MORPHOLOGICAL MODELING OF WORD CLASSES FOR LANGUAGE MODELS

Ulla Uebler and Heinrich Niemann

FORWISS - Bavarian Research Center for Knowledge Based Systems
D-91058 Erlangen, Germany
E-mail: {uebler,niemann}@forwiss.de

ABSTRACT

It is well known that good language models improve performance of speech recognition. One requirement for the estimation of language models is a sufficient amount of texts of the application domain. If not all words of the domain occur in the training texts for language models, a way must be found to model these words adequately. In this paper we report on a new approach of building word classes for language modeling in the bilingual (German, Italian) SPEEDATA project. The main idea is to classify words according to their morphological properties. Therefore we decompose words into their morphological units and put the words with the same prefix or suffix into the same class. Since morphological decomposition is error prone for unknown word stems, we also decomposed words by counting beginnings and endings of different length and used these subunits like prefixes and suffixes. The advantage of this approach is that it can be carried out automatically. We achieved a reduction in error rate from 9.83 % to 5.77 % for morphological decomposition and 5.99 % for automatical decomposition which can be performed without any morphological knowledge.

1. Introduction

Language models can be estimated well if the recognition vocabulary is represented in the training texts of language models. For the words that cannot be estimated appropriately by the available texts ways must be found to get a good estimation, also see [2].

For example, words can be put into word classes together with similar words that are well represented in the training texts. The problem of the SPEEDATA domain (German and Italian) is that only 4100 of 5900 words occur in the training texts for the language models. One possibility is to model words in word classes in order to obtain a good estimation for the words in the utterance.

In the following, we will describe our approach to language modeling in order to improve performance of the SPEEDATA system. Then, there will be a short introduction into the baseline system of the SPEEDATA project. In the next section, we will describe the SPEEDATA baseline system. Section 3 introduces our approach to building language models that model words which do not appear in the training texts. In section 4, we will present results we have obtained with different approaches of building language models, and we will finish with a conclusion of this work.

2. The baseline system

The project SPEEDATA was established in order to develop a system enabling land register offices of the bilingual region of South Tyrol [1] to enter data of the historic master book into electronic data bases. One characteristic of this task due to juridical terms is that data sometimes must be entered in both Italian and German. The region of South Tyrol is a bilingual region in Italy, where both languages are official languages. Speakers with German as mother tongue speak Italian with only little accent, whereas Italian native speakers often have only a rudimentary knowledge of German. For the German language, there is a big variety of dialects among the villages of the region as well as a high degree in variation among and within the speakers between standard German and the respective dialect.

For the data-entry task, data-entry forms have been developed together with the users of the system in order to enter data in the most efficient way. Starting with a general form, it can be switched to new forms with fields that have to be filled in. Depending on the data field, language models are either modeled by a word list, grammar or statistically trained with a training text. For example, family names are modeled as one of a list of names. For a date, we model a grammar consisting of numbers between 1 and 31, followed by a number between 1 and 12, or the name of the respective month, followed by the year. Descriptions of land and houses are entered as complete texts as well as rights of owners and relations among owners and neighbors. Different language models have been trained for each of those text types. This paper will deal with the part of the statistically trained language models.

The texts that are used for the estimation of the language models have been typed in with the computer interface SPEEDATA is based on. For each language, texts with 70000 words including 4100 different words were collected. The recognition vocabulary for statistically trained language models, however, consists of 5900 words, so nearly one third of the words cannot be modeled with these training texts, and word classes have to be introduced in order to model the non-present words.

The words that are not represented in the training texts of language models, are mostly proper names, i. e. names of owners of houses etc. These proper names can be obtained from lists of owners that have been entered to the system until now. Another type of words that are not modeled well enough in language

models are domain dependent words like parts of houses or rights concerning land.

3. Building word classes

One way of language modeling is the use of word classes for language models. In this approach, the transition propabilities of the language model are estimated among word classes instead of words. The probability of words of one word class are equally distributed in this approach.

In our experiments, we compare different approaches of modeling word classes within language models. There are mainly three different ways:

1. frequency of words,
2. semantic word classes,
3. morphological word classes.

In the frequency approach, each word that appears more often than n times is modeled as a word class on its own. This means that there are almost as many word classes as words. The words appearing less than n times (or not at all) in the training text are modeled altogether in one word class. This approach guarantees a precise modeling of words that are well represented in the training texts. On the other hand, for words that appear only rarely in the training text, the modeling may not be appropriate since words of all different kinds are put together in the same class.

A second way of modeling words into word classes is to design word classes depending on their semantic role in the context of data-entry. Since the missing words are mostly proper names which are available in terms of 46 lists of first names, last names etc., we put first words like proper names into one class with equal distribution and then count the rest of the words of the training texts. 2800 words can be found in word lists, 900 of them also appear in the training texts. With these last words we estimate the distribution of the other words of the word lists. The modeling of word classes like proper names can be realized easily this way.

Building word classes this way can be more difficult for words like rights and obligations. It is difficult to extract an appropriate list of words from texts since semantically similar words can appear in many contexts. Manually, one new word class was extracted from the training texts. This class with parts of houses like *shower, bathroom* consists of 350 words for German. However, it is impossible to select words that would appear in similar contexts without an enormous amount of time.

The third approach of modeling word classes is to model words according to morphological structure. For morphological decomposition we used MALAGA, developped at the department of Computational Linguistics at the University of Erlangen. For these experiments we decided to limit morphological decomposition to prefixes and suffixes for two reasons: firstly, the syntactical characteristics of a German word are described in the suffix, whereas German prefixes also describe words in a way e. g. *ge-* often indicates perfect tense. Secondly, it is almost impossible for morphological tools to decompose a word correctly if the word stem is not known. The length of prefixes and suffixes depends on the fineness of the morphological decomposition, thus in some experiments the endings *-e, -er, -en* are not considered as suffixes but *-ter* or even *-güter* which is already a word on its own (=*manor*) may be considered as suffix. Words with the suffix *-güter* all mean manors with the type of the manor further being specified by the first part of the word. On the other hand, words beginning with *Änderungs-* (=*change*) can be modeled in the same class, since they all describe nouns that have something to do with a change. Thus, for some words, morphological decomposition is in some way also a semantic classification.

When applying linguistic tools for decomposing words into their morphological units, a minimum length for a part of word is set. Furthermore, it must be said, that tools for morphological decomposition may produce a wrong decomposition if word kernels, suffixes or prefixes are not known or if the word is not of German origin. Thus, a decomposition of words with a beginning of A-, which may mean a lack of something for Latin derived words, is for most other words a decomposition error. Similarly, sorting words according to their ending of -en or -er may characterise verbs in their infinitive or, in the second case, nouns describing a male person *Besitzer*. This decomposition is also error-prone, since several words (e. g. *jung*) show an ending e. g. *-ung* without belonging to the respective class (nouns with feminine gender). Thus, in order to model words more appropriately, the context must be enlarged. On the other hand, the context may not be too large in order to avoid a too high specialisation of the words modeled into the same class like only specific types of contract (*-ungsvertrag*) instead of *-vertrag*. Also, some prefixes and suffixes could not be decomposed for some words, therefore a post-processing for a consistent decomposition must be carried out.

We also allowed prefixes and suffixes longer than one character. Figure 1 shows the distribution of prefixes and suffixes for both languages (5900 words) depending on the necessary minimal occurence of a prefix or suffix in order to become a word class.

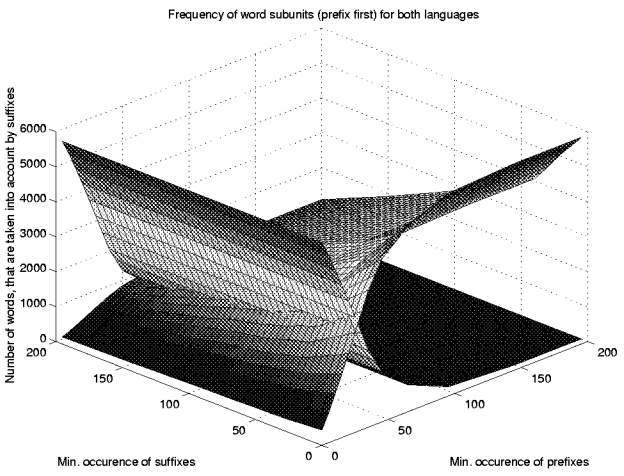

Figure 1: Occurence of prefixes and suffixes for Italian and German

In the figure there are two planes, for the number of prefixes and suffixes, respectively. For example, at $(x, y) = (5, 100)$

for *minocc(prefix), minocc(suffix))* there are 3359 words that are assigned to a word class according to a prefix and 104 words according to the suffix. In this figure, first the words are assigned to word classes according to their prefix, then the remaining words to their suffix, if there are enough occurences of the words for a classification.

Although the first choice is for prefixes, more words are assigned to word classes according to suffixes, thus suffixes seem more structured.

The number of assigned words approaches zero near a minimum occurence of 80 for prefixes. The decrease in occurence is much slower for suffixes, i. e. there are few suffixes that occur even 200 times in the lexicon. Since prefixes are chosen first in this example, their values are constant on the axis of suffixes, whereas the plane of suffixes depends on the number of words left by the choice of prefix.

The line where there is the same amount of words clustered according to the prefix as to the suffix, is roughly at a minimum occurence of 50 for prefixes and almost independent of the minimum occurence of suffixes, although, of course, the number of classified words decreases when a higher minimum occurence is required.

Looking at the languages separately, the rough structure is similar for Italian and German, although some differences can be found. For Italian, in the front left corner of the figure (prefix 0-50, suffix 0-150), there are more words classified to suffixes than for the German language. This means that for the remaining words after prefix classification, in German there are no suffixes left that would allow an occurence high enough to classify words to suffixes or that for the Italian language more words have been classified according to prefixes.

For German, in the opposite corner (prefix 100-200, suffix 100-200), there is still a high number of words that can be modeled, whereas for Italian the number of words classified according to the suffix reaches zero. One reason for that may be that the German language has more characteristic prefixes and suffixes for classification or that the used morphological tool is more sophisticated for German than for Italian since mainly it is used for the German language.

The sum of words modeled by either prefix or suffixes can be seen in Figure 2 (rotated by 180° with respect to Figure 1). Of course, most words can be classified with a low threshold for occurence of prefix and threshold. Furthermore, it can be seen that the number of words is not symmetric between the prefix and suffix axis. With an occurence of suffixes between 100-200 still around 300 words can be modeled, whereas for an occurence of prefixes between 100-200 almost the minimum is reached.

The remaining words are not assigned any class and will be further processed. Optionally, one new class for abbreviations can be created. Then, words can be put into classes according to the frequency approach or, alternatively, two classes will be established if the word begins with a capital letter (noun, proper name) or with a small character.

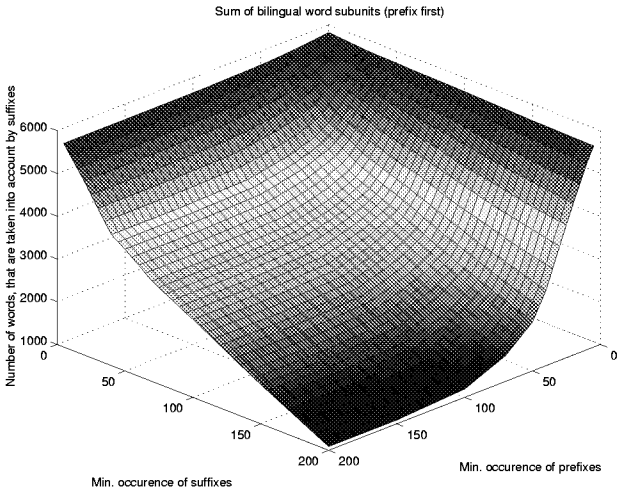

Figure 2: Occurence of prefixes and suffixes for Italian and German

By setting different thresholds for the minimum occurence of prefixes and suffixes and different procedures for the handling of the remaining words, we provide different word classes and therefore different language models for the recognition experiments.

Since morphological decomposition sometimes produces wrong decomposition and needs at least some manual corrections for word stems that are not in the lexicon, we have also automatically produced endings and beginnings by cutting off the first n characters and counting if they occur more often than m times in the lexicon. In many cases, there is an overlap with the manual categories, in some cases classes are built with a prefix plus a further character, if this combination appears more often than the threshold, e. g. some words of the German prefix *Ent-* is modeled in the class *Entw-*.

4. Results

For recognition we used the Isadora system [4], for language models we used bigrams for recognition and performed a polygram [3] verification for the best scoring sentences. We did experiments on several different sets of parameters, that is language model parameters themselves and for the morphology experiments with different thresholds. For the results, we will only give the best results of each category of experiment.

The first entry of Table 1 shows our standard way of training language models with the frequency approach. This approach leads to a word error rate of 9.83 %. Evaluated according to the languages, it can be seen that recognition is much better for Italian than for German, which is due to the dialects and accents of the speakers in German and to the fact that German seems to be easier to recognize. These results are only evaluated for statistically trained language models, and can therefore not be compared to results for the complete SPEEDATA evaluation which also contains word lists and finite state grammars.

Using word classes available from lists like proper names, recognition improves to 9.25 % word error rate. When using an ad-

Word error	Italian + German	Italian	German
frequency	9.83	7.82	12.40
semantics	9.25	8.05	11.88
suffix	6.60	4.04	9.88
prefix	5.82	4.04	8.09
prefix and suffix	8.47	3.97	14.24
prefix and suffix and abbrev.	5.77	3.74	8.38
prefix and suffix and abbrev. and case	11.42	5.82	18.60
automatical prefix and suffix	5.99	4.34	8.14

Table 1: Word error rate with different strategies for the design of word classes

ditional word class for the parts of houses, recognition decreases slightly, but still remains better than with the frequency approach.

The following five lines in Table 1 show the recognition results for the morphological approach. The first two lines refer to using either prefix or suffix for classification. Best results for prefixes were achieved with a threshold of 10 occurences for a prefix to become a word class, whereas for suffixes 2 occurences minimum yielded best results for the suffix only experiments.

Using both prefixes and suffixes, recognition decreases to 8.47 % word error, but is still better than the frequency approach (minimum occurence 10 each). Modeling abbreviations into one class, and setting the threshold to 2, best results are obtained with 5.77 % error rate. When modeling rough classes instead of counting the remaining words, the error rate increases to 11.42 % which is even more than the counting only approach.

The estimation of prefixes and suffixes leads to a word error rate of 5.99 %. This is slightly worse than the prefix or prefix+suffix+abbreviations approach, but better than all other approaches. The automatic approach cannot be best, since some morphological parts are not found. On the other hand, many errors that have been made by morphological decomposition due to unknown parts of the word do not occur, and therefore results are probably better than with most of the other morphological approaches.

Most of the words that were morphologically assigned to word classes are German words, and it is therefore very astonishing that the performance increased even more for Italian than for German. For Italian, the error rate is almost only half compared to our standard approach for most of the experiments. German, instead worsens in some cases, probably due to wrong classification of different words that were merged to the same class. For the best results on morphology, with a morphological tool and for automatic detection of prefixes and suffixes, the error decreases by 4 % absolute.

5. Conclusion

For this special task in SPEEDATA with only two thirds of the lexicon appearing in the texts used for estimating the language model we looked for an approach to model these words appropriately. Since some semantic word categories were already known, we first used them, and recognition improved. Using instead prefixes and suffixes for establishing word classes, the error rate was reduced significantly. Since morphological decomposition needs all word stems in the lexicon and still decomposes incorrectly, a semi-automatic correction became necessary for some words. In order to save correction time we estimated prefixes and suffixes automatically by counting the beginnings and endings of the words and we achieved almost the same performance. With some refinements of the automatic decomposition the result perhaps could be improved even more.

6. REFERENCES

1. U. Ackermann, B. Angelini, F. Brugnara, M. Federico, D. Giuliani, R. Gretter, and G. L. H. Niemann. SpeeData: Multilingual Spoken Data Entry . In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, 1996.
2. C. Crespo, D. Tapias, G. Escalada, and J. Alvarez. Language Model Adaptation For Conversational Speech Recognition Using Automatically Tagged Pseudo-Morphological Classes. In *Proc. ICASSP'97*, volume 2, page 823, Munich, Germany, April 1997.
3. T. Kuhn, H. Niemann, and E. Schukat-Talamazzini. Ergodic Hidden Markov Models and Polygrams for Language Modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 357–360, Adelaide, Australia, 1994.
4. E. Schukat-Talamazzini and H. Niemann. ISADORA — A Speech Modelling Network Based on Hidden Markov Models. *Computer Speech & Language*, 1993.