OPTIMIZED POS-BASED LANGUAGE MODELS FOR LARGE VOCABULARY SPEECH RECOGNITION

Petra Witschel

Siemens AG, Corporate Technology
Otto-Hahn-Ring 6, 81739 Munich, Germany
Petra.Witschel@mchp.siemens.de

ABSTRACT

Stochastic language models based on word n-grams require huge amount of training material and of storage especially for large vocabulary systems. Using n-grams based on classes much less training material is necessary and higher coverage can be achieved. Building classes on basis of linguistic characteristics (POS) has the advantage that new words can be assigned easily. Until now for POS-based language models class sets have usually been defined by linguistic experts. In this paper we present an approach where for a given number of classes a class set is generated automatically such that entropy of language model is minimized. We perform experiments on German medical reports of about 1.2 million words of text and 24 000 words of vocabulary. Using our approach we generate an exemplary class set of 196 optimized POS-classes. Comparing the optimized POS-based language model to the language model based on 196 normally defined classes we get an improvement up to 10% in test set perplexity.

1. INTRODUCTION

To improve recognition accuracy for large vocabulary speech recognition systems language models based on n-grams over classes are used. Using n-grams based on classes instead of words much less training material is necessary and higher coverage can be achieved.

Approaches of automatically generating word classes by clustering words in respecting a statistical criterion can be found e.g. in [1], [3], [4], [6], [8] and [9]. In [5] a clustering algorithm based on words is described for English which uses a small number of POS-based classes as additional initialization classes.

In our approach we derive classes from POS (see [10]) which are described via characteristics of words in terms of linguistic features and values. In contrary to automatic clustering techniques constructing POS-based language models requires lexically given, linguistic characteristics for each word. For this task we use a linguistic knowledge base of high coverage: a lexicon for German language [2]. Using this additional linguistic information inserting new words can be performed with more sophisticated adaptation methods. New words are put into relevant POS-classes according to their linguistic characteristics (see [11]).

Language models based on small number of POS show higher perplexity. Language models based on more detailed linguistic features result in a larger number of classes and perplexity gets lower. Using all lexically given features and values often results in a large number of classes. Respecting the available amount of text material and memory often it is necessary to reduce the number of classes.

Until now for POS-based language models class sets have usually been defined by linguistic experts according to linguistic aspects (see e.g. [8]). In this paper we present an approach where for a given number of classes a class set based on POS characteristics is generated automatically such that entropy of the resulting language model is minimized.

2. LANGUAGE MODEL

2.1. POS-based Language Models

The general task of a language model is to estimate for a given word chain $W = w_0 \ldots w_n$ the a priori probability P(W). In the case of bigram models P(W) is approximated as follows:

$$P\left(w_0 \ldots w_n\right) \approx \prod_{i=1}^{n} P\left(w_i | w_{i-1}\right) \tag{1}$$

POS-based stochastic language models (see [10]) assign words into classes according their linguistic characteristics. Therefore a word may belong to several classes. This results in the following approximation.

$$P(W) \approx \prod_{i=1}^{n} \sum_{C(w_i)} \sum_{C(w_{i-1})} P\left(w_i \mid C\left(w_i\right)\right) \cdot$$

$$P\left(C\left(w_i\right) \mid C\left(w_{i-1}\right)\right) \cdot P\left(C\left(w_{i-1}\right) \mid w_{i-1}\right) \tag{2}$$

The summation over the classes $C(w_i)$ and $C(w_{i-1})$ concerns all classes the word w_i or the word w_{i-1} belongs to. $P(w_i \mid C(w_i))$ and $P(C(w_{i-1}) \mid w_{i-1})$ are referred as "word probabilities" and the $P(C(w_i) \mid C(w_{i-1}))$ as "bigram probabilities" in this paper.

Quality of language models is measured via test set perplexity:

$$PP = 2^{H(LM)} \tag{3}$$

Language model entropy H(LM) is given as

$$H(LM) = -\left(\frac{1}{n}\right) \cdot \log \widehat{P}(W) \tag{4}$$

With $\widehat{P}(W)$ calculated via approximation (2) and n is the size of a sample text (e.g. the training set).

2.2. Constructing Class Sets According to Linguistic Knowledge

Linguistic features and values are used to define the characteristics of different POS-classes. Features and values are taken out of a large linguistic lexicon for German [2]. Some exemplary features and values are listed in table 1.

Feature	Values
main category	noun, verb, adjective, determiner,...
number	singular, plural
person	1th, 2nd, 3rd
gender	masculine, feminine, neuter
case	nominative, genitive, dative, accusative
inflection	weak, strong
degree	positive, comparative, superlative

Table 1: Exemplary Features and Values

Class sets defined by linguistic experts are typically constructed in such a way, that linguistically obvious context dependencies are modelled. E.g. in German in a noun phrase the congruence characteristics of the determiner (realized as the ending of a determiner or an adjective) are applied to the complement of the noun phrase (the noun). As an example class C_i may be defined using the features: “main category, gender, case, number” and the values: “noun, masculine, nominative, singular”. The class C_j may consist of all adjectives of masculine gender, case nominative, number singular and inflection strong. Under this conditions context dependencies are expected to be high. This is reflected in a relatively high bigram probability value $P(C_i \mid C_j)$.

3. OPIMIZATION APPROACH

In our approach we calculate POS-based language models which are based on optimized class sets. The optimized class set for a specific domain and training corpus is generated automatically such that entropy of the resulting language model is minimized. A domain specific training corpus and a linguistic lexicon are used as knowledge bases. To get the optimized class set we first are constructing the maximal class set (see section 3.1). The classes of the maximal class set are clustered (see section 3.3) such that optimization criterion (see section 3.2) is fulfilled. On basis of optimized class set the optimized language model is calculated. An overview of the different components of language model optimization is given in figure 1.

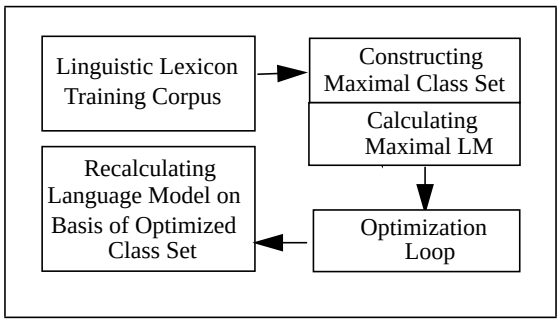

Figure 1: Components of Language Model Optimization

3.1. Constructing the Maximal Class Set

To find the maximal class set we determine all different feature and value combinations for the domain vocabulary contained in the linguistic lexicon. Each of such feature and value combinations is taken to perform the characteristics for a new class of the maximal class set. For domains of different medical fields maximal class sets range from 842 to 1023 numbers of classes. To give an example for German we consider the word “kleineres” (in English: “smaller”). Lexicon look up results in two combinations of linguistic features and values for adjectives of comparative degree which differ in case (nominative and accusative). The two combinations are represented as classes in maximal class set. They contain words like:

“größeres” , “kleineres”, “älteres”, “letzteres”, “stärkeres”,...
English: “bigger”, “smaller”, “older”, “latter”, “stronger”,...

Table 2: Exemplary Classes of Maximal Class Set

As second step a language model (maximal language model) is trained on the domain specific training corpus using the maximal class set.

3.2. Optimization Criterion: Minimization of Language Model Entropy

Calculating the optimized class set in optimization loop means that we are looking for a mapping $\underline{\mathrm{OPT}}_M$ such that language model entropy H(LM) (see formula (4)) is minimized:

$$\underline{OPT}_M = argmin_{OPT_M \in \Phi_M} H\left(LM\left(OPT_M\right)\right) \tag{5}$$

Φ_M is the set of possible mappings OPT_M which reduce a larger class set to a class set of M classes by clustering. Clustering two classes is done in constructing a resulting class as the union of the two original classes. This means the resulting class consists of all words of the two original classes. Language model LM is calculated on basis of the class set generated via OPT_M.

3.3. Optimization Strategy

To reduce number of classes of maximal class set classes are clustered. Respecting all possible clustering combinations is too expensive in practical use. Therefore in our approach the following suboptimal optimization strategy (see figure 2) is used. N is the number of maximal classes and M the target number of classes of optimized class set. Probability values of maximal language model are used. The most probable M classes of maximal class set are taken as basis classes to start optimization loop.

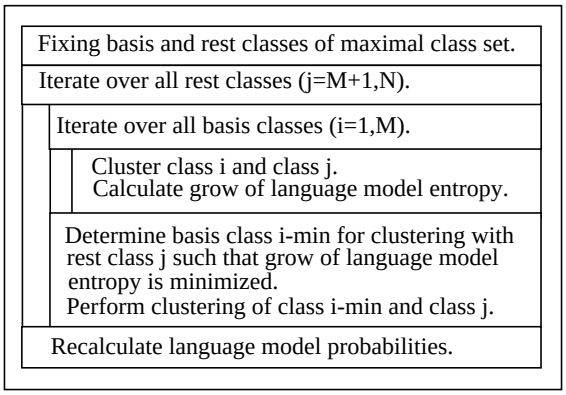

Figure 2: Optimization Strategy

In each iteration of the optimization loop one of the rest classes (N-M) is clustered to such one of the basis classes that the grow of entropy value (see (4)) of the resulting language model is minimized. Putting two classes together only those probabilities are recalculated, which are necessary to fix the grow in language model entropy. The “unknown class” and the class for “beginning of sentence” are not involved in the clustering process.

The resulting optimized classes are sometimes but not necessarily homogeneous according to a linguistic point of view. Words of our exemplary maximal classes (see table 2) are clustered to classes with corresponding linguistic characteristics but containing words of positive degree. Resulting optimized classes therefore contain words of positive and of comparative degree:

“großes”, größeres”, “kleines”, “kleineres”, “alt”, “älteres”,...
English: “big”,”bigger”,”small”,“smaller”,“old”, “older”,...

Table 3: Exemplary Optimized Class

4. CORPORA

Our experiments are performed on a medical domain. The training corpus contains reports of medical examinations about computer tomography (CT). Characteristics of the training corpus are listed in table 4. Before language model training the text corpus has to be normalized. This task is supported via our domain tool. Exemplary aspects are: finding limits of each sentence, normalizing orthography at the beginning of each sentence, representation of numbers according to spoken speech units, separating punctuation marks from words in context and representing ambiguous word units according to sentence context. Moreover to each word of the normalized text an unique class of the given class set has to be assigned. This tagging procedure is performed using our automatic tagging tool (see [10]).

For evaluation we use three test corpora. Test set 1 is our usual evaluation set for CT language models. Test set 2 contains all sentences spoken by speaker C. In test set 3 the utterances spoken by speaker D can be found. OOV rate of test sets is calculated on basis of a recognition vocabulary of 20 730 words.

CT Domain	Words of Text	Words of Vocabulary	OOV Rate
Training Corpus	about 1.2 million	23 938	
Test Set 1	2 414	598	1.2%
Test Set 2	1163	399	1.4%
Test Set 3	1059	336	1.9%

Table 4: Characteristics of Corpora

5. EXPERIMENTAL RESULTS

In our experiments we calculate different CT domain language models. The recognition vocabulary always consists of 20 730 words. The remaining vocabulary is taken to perform the training of the “unknown” class of the language model. First we determine a maximal class set of 1023 classes. On basis of this maximal class set the language model “LM-1023max” is calculated. As second step we reduce the maximal class set using our optimization approach. Different reduced class sets consist of 500, 250, 196, 100 numbers of classes. These class sets are used to calculate the language models “LM-500op”, “LM-250op, “LM-196opt”, “LM-100op”. To measure quality of optimization we calculate test set perplexity using the three test sets described in table 4. Comparing e.g. the optimized POS-based language model (“LM-196op”) to a language model with 196 classes defined by linguistic experts (“LM-196li”) we get an improvement of 10.3% in perplexity for test set 1. For “LM-500op” this results in an improvement of 12.5%. In table 5 test set perplexities are listed for the different language models and test sets.

LM	PP* Test set 1	PP* Test set 2 Speaker C	PP* Test set 3 Speaker D
LM-1023max	58.5	75.5	55.9
LM-196li	67.2	84.6	59.3
LM-500op	58.8	75.9	56.0
LM-250op	60.0	77.2	56.3
LM-196op	60.3	78.1	56.9
LM-100op	66.4	85.4	63.0

Table 5: Test Set Perplexity (PP*) of Language Models

Recognition tests have been performed using our large vocabulary speech recognizer in off-line evaluation mode. More detailed informations about the recognizer can be found in [7]. The test set comprises 2 speakers (speaker C, D). The speakers are doctors, who dictated how they are used to. This means they spoke like dictating for a human secretary. Therefore the recordings contain noise, hesitations, repetitions and restarts, especially for speaker D. Characteristics of the different test sets are listed in table 4.

For the recognition experiments unknown vocabulary is put into the "unknown" class of the language model and into the pronunciation lexicon. The bigram probabilities of "unknown" class have been calculated during off-line training of the language model. Results show no improvement for optimized language models. And especially optimized language models turn out to be a little bit more sensitive for spontaneous speech phenomena like hesitations. Results are listed in table 6.

LM	Speaker C WER*	Speaker D WER*
LM-196li	7.8	11.6
LM-250op	7.8	12.6
LM-196op	8.0	12.7
LM-100op	8.6	14.4

Table 6: Word Error Rate (WER*)

6. CONCLUSION AND FUTURE WORK

The experiments have shown that we have found a method to generate a class set automatically which is based on linguistic features and values and which moreover minimizes entropy of the language model. In our experiments language models based on optimized class sets turned out to reduce test set perplexity up to 10.3% compared to language models defined by linguistic experts according to linguistic aspects. Recognition experiments however show no improvement in recognition rate for optimized language models. And especially optimized language models turned out to be a little bit more sensitive for spontaneous speech phenomena. For this reason more detailed experiments will be performed e.g. in respecting in addition semantic features and values as linguistic characteristics and in fixing the best number of classes for optimized class set depending on the specific domain.

Nevertheless incorporating this method into our domain development process the process will become more automated. Especially we hope that for generating language models for "foreign" languages no linguistic expert will be necessary to fix the POS-based class set. Our next work will be to apply our adaptation techniques to optimized POS-based language models. So that optimized language models can profit of the advantage of putting new words into linguistically well determined POS-based classes.

7. REFERENCES

1. P. F. Brown, V. J. D. Pietra, P. V. deSouza, J. C. Lai, R. L. Mercer: „Class-Based n-gram Models of Natural Language“, Computational Linguistics, 1992, pp. 467-479.

2. F. Guenthner, P. Maier: "Das CISLEX-Wörterbuchsystem'', in "Lexikon und Text:", editors. H. Feldweg, E. W. Hinrichs, Max Niemeyer Verlag, Tübingen, 1996, pp. 69-82.

3. M. Jardino, G. Adda: "Automatic Word Classification Using Simulated Annealing", ICASSP, Minneapolis, 1993, V. II, pp. 41-44.

4. R. Kneser, H. Ney: "Improved Clustering Techniques for Class Based Statistical Language Modelling", 3rd Eurospeech, Berlin, 1993, pp. 973-976.

5. S. Martin, J. Liermann, H. Ney: "Algorithms for Bigram and Trigram Word Clustering", Speech Communication 24, 1998, pp. 19-37.

6. M. K. McCandless, J. R. Glass: "Empirical Acquisition of Language Models for Speech Recognition", ICSLP, Yokohama, 1994, pp. 835-838.

7. M. Niemöller, A. Hauenstein, E. Marschall, P. Witschel, U. Harke: "A PC-Based Real-Time Large Vocabulary Continuous Speech Recognizer for German", ICASSP, München, 1997, pp. 1807-1810.

8. T. R. Niesler, E. W. D. Whittaker, P. C. Woodland: "Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition", ICASSP, Seattle, 1998, pp. 177-180.

9. J. P. Ueberla: "More Efficient Clustering of N-Grams for Statistical Language Modelling", 4th Eurospeech, Madrid, 1995, pp. 1257-1260.

10. P. Witschel: "Constructing Linguistic Oriented Language Models for Large Vocabulary Speech Recognition", 3rd Eurospeech, Berlin, 1993, pp. 1199-1202.

11. P. Witschel, H. Höge: "Experiments in Adaptation of Language Models for Commercial Applications", 5th Eurospeech, Rodes, 1997, pp. 1967-1970.