>
<td>10k</td>
<td>94.9%</td>
<td>90.6%</td>
<td>86.1%</td>
<td>87.9%</td>
</tr>
<tr>
<td>20k</td>
<td>97.5%</td>
<td>94.6%</td>
<td>90.0%</td>
<td>92.0%</td>
</tr>
<tr>
<td>40k</td>
<td>99.0%</td>
<td>97.3%</td>
<td>93.9%</td>
<td>95.2%</td>
</tr>
<tr>
<td>60k</td>
<td>99.6%</td>
<td>98.3%</td>
<td>95.1%</td>
<td>96.6%</td>
</tr>
<tr>
<td>80k</td>
<td>99.7%</td>
<td>98.9%</td>
<td>95.7%</td>
<td>97.3%</td>
</tr>
</tbody>
</table>

The table shows that Dutch lies in between French and German with respect to lexical coverage. The difference with German is partly due to the fact that, in contrast with the Dutch results, upper and lower case were distinguished for German. Yet, a more important factor is that German still has case declension for articles, adjectives and nouns. This phenomenon is no longer present in Dutch. As to compounding, Dutch and German are very similar in that many combinations are allowed.

In order to assess the degree of compounding in Dutch, we compared the OOV rate (for the given lexicon sizes) of the original 34.7M words newspaper text to a decompounded version of the same text. Decompounding was performed by the module described in section 3. The results are shown in figure 1. In contrast with OOV rates for decompounding in Dutch reported in [5], we do get a considerable and consistent improvement when applying decompounding. As this improvement is similar to results for German reported in [6], we suppose that the decompounding algorithm used in [5] was not optimal.

3. A HYBRID APPROACH TO COMPOUNDING

3.1. Compounds in LVCSR

Section 2 shows that coping with (orthographic) compounds is indispensable to ensure good lexical coverage in Dutch LVCSR. We propose the following method. First, all lexicon words are decompounded, resulting in a reduced lexicon which consists of non-compounds only. Then, automatic speech recognition is performed on the basis of this non-compound lexicon. Finally, automatic recomposing is applied to the recognition output by means of a compound module.

The compound module proposed in this paper first uses rules to generate possible compound candidates. The rules cover all possible combinations to produce compounds and thus overgenerate. So the set of candidates is subsequently pruned on the basis of statistics.

3.2. Rule-Based Compounding

The Dutch word formation rules applied by the compound module were derived from [7] and [8]. These rules concern two input units (a modifier and a head) and condition on the morpho-syntactic properties (POS tags, gender, flexion, etc.) of the units. The following is an example rule, stating that in Dutch a compound noun (e.g. *wandelstok*—E: *walking stick*) can be formed on the basis of a verb modifier in its base form (*wandelen*) and a noun head (*stok*)

\[ \text{Noun}_{\text{comp}} \rightarrow \text{Verb}[\text{VFORM} = \text{base}]_{\text{mod}} \text{Noun}_{\text{head}} \]

A detailed description of the 15 rules employed can be found in [9]. If the compound consists of more than 2 word parts, the rules are applied recursively.

3.3. Statistical Pruning

Since the rules mentioned in section 3.2 tend to overgenerate, the set of candidate compounds is pruned by means of statistical information. Two statistical measures were taken into account: the Relative Frequency Threshold and the Compound Probability.

The Relative Frequency Threshold (RFT) copes with homonyms having different morpho-syntactic characteristics. For example, the Dutch word *bij* can be both a noun (E: *bee*) and a preposition (E: *by, at*). [9] spells out in detail how different word frequency lists for Dutch—based on 83.5 M words in total—were combined into a frequency list with 385K entries. From this list we derived the frequency for each pair of a word and its corresponding morpho-syntactic information relative to the frequency of its homonyms. If the relative frequency of a word–syntactic information pair
does not exceed the RFT, the pair is not considered for compounding. The RFT is applied to the modifier as well as the head. Its optimal value for the experiment set forth in section 4 was empirically determined as 0.05.

In order to reduce the set of compound candidates further, a Compound Probability (CP) is proposed. The CP for a compound with modifier \( i \) and head \( j \) is based on the following ratio:

\[
\frac{P(w_1 = \text{mod}_i, w_2 = \text{head}_j)}{P(w_1 = \ast, w_2 = \text{head}_j)}
\]

with \( P(w_1 = \text{mod}_i, w_2 = \text{head}_j) \) the unigram probability that the concatenation of words \( w_1 \) and \( w_2 \) occurs in a text, and \( P(w_1 = \ast, w_2 = \text{head}_j) \) the probability that any compound with head \( w_2 \) occurs (i.e. combined with any modifier).

In order to cope with unseen events we applied linear interpolation for smoothing. The CP can then be estimated as follows:

\[
\frac{\text{Count}(w_1 = \text{mod}_i, w_2 = \text{head}_j)}{\text{Count}(w_1 = \ast, w_2 = \text{head}_j)} \times (1 - D_{\text{head}_j}) + \frac{\text{Count}(w_1 = \text{mod}_i, w_2 = \ast)}{\text{Count}(w_1 = \ast, w_2 = \ast)} \times D_{\text{head}_j},
\]

where \( \text{Count}(w_1 = \text{mod}_i, w_2 = \text{head}_j) \) represents the compound’s frequency as provided by the frequency list, \( \text{Count}(w_1 = \ast, w_2 = \text{head}_j) \) the sum of the frequencies of all compounds in the list with \( w_2 \) as head, \( \text{Count}(w_1 = \text{mod}_i, w_2 = \ast) \) the sum of the frequencies of all compounds in the list with \( w_1 \) as modifier, and \( \text{Count}(w_1 = \ast, w_2 = \ast) \) the sum of all frequencies of compounds in the list. \( D_{\text{head}_j} \) is the discount parameter, representing the probability mass for words that are not in the frequency list:

\[
D_{\text{head}_j} = \frac{N_{\text{mod}_i}}{\text{Count}(w_1 = \ast, w_2 = \text{head}_j) + 1}.
\]

\( N_{\text{mod}_i} \) being the number of different modifiers occurring with the given head \( w_2 \).

Fixing the CP threshold value at 0.005 yielded optimal results. If the CP for a hypothesized compound exceeds this threshold, a compound is returned. Otherwise, the words are retained as separate orthographic units.

### 3.4. Decompounding and Recompounding

Both processes of decompounding (lexicon reduction) and recompounding (of the recognition output) are handled by the compound module. As explained above, the module deals with recompounding directly. For decompounding, the current word is first looked up in the recognition lexicon, which consists of non-compounds only. If no match occurs, all possible split-ups of the word in two, three or four parts are hypothesized. Each time the resulting word parts are fed into the module and recompounding is tried. If no recompounding is possible, the original word is retained. Otherwise, the most probable candidate, i.e. the one with the highest CP, is chosen from among the set of successful recom poundings.

### 4. SPEECH RECOGNITION EXPERIMENTS

#### 4.1. Experimental Set-up

We evaluated the compound module in an automatic speech recognition experiment. The test set consisted of 21 speakers, uttering 4797 words in 420 sentences of radio broadcast news taken from the Spoken Dutch Corpus [10]. We used the large vocabulary continuous speech recognition system developed by the ESAT-PSI speech group at the K.U.Leuven. For all experiments context-dependent acoustic models were employed which were trained on 38 hours of read-aloud text from the Spoken Dutch Corpus. None of the speakers in the train set appeared in the test set.

First, a reference experiment was set up using a 40K lexicon. This lexicon was compiled by simply selecting the most frequent words from the normalized newspaper corpus De Standaard (34.7M words). A Good-Turing back-off trigram language model was trained on the same corpus.

Then, we integrated the compound module into the experiment. The 40K words of the reference lexicon were decompounded by the module, resulting in a 28K lexicon of non-compounds. So a 30% decrease in lexicon size was achieved. The De Standaard newspaper corpus was decompounded as well, adding 2.9M word tokens. On the decompounded corpus we trained a language model similar to the one in the reference experiment. The new language model and the 28K lexicon were used in automatic recognition. In a post-processing step the recognized output string was then fed into the compound module for fully automatic recompounding. In addition, we produced a manual recompounding of the recognized output string in order to determine the theoretically achievable improvement limit and to assess the accuracy of the automatic compound module (for recompounding).

<table>
<thead>
<tr>
<th>Experiment</th>
<th>WER</th>
<th>Rel.</th>
<th>OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>14.8%</td>
<td>—</td>
<td>3.3%</td>
</tr>
<tr>
<td>Fully Automatic</td>
<td>13.2%</td>
<td>11%</td>
<td>2.0%</td>
</tr>
<tr>
<td>Manual</td>
<td>11.4%</td>
<td>23%</td>
<td>2.0%</td>
</tr>
</tbody>
</table>

Table 2. A comparison of recognition performance.
4.2. Discussion of Results

The results of the experiments are summarized in table 2. The table shows that speech recognition in Dutch significantly benefits from an automatic approach to compounds. By automatically recomposing the recognition output produced on the basis of the 28K non-compound lexicon, we achieved an 11% relative improvement on the word level. Moreover, a manual recomposing of the same recognition result showed that in theory a 23% relative improvement is possible. The improvement is mainly due to the reduction in OOV rate. Note that a 1.3% decrease in OOV rate can lead to a 3.4% decrease in WER. This high factor (2.6) can be explained by the fact that on top of the typical number of errors per OOV word due to the detrimental effect of an OOV on the recognition of neighbouring words (see section 1), a compound OOV is typically recognized as a sequence of its constituents, automatically introducing 2 or more errors in the WER metric. Moreover, the use of the non-compound lexicon does not introduce many new errors: the additional acoustic confusability is minimal as recognized compound constituents, unlike morphemes for flection and derivation, are not typically small phonetic strings. We also checked the 2.0% remaining OOV words and found out that 70% of them are proper nouns. Leaving proper nouns aside would result in a 0.6% OOV rate.

A total of 315 compounds was constructed by manual recomposing. The compound module processed 263 of them (83.5%) correctly. The main type of errors made by the module consist of the overgeneration of compounds with a preposition as modifier and a noun as head, which can be explained by the fact that this type of compounds is very frequent while the bigram frequency of preposition followed by noun is high as well. An example is the word opkomst (E: attendance) and the phrase op komst (E: on its way), both occurring often in running text.

5. CONCLUSIONS AND FUTURE RESEARCH

We conclude that an accurate approach to compounds leads to a substantial improvement in word error rate for LVCSR in Dutch. We assume similar results are possible for other morphologically rich languages such as German, Swedish or Norwegian. The achieved reduction in OOV rate results in a larger WER improvement due to the bad contextual influence of misrecognized OOV words and the typical misrecognition of OOV compounds as a sequence of 2 or more words.

Future research will focus on better statistical measures for the compound module. Section 4 showed that the current system yields an 11% relative improvement whereas 23% is theoretically possible. We assume further research on the automatic compound module will bring the WER closer to this theoretical limit. In addition, we aim at a better integration of the compound module in the recognition process. In the current implementation recompounding occurs in a post-processing step. A more intelligent combination of compound module and language model would result in a one-step recognition of compounds, doing away with post-processing altogether.

6. REFERENCES


