A mixed word / morphological approach for extending CELEX for high coverage on contemporary large corpora

Joris Vaneyghen[*], **Guy De Pauw**[†], **Dirk Van Compernolle**[*], **Walter Daelemans**[†]

[*]K.U.Leuven/ESAT/PSI
Kasteelpark Arenberg 10 , 30001 Leuven, Belgium
{joris.vaneyghen, dirk.vancompernolle}@esat.kuleuven.be

[†]University of Antwerp/CNTS
Universiteitsplein 1, 2610 Antwerpen, Belgium
{guy.depauw, walter.daelemans}@ua.ac.be

Abstract

This paper describes an alternative approach to morphological language modeling, which incorporates constraints on the morphological production of new words. This is done by applying the constraints as a preprocessing step in which only one morphological production rule can be applied to an extended lexicon of known morphemes, lemmas and word forms. This approach is used to extend the CELEX Dutch morphological database, so that a higher coverage can be reached on a large corpus of Dutch newspaper articles. We present experimental results on the coverage of this extended database and use the extension to further evaluate our morphological system, as well as the impact of the constraints on the coverage of out-of-vocabulary words.

1. Introduction

Most applications involving language modeling, such as speech recognition, are very vulnerable to the problem of out-of-vocabulary (OOV) words, as these tend to trigger multiple errors in a row. This is especially true for languages such as Dutch, that have a rich morphology and are therefore difficult to cover with a static word lexicon. To tackle this issue, our speech recognition system for Dutch (FLaVoR) uses language models that operate on the level of the morpheme instead of on the word level, thereby providing a dynamic and productive lexicon that is able to deal with OOV words.

If we want to train accurate morphological language models, it is necessary to have reliable morphological analyses for the words in the training text corpus. Unfortunately, even state-of-the-art morphological systems tend to highly overgenerate on the word level. In the context of a language model (LM), this type of overgeneration may in fact have an equally detrimental impact as a large OOV-rate in a static lexicon. In the approach presented in this paper, we therefore choose not to recursively apply morphological productions rules to the smallest possible units (morphemes), but instead apply one single production rule to a large set of known bigger units, containing lemmas and word forms. Because it is not trivial to incorporate this constraint in a LM, we propose a more flexible approach in which the constraint is used as a preprocessing step. This paper introduces this alternative approach and describes experiments on extending the CELEX Dutch morphological database to obtain a higher coverage on contemporary large corpora.

The paper is organized as follows: first we describe the FLaVoR speech recognition architecture for which this approach has been developed. Next, we describe the CELEX Dutch morphological database, after which we detail the method proposed in this paper. After evaluating the CELEX database, we discuss our approach on the extension of CELEX. We conclude with suggestions for future work.

2. Speech Recognition and Morphology

2.1. Morphology in FLaVoR

Most current speech recognizers make use of the standard monolithic HHM-framework in which all knowledge sources (lexicon, acoustic model, language model) are combined in one big search. The main advantage of this approach is that including higher level information from lexicon and language model drastically reduces the search space and therefore indirectly helps in the disambiguation of phonemes. Yet, at the same time this architecture forces all knowledge sources to be extreme simple. As a consequence, there is little room for improving accuracy by using more complex linguistic knowledge sources.

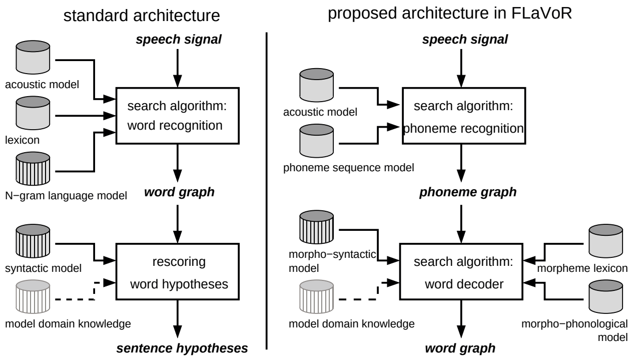

Figure 1: Standard vs. FLaVoR architecture

In the novel FLaVoR-architecture (Demuynck et al., 2003) the search is organized in two layers.

- The first layer is a pure acoustic-phonetic search. A phonetic network enriched with meta-data is generated as output.

- The second layer performs a search on the phonetic network using complex linguistic knowledge sources to decode words.

One of the knowledge sources used in the second layer of the FLaVoR-architecture is a morpho-syntactic model. Morphological processing has only recently been introduced in speech recognition for morphologically rich languages (Szarvas and Furui, 2003; Siivola et al., 2003). Incorporating morphology in language models has two advantages: first, the morphological analyses of words are a source of more general features, that can be exploited by language models. Second, it is possible to recognize new unobserved words, produced by morphological production rules, i.e. we can create a dynamic lexicon.

2.2. Dutch Morphology and the CELEX database

Dutch morphology is characterized by processes of inflection, derivation and mainly compounding. With a static lexicon of 60k words, we typically have to deal with an OOV rate in Dutch of 3.4% [1]. In terms of morphological productivity, this situates Dutch between French and German, that have an OOV rate of respectively 1.7% and 4.9%. For English the OOV rate is typically 0.4%[2].

The only extensive and publicly available morphological database for Dutch is CELEX (Baayen et al., 1995). It contains 381.292 word forms linked to 124.136 lemmas. The lemmas are represented as a hierarchical tree structure representing the internal morphological structure. Every word form is linked to its underlying lemma and is labeled with inflectional features. For example, ‘arbeidsfilosofieën’ (E: labor philosophies) is the plural of the lemma ‘arbeidsfilosofie’ and gets the label ‘m’ (plural). To get useful analyses for word forms, we further segmented these word forms and assigned the corresponding inflectional feature to the segments. For the example of ’arbeidsfilosofiën’ we get the segmentation: `arbeid+s+filosofie + en[m]`

3. Proposed Morphological System

Recent morphological work in the context of speech recognition was done for decomposing Dutch words (Vandeghinste, 2002), and automatically segmenting lemmas (De Pauw et al., 2004). The systems described in these papers provide morphological rules for decomposition and inflection that can be used as a postprocessing tool. We can however also take advantage of these morphological rules directly in the language model. But even though they achieve state-of-the-art accuracy as an analysis tool, these systems tend to be overgenerating in the context of language modeling, making it necessary to formulate some hard constraints on the production of new words.

It is however far from trivial to incorporate these constraints in the LM itself. We therefore propose a more flexible approach in which constraints are applied as a preprocessing step. This approach is independent of the chosen language model and the lexicon of the decoder. We will present experiments with structured language models (SLM) in this paper, as they are able to incorporate morphological production rules, but other LMs can be combined with our approach as well.

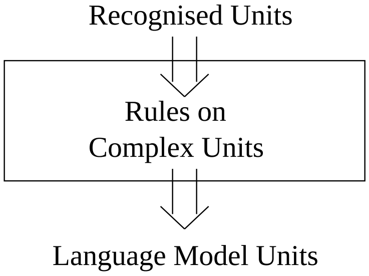

Figure 2: Illustration of the preprocessing on the recognized units. First, hard constraints are expressed by rules on CUs. Second, RUs are mapped on LMUs.

Figure 2 illustrates the preprocessing step between the word/morpheme decoder and the SLM. This preprocessing has two tasks. The first task is mapping the output sequence of the word/morpheme decoder, on a different sequence predicted by the SLM. We will call these sequences respectively the Recognized Units (RUs) and the Language Model Units (LMUs). In normal word based recognition systems there is no difference between RUs ans LMUs, but for our morphological approach we can take advantage of a mapping, especially when a SLM is used on the morpheme level. First of all the mapping between RUs and LMUs introduces some flexibility by making the decoder lexicon independent of the SLM lexicon. Thereby, the segmentation of a word does not have to be the same for the decoder as for the SLM. This last property is interesting for irregular word forms and some spelling/pronunciation difficulties, as there is no segmentation that fits the general structures used by the SLM. For example the irregular word form ‘kunsthistorici’ (E: art historians) will be recognized as the RUs ‘kunst’ and ‘historici’. By mapping the RUs to an alternative segmentation using a meta morpheme as the irregular inflection ending, we get the LMUs ‘kunst’, ‘historicus’ (E: historian) and ‘irr_plural’. In this way the SLM can first compound ‘kunst’ and ‘historicus’ and thereafter, inflect the result.

Another example is the verb ‘zet’ (E: set). To form the singular second and third person of a verb in the present, a ‘t’ is added after the lemma. But for lemmas already ending in a ‘t’, the additional ‘t’ is not spelled or pronounced. It would however still be better in a SLM to have the ‘t’ doubled. We can solve this by optionally mapping the RU ‘zet’ in the LMUs ‘zet’ and ‘t’.

The second task of the preprocessing is to put constraints on creating new words. The basic idea for the constraints is to allow only one production rule applied on known units. These units can be morphemes, lemmas and word forms and can exist of one or more RUs. We will call these units Complex Units (CUs). We allow the combination of a compound or derivation with multiple inflections in one production rule. This makes it easier to put constraints on inflect-

[1] We measured the OOV rate on *De Standaard* (37.4M words), a Flemish newspaper.

[2] Last three OOV rates are adapted from Young et al. (1997) and are also measured on newspaper text for a 60K lexicon.

ing compounds or derivations. For example, inflecting a compound can be seen as first inflecting the right compound before the actual compounding. In this way the unseen word ‘watersportactiviteiten’ (E: water sport activities) can be produced by the rule ‘NOUN + INFLECTED NOUN’ on the CUs ‘water+sport’ and ‘activiteit+en’.

The preprocessing step can be easily implemented in a Finite State Transducer (FST). For each class of CUs, a lexicon is built and represented by a sub-FST. Each path through the sub-FST represents a CU segmented in its RUs and LMUs. The arcs of a path are labeled with input symbols and output symbols, representing respectively the segmentation in RUs and the segmentation in LMUs. Each rule can be implemented by a concatenation of the corresponding sub-FSTs. Finally, putting all concatenations in parallel will form the full FST.

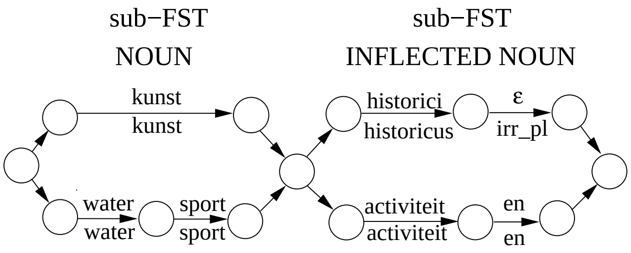

Figure 3: Illustration of the concatenation of sub-FSTs for the rule ‘NOUN + ‘INFLECTED NOUN’. Input labels (RUs) and output labels (LMUs) are respectively situated above and below the arcs.

Either the FST will output one or more sequences of LMUs for some given input sequence of RUs, or it will output nothing if no rule could be applied to some of the corresponding CUs. The probability of a sequence of RUs can be calculated by a weighted sum of probabilities over the LMUs.

$$P(RUs) = \sum_{LMUs} P_{LM}(LMUs)P(RUs|LMUs) \quad (1)$$

If the mapping between RUs and LMUs is carefully chosen so that each segmentation of LMUs corresponds to a unique segmentation of RUs, then the term $P(RUs|LMUs)$ is always 1 and can be omitted in the calculations.

4. Evaluation of CELEX

We used the Mediargus corpus (270M words) containing Dutch newspaper and magazine material. This corpus contains 1.8M unique words. The true lexicon size is probably significantly smaller, as this number contains a large number of spelling errors and duplicate entries due to insufficient normalization.

Table 1 gives an overview of the number of analyzed words and the coverage of CELEX and the extension of CELEX. From the 381K analyses, we can build a lexicon of 302K single word tokens. Although a 60K word lexicon is typically sufficient to have a coverage of 96% for newspapers, the 302K CELEX lexicon only has a coverage of 90.7% on the Mediargus corpus. Furthermore, only 54% of the words in the lexicon occur at least once, indicating a significant discrepancy between Mediargus and the corpora used to construct the CELEX database. CELEX was developed mainly from Dutch sources while Mediargus only contains material from Belgium. Also, a lot of typical newspaper words are not in the CELEX database, for example the word ‘gemeenteraadsverkiezing’ (E: community council elections). Moreover, CELEX includes very few proper names (4.7K) while newspaper material typically contains a considerable number of proper names.

Extending the CELEX database is interesting for two reasons. First, we need reliable segmentations for all of the CUs we want to consider for the preprocessing stage of our morphological system. Second, to have useful training data for a morphological language model, we need reliable analyses for the missing 9.3% of the corpus. Fully automatic analyses are however not accurate enough, so we made extra manually verified analyses of the missing 55K words in a 113K word target lexicon (all words that occur at least 50 times in Mediargus). With the 113K word lexicon we obtain a coverage of 97.3% of the training corpus and a coverage of 97.8% with all CELEX words plus the extra 55K words.

First, the 55K extra words were analyzed automatically. For this purpose, we used an earlier and simpler version of our mixed word/morphological approach. To analyze the words, we first need to segment them, i.e. place morpheme boundaries. For word segmentation we used an adjusted version of the FST described in De Pauw et al. (2004). Because the FST for word segmentation already has the possibility to produce meta morphemes, there was no need to map the RUs on LMUs. Furthermore, the one rule constraint in the preprocessing step is only applied to the creation of new lemmas and not to the inflection of lemmas. The lexicon of LMUs contains the 28.6K morphemes used in CELEX and an additional set of 6.2K frequent multi-morpheme lemmas. Adding this last set to the lexicon substantially reduces the depth of the structures used by the SLM. The SLM we used is a Left Corner Parser (LCP) (Van Uytsel and Van Compernolle, 2005). This SLM considers a word segmentation in LMUs as a miniature sentence and builds a tree structure over these segments. We limit the lexicon of CUs to the set of LMUs. This limitation yields harder constraints during preprocessing, but will also reduce the coverage of new words.

Only the most probable segmentation/analyses for a word form are selected. Furthermore, the structures created by LCP are transformed to the annotation style of CELEX, i.e a structured analysis for the lemma and a further analysis of the inflection endings. Note that because of the one rule constraint, the structured analyses of lemmas have only one level. This considerably simplifies manual verification. After the manual verification, we can automatically replace the frequent multi-morpheme lemmas by their subsequent analyses, resulting in CELEX-like deep structures.

5. Discussion

On the on hand, putting harder constraints in the preprocessing stage will reduce the overgeneration of new words but on the other hand, it will also reduce the coverage of

	CELEX		CELEX
			+ 55K
#words	302K	+55K	357K
#proper nouns	4.7K	+36K	40K
#non proper nouns	297K	+19K	318K
#words ∈ 1.8M lexicon	164K	+55K	219K
#words ∈ 113K lexicon	58K	+55K	113K
coverage	90.7%	+7.1%	97.8%

Table 1: Overview of the number of analyzed words and their coverage of a text corpus.

new words. Constraints can be weakened by adding more production rules or by extending the CU lexicons. Constraints can be strengthened by disallowing less productive rules and by using more detailed production rules and word classes, or by disposing of less frequent CUs.

The most challenging task is to find a good balance between coverage on the one hand and overgeneration on the other. While it is easy to measure the coverage, overgeneration is more difficult to measure. It can however be measured indirectly for example by measuring the performance of the morphological system in disambiguating analyses (segmentations and structured analyses).

For the task of extending the CELEX database, we measured the coverage of new words and the precision and recall of the automatically analyzed words, ie. the ratio of the number of correct analyses to the total number of the proposed analyses for a word and the ratio of the number of correct analyses to the total number of relevant analyses for a word. After verification by hand, we observed that only 46% of the new words can be created by using morphological production rules. From these 46%, 23% could not be automatically analyzed due to the constraints in the preprocessing. Considering the very limited lexicons of CUs we still obtained a good coverage.

Looking at all automatically produced analyses we obtained a labeled precision of 79.2% and a recall of 83%. In previous experiments on segmentation only (De Pauw et al., 2004), we obtained an accuracy of 88.9% of correctly segmented words. Other experiments showed that starting from the correct segmentation, precision and recall of the analyses is around 89%. Considering the results of the previous experiments on the separate tasks, the precision and recall on the combined task is encouraging, especially if we have to deal with a lower coverage of new words.

6. Conclusions and Future Work

In this paper, we presented a new, flexible approach that allows us to incorporate hard constraints in morphological language modeling and applied it to the extension of the CELEX database. We performed a preliminary evaluation of the approach on the extension of CELEX. Even though the constraints had been defined too strictly in the experiments presented here, we were still able to obtain a reasonably good coverage on new words and observed promising results on the accuracy of morphological analyses. Development will continue as we work on a full version of our approach and perform more extensive evaluation. This also involves finding a good balance between coverage and overgeneration on new words.

Acknowledgments

The research described in this paper was funded by IWT in the GBOU programme, project FLaVoR: Flexible Large Vocabulary Recognition: Incorporating Linguistic Knowledge Sources Through a Modular Recognizer Architecture.(Project number 020192). http://www.esat.kuleuven.be/psi/spraak/projects/FLaVoR.

7. References

R.H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. *The Celex Lexical Database (Release2) [CD-ROM]*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, U.S.A.

G. De Pauw, T. Laureys, W. Daelemans, and H. Van hamme. 2004. A comparison of two different approaches to morphological analysis of dutch. In *Proceedings of the ACL 2004 Workshop on Current Themes in Computational Phonology and Morphology*, pages 62–69, Barcelona, Spain, July.

K. Demuynck, T. Laureys, D. Van Compernolle, and H. Van hamme. 2003. Flavor: a flexible architecture for LVCSR. In *Proc. European Conference on Speech Communication and Technology*, pages 1973–1976, Geneva, Switzerland, September.

V. Siivola, T. Hirismaki, M. Creutz, and M. Kurimo. 2003. Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner. In *Proc. European Conference on Speech Communication and Technology*, pages 2293–2296, Geneva, Switzerland, September.

M. Szarvas and S. Furui. 2003. Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 368–371, Hong Kong, China, May.

D. H. Van Uytsel and D. Van Compernolle. 2005. Language modeling with context-sensitive probabilistic left corner parsing. *Computer Speech and Language*, 19(2):171–204, April.

V. Vandeghinste. 2002. Lexicon optimization: Maximizing lexical coverage in speech recognition through automated compounding. In *Proc. 3rd International Conference on Language Resources and Evaluation*, volume IV, pages 1270–1276, Las Palmas, Canary Islands, May.

S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, and P.C. Woodland. 1997. Multilingual large vocabulary speech recognition: the European SQUALE project. *Computer Speech and Language*, 11(1):73–89.