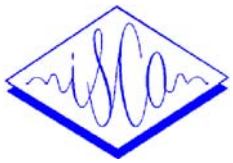

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

EXPERIMENTS IN ADAPTATION OF LANGUAGE MODELS FOR COMMERCIAL APPLICATIONS

Petra Witschel, Harald Höge

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

ABSTRACT

To improve recognition accuracy for large vocabulary speech recognition systems we use language models based on linguistic classes (extended POS). In this paper an adaptation technique is presented, which profits from linguistic knowledge about unknown words of new domain. Switching from basis domain to new domain we keep the bigram probabilities of linguistic classes fixed and adapt only monograms of word probabilities. In our experiments we use three different corpora: financial columns of a newspaper corpus and two medical corpora (computer tomography and magnetic resonance). Adapted language models show an improvement of test-set perplexity of 48% to 51% compared to the case of putting unknown words into the language model "unknown" class.

1 INTRODUCTION

Stochastic language models are used in large vocabulary speech recognition systems to improve recognition accuracy. For research purposes training of language models is performed on text databases of often more than several 10 millions of words, which mostly consist of newspaper material.

Current approaches in adaptation almost are based on n-gram language models, where n-grams consist of words (see e.g. [1], [2], [3], [4], [9], [12]). Those language models have high memory requirements and need large corpora for training. Adaptation means to reestimate word n-gram probabilities of language models trained on large, often domain independent corpora using statistical information of domain or user specific small text material. Language models based on n-grams of classes need much less training material. In [3] and [9] on small text material of new domain class probabilities are calculated to reestimate word n-grams of basis domain. Constructing classes with automatic clustering techniques (see e.g. [6], [8]) those language models usually have one single "unknown" class for putting unknown new domain words into. In [13] an automatic clustering approach for domain adaptation is presented using a modified optimization criterion.

Commercial applications, which fit to current speech technology, are restricted to more specific domains, like computer tomography diagnosis reports, for which only fewer text data (e.g. 1 million words) are available. For this purpose language models are required where low perplexity can be achieved with a small amount of training data. To accommodate language models for "similar" domains or even more specific tasks (e.g. reports of one specific hospital) adaptation is needed. In this situation often not more than 200 000 or less words of text are available

In our approach we use language models based on linguistic classes (extended POS, see [14]). Because we estimate probabilities for class n-grams we need less training material and memory. Moreover we profit of the fact that these class n-gram probabilities are word independent. A class is defined via given linguistic characteristics. Class based n-grams represent some kind of grammatical structure of the training corpus. For adapting domains with "similar" text structure we let the class n-gram probabilities unchanged and estimate only word monograms. We can easily assign new domain words to the relevant, linguistically defined classes. This is done according to linguistic characteristics of words, which we take out of a lexicon for German language [5].

2 LANGUAGE MODEL

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

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

Linguistic oriented, stochastic language models (see also

[14]) assign words into classes according their linguistic characteristics. So one word can belong to several classes. This leads to the following approximation.

$$P(W) \approx \prod_{i=1}^{n} \sum_{C_i} \sum_{C_{i-1}} P\left(w_i|C_i\right) \times P\left(C_i|C_{i-1}\right) \times P\left(C_{i-1}|w_{i-1}\right) \quad (2)$$

The summation over the classes C_i and C_{i-1} concerns all classes word w_i or word w_{i-1} belongs to. $P(C_{i-1}|\, w_{i-1})$, $P(w_i|\, C_i)$ are referred as "word probabilities" and $P(C_i|C_{i-1})$ as "bigram probabilities" in the following paper. The class bigram probabilities $P(C_i|C_{i-1})$ stand for some kind of statistically defined, linguistic grammar of the training corpus.

According to linguistic features: $(f_1,...,f_m)$ and values: $(v_{11},..,v_{1j})$ to $(v_{m1},..,v_{mj})$ to each word one or more classes are assigned. The feature and values we take out of a linguistic knowledge base of high coverage: a lexicon for German language [5]. The mapping F from linguistic features and values to classes is called classifier.

$$F\left(\left(f_1, v_{11} \cdots v_{1j}\right) \cdots \left(f_m, v_{m1}, \cdots v_{mj}\right)\right) = \left(C_1, \ldots, C_k\right) \quad (3)$$

In our experiments we use a classifier which generates up to k=195 classes. Exemplary linguistic classes are:
C: "noun, masculine, nominative, singular",
C: "adjective, masculine, nominative, singular, strong".

A special class is provided ("unknown" class, $C_{unknown}$), into which the user of a recognition system can insert new words. This is done without performing any additional training procedure. The class bigram probabilities $P(C_{unknown} \mid C_{i-1})$ and $P(C_i \mid C_{unknown})$ are estimated during the off-line training of the language model. To estimate these probabilities training material was prepared so that about 10% of the least frequent words of vocabulary are taken as "unknowns". The word probabilities $P(C_{unknown}|w_i)$ are fixed equal to 1.0 and the probabilities $P(w_i|\, C_{unknown})$ are taken equally distributed for unknown words w_i.

We start with a model trained on a large corpus (basis domain), which is "similar" to the new domain corpus. We call this the basis language model.

3 ADAPTATION APPROACH

For adaptation the class bigram probabilities $P(C_i|C_{i-1})$ (see formula (2)) are derived from basis language model. Only word probabilities for new domain words are to be calculated. New domain words are assigned to existing linguistic classes according to their linguistic characteristics. We apply mapping F (see formula (3)) to the features and values, which we get from our background linguistic lexicon for German [5]. Then word probabilities $P(w_i|C_i)$ and $P(C_{i-1}|w_{i-1})$ (see formula (2)) have to be estimated. For linguistically nonexistent class assignments the word probabilities are fixed to zero. Otherwise the $P(w_i|C_i)$ are estimated on basis of new domain corpus. For this in new domain text to each word an unique linguistic class (respecting the sentence context) has to be assigned. To achieve this without a rule-based parser we use our automatic, statistically based tagging tool [14]. Unseen events are estimated following the smoothing algorithm of [7]. The $P(C_{i-1}|\, w_{i-1})$ are recalculated according to formula (4).

$$P\left(C_i|\, w_i\right) = K \times P\left(w_i|C_i\right) \times P\left(C_i\right) \quad (4)$$

with normalizations factor

$$K = \frac{1}{\sum_{i=1}^{n} P\left(w_i|\, C_i\right) \times P\left(C_i\right)}$$

and $P(C_i)$ is taken form basis language model. The number of word probabilities to be estimated is proportional to the size of the vocabulary (which is in the range of 2 for German).

Word probabilities for "default words" are taken unchanged out of the basis language model. E.g:
numbers (e.g. "twenty_one"),
command words (e.g. "new_paragraph").

4 CORPORA AND BASIS LANGUAGE MODELS

Our experiments are performed on three different German text corpora. A newspaper corpus (SZ-domain: Süddeutsche Zeitung, financial columns) and two corpora about diagnosis of medical examinations (CT-domain: computer tomography, MR-domain: magnetic resonance). Attributes of the corpora are listed in table 1. Computer tomography corpus shows a more rigid grammar with short sentences. The mean length of a sentence range form 22 in SZ corpus to 8 in CT and 10 in MR.

	SZ	CT	MR
Words of text	1,79 Mio	1,2 Mio	961 896
Sentences	89 481	138 697	91 016
Vocabulary	102 675	24 122	26 577

Table 1: Text corpora

For each domain a language model is trained using the same 195 morpho-syntactic, semantic classes (see table 2). For our investigations we regard a 10 000 vocabulary recognizer task. These words were selected via frequency list of complete domain vocabulary (see also [11]). The probability values to estimate (see "Parameters" in table 2) consist of 38 925 bigram probabilities $P(C_i|C_{i-1})$ and e.g. for computer tomography of 23 754 values $P(w_i| C_i)$. The probabilities $P(C_{i-1}| w_{i-1})$ are recalculated from $P(w_i| C_i)$, see formula (4). This results in a number of 62 679 parameters. For a word based bigram language model of 10 000 words of vocabulary 10^8 bigram probabilities would have to be calculated.

Basis LM	SZ	CT	MR
Vocabulary	10 326	10 194	10 793
OOV	9.1%	1.6%	2.3%
Training: Words of Text	1 765 793	1 183 014	958 836
Classes	195	195	195
PP*	319	54	98
Parameters	68 904	62 679	63 641

Table 2: Basis language models
*Testset perplexity

5 EXPERIMENTAL RESULTS

In order to measure usefulness of our adaptation scheme we generate reference language models, which reflect real life situation. Assuming an user of a recognition system adding a hundred of new words to recognizer language model. Without using adaptation technique new words are inserted to the "unknown" class. Following section 2 (with $P(w_i| C_i)$ taken equally distributed) no additional probabilities have to be estimated. In our experiments we extract words of a fixed CT testset, which are unknown in relation to SZ basis language model or to MR basis language model. For these "unknowns" the numbers are given in table 3 (column 1 and 2, "Unknowns"). We insert them in the "unknown" class of the basis language model. For the resulting testset perplexities see table 3, "PP* Unk".

First adaptation is performed with basis language model trained on SZ corpus (see table 2, column 1). Computer tomography domain is used for adaptation (see table 3, column 1, "...Adapt."). For getting results of adaptation between different kinds of medical reports we generate a basis model on basis of MR corpus (see table 2, column 3). Adaptation is performed to CT domain (see table 3, column 2, "...Adapt."). In addition a CT basis language model is adapted to MR domain (see table 3, column 3).

In theses experiments MR and CT turned out to be of "similar" text structure. Adaptation shows an improvement of perplexity of 48% to 51% compared to the case of putting unknown words into "unknown" class (see table 3, columns 2 and 3: "PP Adapt." compared to "PP Unk."). Looking at trained language models (table 2) adaptation results in an increase of testset perplexity in the range of 21% to 25% in both cases. Adapting "dissimilar" domains (CT and SZ) testset perplexity increases 150% (table 3, column 1).

In case of adaptation much less probabilities are to be estimated. The reduction is in the range of 59% to 62% in comparison to the number of parameters for training without adaptation (table 2 and table 3: "Parameters" compared to "Param. Adapt.").

LM	CT	CT	MR
Basis LM	SZ LM	MR LM	CT LM
Unknowns	398	123	185
PP* Unk.	--	132	244
PP* Adapt.	135	68	119
Param. Adapt.	23 621	23 621	25 616

Table 3: Adaptation results
*Testset perplexity

However using small corpora has an influence on recognizer vocabulary. E.g. a computer tomography corpus of 188 500 words of text contains a vocabulary of 10 165 words including also words of frequency 1. Thus OOV rate is increased from 1.6% to 3.1%.

6 CONCLUSION AND FUTURE WORK

Our experiments have presented two domains of which text structures given via class bigram probabilities seem to be "similar" to each other (CT and MR domain). These domains were shown to be suitable for adaptation. The testset perplexity of the resulting adapted language model was calculated and it seems to be in a practicable range. On the other hand we have presented two domains which are obviously "dissimilar" to each other (SZ and CT domain). Adaptation of "dissimilar" domain language models were shown in the experiments to result in an unsatisfying testset perplexity.

For commercial applications language models of high coverage are necessary. With bootstrapping techniques the domain vocabulary can be gradually extended. For this purpose we are developing algorithms to carry over basis language model vocabulary to new domain language model classes. Sophisticated weights will be nec-

essary to fix the proportion of basis domain and new domain.

Our future intention is to find out, if we can result in a small number of basis language models, on which we can achieve good adaptations for many different domains. Some kind of perplexity measure based on class bigram probabilities of basis language model could be used to find for a new domain text automatically the optimal basis language model of high "similarity" in text structure.

REFERENCES

[1] S. Besling: "Confidence-Driven Estimator Perturbation: BMPC", Proc ICASSP 1997, Munich, pp. 803-806.

[2] P.R. Clarkson, A.J. Robinson: "Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache", Proc. ICASSP 1997, Munich, pp. 799-802.

[3] C. Crespo, D. Tapias, G. Escalada, J. Álvarez: "Language Model Adaptation for Conversational Speech Recognition Using Automatically Tagged Pseudo-Morphological Classes", Proc. ICASSP 1997, Munich, pp. 823-826.

[4] M. Federico: "Bayesian Estimation Methods for N-Gram Language Model Adaptation", Proc. ICSLP 1996, Philadelphia, pp. 240-243.

[5] F. Guenthner, P. Maier: "Das CISLEX-Wörterbuchsystem", CIS-Bericht 94-76-CIS, Universität München, 1994.

[6] M. Jardino, G. Adda: "Language Modelling for CSR of Large Corpus Using Automatic Classification of Words", Proc. 3rd EUROSPEECH 1993, Berlin, pp. 1191-1194.

[7] S.M. Katz: "Estimation of Probabilities from Sparse Data for the Language Model Component of Speech Recognizer", IEEE Trans. ASSP-35(3), pp. 400-401, March, 1987.

[8] R. Kneser, H. Ney: "Improved Clustering Techniques for Class-Based Statistical Language Modeling", Proc. Eurospeech 1993, Berlin, pp. 973-976.

[9] R. Kneser, J. Peters: "Semantic Clustering for Adaptive Language Modelling", Proc ICASSP 1997, Munich, pp. 779-782.

[10] H. Masataki, Y. Sagisaka, K. Hisake, T. Kawahara: "Task Adaptation Using MAP Estimation in N-Gram Language Modeling", Proc ICASSP 1997, Munich, pp. 783-786.

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

[12] P.S. Rao, M.D. Monkowski, S. Roukos: "Language Model Adaptation via Minimum Discrimination Information", Proc. ICASSP 1995, Detroit, pp. 161-164.

[13] J.P. Ueberla: "Domain Adaptation with Clustered Language Models", Proc. ICASSP 1997, Munich, pp. 807-810.

[14] P. Witschel: "Constructing Linguistic Oriented Language Models for Large Vocabulary Speech Recognition", Proc. 3rd EUROSPEECH 1993, Berlin, pp.1199-1202.