automatic speech recognition,
medical information systems,
language modeling

Jerzy SAS[1]

BUILDING COMPACT LANGUAGE MODELS FOR MEDICAL SPEECH RECOGNITION IN MOBILE DEVICES WITH LIMITED AMOUNT OF MEMORY

The article presents the method of building compact language model for speech recognition in devices with limited amount of memory. Most popularly used bigram word-based language models allow for highly accurate speech recognition but need large amount of memory to store, mainly due to the big number of word bigrams. The method proposed here ranks bigrams according to their importance in speech recognition and replaces explicit estimation of less important bigrams probabilities by probabilities derived from the class-based model. The class-based model is created by assigning words appearing in the corpus to classes corresponding to syntactic properties of words. The classes represent various combinations of part of speech inflectional features like number, case, tense, person etc. In order to maximally reduce the amount of memory necessary to store class-based model, a method that reduces the number of part-of-speech classes has been applied, that merges the classes appearing in stochastically similar contexts in the corpus. The experiments carried out with selected domains of medical speech show that the method allows for 75% reduction of model size without significant loss of speech recognition accuracy.

1. INTRODUCTION

Automatic speech recognition (ASR) is a mature technique of man-machine interface for more than decade. ASR is widely used in medical information systems, in particular in diagnostic image reporting [4, 12, 13]. Recently, due to rapid development of mobile technology, smartphones and tablets are utilized in medical information systems as handy wireless terminals used to enter and access patient-related data. Usage of mobile devices is particularly convenient if ASR is a method of data entering and navigating. Application of ASR to Polish language however raises more problems than it is in the case of English. This is mainly because Polish, similarly to other Slavic languages, is highly inflectional, what leads to much bigger dictionary. Analysis presented in [17] shows that in order to obtain similar out-of-vocabulary coverage (99%) the size of dictionary for Russian has to be almost 7 times bigger than the dictionary for English. Due to lexical and syntactic similarities, we can expect similar relations for Polish. Typical approach to ASR is based on n-gram *language mode*l (LM), usually for *n=2* (*bigram* model). The role of the LM in speech recognition consists in providing prior word occurrence probabilities $p(w_i)$ and the conditional probabilities $p(w_i \mid w_{i-1})$ of word w_i occurrence in the text, provided that the previous word is w_{i-1} for all words w_i, w_{i-1} from the dictionary of permissible words. As the size of dictionary grows, the number of frequently appearing bigrams (pairs of adjacent words) also grows in the language model and in result, the total size of data structures build in RAM for LM representation increases enormously.

Although contemporary mobile devices are equipped with multi-core powerful processors, the amount of available RAM is still short of needs of memory-hungry applications like ASR decoders. The significant fraction of ASR decoder memory is allocated to language model structures. In order to successfully implement ASR in mobile devices, the size of LM must be reduced so as to fit within the limitations of available RAM.

In this article, the works aimed on elaboration of compact language model for domain specific ASR recognition in medical applications are described. Our aim is to create LMs for selected domains of speech application in medicine that are compact enough to be used in mobile device environment and that

[1] Instutute of Informatics, Wroclaw University of Technology, 50-370 Wroclaw, ul.Wyb. Wyspianskiego 27, email: jerzy.sas@pwr.wroc.pl.

assure sufficiently high accuracy of ASR. Reduction of LM size in most cases leads to degradation of LM properties. Therefore the appropriate balance must be kept between the model size and resultant ASR accuracy. The method proposed here consists in combining ordinary *word-based* bigram model with *class-based* LM. In the class-based models, original words are (possibly ambiguously) assigned to classes and the model provides prior and conditional probabilities of classes. Class-based models are known to be much smaller that corresponding word-based LMs, but the ASR accuracy achieved with them is usually lower. In order to obtain appropriate performance/size balance, some bigrams in word-based model can be removed and the probabilities associated with them can be derived from class-based LM. Because most of memory used to store the LM is consumed by explicit bigram representation, reduction of bigrams take effect of almost proportional reduction of required memory size. The problem that needs to be solved is how to select candidates for removal, so as to obtain minimal degradation of ASR accuracy. The approach being described here employs a criterion which evaluates the importance of a bigram for ASR and the similarity of its probability estimation by words-based and class-based LMs.

The article is organized as follows. The next section presents results of related works aimed on stochastic LM construction. Particular attention is paid to class-based models. Section 3 describes the proposed concept of class-based and word-based LMs combination. Some details of class-based model construction using part-of-speech (POS) tagging are also presented there. Section 4 describes the experiments aimed on finding optimal bigrams removal rate and on overall assessment of the proposed method in selected areas of ASR application in medicine. Finally, some conclusions are drawn and practical recommendations for LM construction for ASR in mobile devices are given.

2. RELATED WORKS

We are considering here the typical approach to ASR based on Hidden Markov Model (HMM) of the speaker acoustic properties. The HMM-based approach is described in depth in many published works, for example in the monograph by Jelinek [5]. HMM-based approach is utilized in the majority of research-oriented and commercial ASR systems, as e.g. systems described in [6] and [19]. Accurate acoustic models and language models are crucial for the performance of ASR and in result - for practical usability of this technique. As has been pointed out in the introduction, language models provide prior probabilities of words appearance as well as conditional probabilities of words occurring after determined predecessors. The probabilities are estimated by smoothing maximal likelihood estimates obtained by counting adjacent word occurrences in representative text corpora. Conditional probabilities of successive words $p(w_i \mid w_{i-1})$ can be also estimated taking longer distance co-occurrences into account. It is especially important for loose word order languages like Polish. The problem of taking long distance co-occurrence for Polish ASR has been considered in [14]. The main problem when estimating word bigram probabilities is data sparseness. Even huge text corpora are not sufficient in order to reliably estimate the conditional probabilities $p(w_i \mid w_{i-1})$ for all pairs of words from the dictionary. Smoothing techniques are commonly applied in order to overcome data sparseness problem. The wide review of smoothing methods for LM creation is presented in [3]. Another way of data sparseness problem solution is the application of class-based model. In class based models, words are grouped into classes, so that classes are represented much more frequently that individual words and their probabilities can be estimated more reliably. The fundamental principles of class-based n-gram models have been introduced in [1] and [16]. Although the class-based LMs are more compact than word-based models (due to much lower number of classes) and the class probabilities are estimated more reliably, practical experiments show their lower effectiveness in ASR. One of possible grouping is by *part-of-speech tagging* (POS), where groups are determined by combinations of tags assigned to words, as proposed in [9] and [10]. The next step in improving LMs is to apply combination of word-based and class-based LMs. In [9] authors show that the linear combination of both types of models can result in observable decrease of word error rate in ASR. Similar experiment with Czech and Slovak languages presented in [2] confirms that pure class-based LM performs worse than word-based one and that the linear combination of models outperforms both interpolated LM components. The usefulness of linear interpolation of LMs is also confirmed by results shown in [15] for

Lithuanian language. The linear LM combination can be adaptive, i.e. the combination coefficient can be set individually for each bigram as proposed in [7].

All approaches presented above use complete word-based model in the combination with class-based one, thus they are not useful as far as model size reduction is the ultimate goal. Therefore, another concept is followed in this work, which backs-off the calculation of certain word bigrams to class-based model. The general concept used here is similar to the one described in [10]. The novelty of the approach proposed in this article in relation to [10] consists in applying different criterion of word bigram back-off to class based model and different model reduction procedure, which seems to be more closely aimed on the reduction of the model data structures size in memory. Another element of novelty is that experiments carried out here are based on Polish language, for which earlier conclusions drawn for English may be not quite applicable.

3. COMBINATION OF WORD-BASED AND CLASS-BASED LANGUAGE MODELS

Our aim is to create LM that occupies the amount of memory not greater that specified limit and assures maximal possible ARS accuracy. Unfortunately, the solution to such specified problem seems to be not feasible due to: a) impossibility to formally compute actual ASR accuracy for a model, b) lack of truly optimal methods of language model construction for the sake of ASR. Therefore, the formal requirement must be weakened in order make it more practical. Firstly, we replace the requirement to find the model that maximizes overall ASR accuracy by maximizing the accuracy of speech recognition in the selected set of test utterances. Secondly, we will apply the suboptimal algorithm of model building that reduces the model size by removing some word bigrams that are expected to have minimal impact on the overall ASR accuracy. The obtained procedure obviously in the majority of cases will not create the best possible model that fits to the required memory size limit, but we believe the obtained model will be near-optimal. From the practical point of view, the results will be satisfactory if the ASR accuracy with the obtained LM will be close to the accuracy achievable with the original word-based "big" model.

3.1. COMPACTION OF THE WORD-BASED MODEL

The proposed approach consists in stepwise removal of individual bigrams from the original word-based LM. Because the majority of amount of memory occupied by LM is assigned to bigram probability storage, removal of bigrams effectively reduces the model size. The method combines two component LMs: word-based model LM_W and class-based model LM_C. The component models are created using typical methods. LM_W is created using modified Knesser-Nay smoothing described in [3]. Class-based model is created using POS classes assigned to corpus words by a tagger. Details related to application of POS tagging to LM building for Polish are presented in the next subsection. The method starts with the ordinary LM_W extended with class bigrams taken from LM_C. Each class bigram is a triple:

$$(c_1, c_2, \log(p(c_2 \mid c_1)), \tag{1}$$

where c_1 and c_2 are POS classes and $p(c_2|c_1)$ is the probability that the word from class c_2 appears as the successor of the word belonging to the class c_1. Additionally, in the section of unigrams of LM_W for each word w the set of classes the word belongs to and corresponding conditional probabilities are added:

$$\Theta_w = \{(c_{i_1}, p(c_{i_1}, \mid w)), ..., (c_{i_{1n}}, p(c_{i_n}, \mid w))\}, \tag{2}$$

where n is the number of POS classes that the word w belongs to. Let us denote the initial extended model by LM_E

The method of LM size reduction consists in removal of bigrams appearing explicitly in LM_E until the required final size of RAM data structures necessary to store LM is reached. The removed word bigrams are backed off to class-based bigrams, i.e. if the required word bigram (w_i, w_j) does not appear in the final LM then its conditional probability $p(w_j|w_i)$ is calculated using class-based data. The main problem that needs to be solved is how to select bigrams for removal. The experiments with word-based

and class-based models comparison presented in [2] and our own experiments show that the performance of class-based model in ASR is usually lower in comparison to corresponding word-based LM. Obviously, we want to minimize the deterioration the model performance being the result of bigrams removal as little as possible. The bigram removal will affect the final model performance only marginally if the following conditions are satisfied:

- the removed bigram is very rare, i.e. it is very unlikely that it will appear in the utterance being recognized,
- the bigram probability in LM_W is assumed to be imprecise (usually, due to little number of bigram occurrences in the corpus used to build the model),
- the probability $p(w_j|w_i)$ computed in LM_W is close to the corresponding probability computed using LM_C.

The following formula can be then proposed as the measure of the LM deterioration resulting from the bigram removal from the model:

$$\tau(w_i, w_j) = (1 - \min(\varepsilon(w_i, w_j), 1.0))\, p_{LM_W}(w_i, w_j) \,|\, p_{LM_W}(w_j \mid w_i) - p_{LM_C}(w_j \mid w_i) \,| \tag{3}$$

where $p_{LM_W}(w_j \mid w_i)$ and $p_{LM_C}(w_j \mid w_i)$ are conditional probabilities and $p_{LM_W}(w_i, w_j)$ is the absolute probability of the bigram (w_i,w_j) calculated in LM_W and LM_C models correspondingly. Taking into account the observations of other researchers, we assume here that the bigram probability calculated using LM_W is more accurate than the probability computed using LM_C and it is used here as a kind of baseline for comparison. The product $p_{LM_W}(w_i, w_j) \,|\, p_{LM_W}(w_j \mid w_i) - p_{LM_C}(w_j \mid w_i) \,|$ accounts for the importance of the difference in bigram probabilities computed by both models. The higher is its value, the stronger is the impact of the bigram removal on the final model and the weaker is the indication for such bigram removal. The symbol $\varepsilon(w_i,w_j)$ denotes the radius of the confidence interval in the estimation of bigram probability $p(w_j|w_i)$ in LM_W. The term $(1 - \min(\varepsilon(w_i, w_j), 1.0))$ assesses the accuracy of the probability estimation. For bigrams actually occurring in the corpus, the probability is estimated by smoothing the maximum likelihood estimation:

$$\hat{p}(w_j \mid w_i) = \frac{n(w_i, w_j)}{n(w_i)}, \tag{4}$$

where $n(w_i,w_j)$ is the number of word sequences (w_i,w_j) appearing in the corpus and $n(w_i)$ is the number of w_i word occurrences. In fact, we are estimating the binominal proportion in the series Bernoulli experiments where the number of experiments (observations) is $n(w_i)$ and we consider the appearance of the word w_j next to w_i as the success. Hence, the problem of $p(w_j|w_i)$ estimation is equivalent to the estimation of the success probability in the binominal distribution. The radius of the confidence interval can be used to evaluate the probability estimation accuracy in LM_W model. The due to known deficiencies of the most popular Wald formula based on Bernoulli approximation with normal distribution, according to considerations in [8], the more precise Wilson formula is applied here:

$$\varepsilon(w_i, w_j) = \frac{k\sqrt{n(w_i)}}{n(w_i) + k^2} \sqrt{\frac{n(w_i, w_j)}{n(w_i)} \frac{n(w_i) - n(w_i, w_j)}{n(w_i)} + \frac{k^2}{4n(w_i)}}\ , \tag{5}$$

where $k = \Phi^{-1}(1 - \alpha/2)$. Φ is the accumulated normal distribution and $\alpha=0.95$ is the assumed confidence level.

The complete LM model compaction procedure is very straightforward; and consists of the following steps:

- create LM_W and LM_C models using typical LM building methods,
- merge LM_W and LM_C by extending LM_W with class bigram section and by associating class-related probabilities (1) and (2) to the unigram section of LM_W, the result is the merged model LM_E,
- evaluate the RAM size $S(LM_E)$ necessary to store it in the memory,

- estimate the number of bigrams that need to be removed in order to obtain the required RAM size S that can be allocated to LM as: $n_r = (S(LM_E) - S)/\delta$, where δ is the average number of bytes occupied by the single bigram in RAM,
- sort all bigrams (w_i,w_j) in LM_E by their criterions $\tau(w_i,w_j)$ in the ascending order,
- select the first n_r bigrams from the sorted list and remove them from the model LM_E.

In result we will obtain the model that will fit into available assumed RAM area and only these bigrams explicitly appearing in LM_W will be removed that are expected to have low impact on the LM performance in ASR procedure.

3.2. CLASS-BASED MODEL BUILDING

The class-based LM is created using part-of-speech tagging. The classes are defined by sets of words that are assigned the same combinations of POS tags by a tagger. In order to assign tags to the words in the corpus, TaKIPI tagger described in [11] was used. TaKIPI tagger makes it possible to unambiguously (but not necessary correctly) assign tags to words, taking into account their context in the sentence. The used method of tags coding (*tagset*) with character sequences is described in [18]. In order to build the class-based model the following operations are carried out:

- the original domain-specific corpus is passed through the tagger; in result the tags combination is assigned unambiguously to each word appearance in the corpus,
- each tags combination that appears in the corpus is initially assumed to be an individual class; in result each word appearance is unambiguously assigned to a class,
- specific pseudo-word is assigned to each class; the pseudo-word word is created by concatenating POS tag symbols produced by the tagger for the word appearance in the corpus.
- the original corpus is converted into the class-based corpus; in the class based corpus original words are replaced by pseudo-words corresponding to classes assigned to them,
- finally, the class-based corpus is passed through usual smoothing and discounting procedure and the class-based model is created in the same way that is usually applied do word-based models.

In the case of both word-based and class-based LMs the same modified Knesser-Ney smoothing/discounting procedure is applied. In order to obtain maximally compact class-based LM representation in the memory, it is desirable to index classes using short integer numbers. If the number of classes is not greater than 256 then only one byte for class index is necessary. Unfortunately, Polish is highly inflected language and the number of classes significantly exceeds this limit. For domain-specific corpora for medical speech the number of classes based on POS tagging varies from more than 500 to more than 1000 (see Table 1 for details). The class merging procedure was applied to reduce the number of classes. The applied method is based on the similarity of the class occurrence context. Let us consider a class c_x. Define the vector of context probabilities for this class as:

$$\pi(c_x) = (\hat{p}(c_1 \mid c_x),..., \hat{p}(c_N \mid c_x); \hat{p}(c_X \mid c_1),..., \hat{p}(c_x \mid c_N)), \tag{6}$$

where $\hat{p}(c_i \mid c_x)$ is the unsmoothed maximum likelihood estimate of the probability that the class c_x is followed by the class c_i and N denotes the current number of classes. The probability estimate is calculated using the formula analogous to (4) but applied to class occurrence numbers. Two classes have similar occurrence contexts if their vectors (6) are similar. We used typical Euclidean distance between vectors $\pi(\bullet)$ to measure classes dissimilarity. The reduction of class number is achieved by iterative merging least frequently appearing class with the more frequently populated one, which is most similar to the class being processed. The algorithm is then as follows:

- start with the full set of classes created by word corpus POS tagging,
- sort classes by its number of occurrence $n(c)$ in increasing order,
- compute vectors $\pi(c)$ for all classes
- repeat until required number of classes is obtained,
    - select the least frequently appearing class c_i,

- find the class c_j, $j \neq i$, that minimizes the Euclidean distance between $\pi(c_i)$ and $\pi(c_j)$,
- create the new class c^* by merging c_i and c_j, i.e. update the bigram counts in the following way:

$$\begin{aligned} n(c^*) &= n(c_i) + n(c_j), \\ n(c^*, c_k) &= n(c_i, c_k) + n(c_j, c_k), k \neq i, j, \\ n(c_k, c^*) &= n(c_k, c_j) + n(c_k, c_j), k \neq i, j, \\ n(c^*, c^*) &= n(c_i, c_i) + n(c_i, c_j) + n(c_j, c_i) + n(c_j, c_j) \end{aligned} \tag{7}$$

- update vectors $\pi(c)$ for all classes except of c_i and c_j using new bigram counts (7),
- remove classes c_i and c_j from the sorted sequence and put the new class c^* at the appropriate position in the sorted class sequence.

As a by-product of the class merging procedure we obtain unigram and bigram counts $n(c_i)$, $n(c_i,c_j)$ which can be directly used in building smoothed class-based LM for the reduced number of classes. Knesser-Nay smoothing is used again to obtain final class-based LM.

3.3. COMPUTING WORD SUCCESSION PROBABILITIES WITH COMBINED MODEL

The combined LM is used in speech recognition to compute the probabilities of words $p(w_i|w_j)$, i.e. the probabilities that the next word in the sequence is w_i, provided that the previous word is w_j. If the bigram (w_j,w_i) is explicitly represented in the model (i.e. this bigram occurred in the corpus and was not removed in result of model compaction procedure described in section 3.1) then the probability is explicitly stored in the combined model and can be immediately accessed. Otherwise, back-off to the reduced class-based model is applied and the probability is calculated using the class-based model. Let $\aleph_W(w)$ denotes the set of words v such that bigrams (w, v) are explicitly represented in word part of the model LM_E. The probability $p(w_j|w_i)$ is calculated in the following way:

$$p_{LM}(w_j \mid w_i) = \begin{cases} p_{LM_W}(w_j \mid w_i) & \text{if } w_j \in \aleph_W(w_i) \\ \alpha(w_i)\, p_{LM_C}(w_j \mid w_i) & \text{otherwise} \end{cases}, \tag{8}$$

where $p_{LM_W}(w_j \mid w_i)$ and $p_{LM_C}(w_j \mid w_i)$ are probabilities of the bigram (w_i,w_j) calculated in LM_W and LM_C models correspondingly. $\alpha(w_i)$ denotes the factor necessary to assure the summation of probabilities $\sum_j p(w_j \mid w_i)$ to unity. It is calculated as follows:

$$\alpha(w_i) = \frac{1 - \sum_{w_j \in \aleph_W(w_i)} p_{LM_W}(w_j \mid w_i)}{\sum_{w_j \in \Re - \aleph_W(w_i)} p_{LM_C}(w_j \mid w_i)}. \tag{9}$$

$\Re$ denotes here the set of all words appearing in the model. If the class is unambiguously determined for each word independently on its context then the probability $p_{LM_C}(w_j \mid w_i)$ can be calculated as:

$$p_{LM_C}(w_j \mid w_i) = p(c(w_j) \mid c(w_i))\, p(w_j \mid c(w_j)). \tag{10}$$

The motivation for this formula can be found in [9] and [19]. However in our case, despite the unambiguous class assignment to words in the specific context, it may happen that the same word standing in various contexts is assigned to various classes by the tagger. Therefore, if the wide context of the word is not known the class for the word cannot be determined unambiguously. This is the situation that we have during the speech recognition, so the formula (10) cannot be directly applied. By adapting

general considerations concerning n-grams and ambiguous word-class relation presented in [9] and by assuming that the probability of the class of the word w_j depends only on the preceding class being the actual class of the word w_i the formula (10) can be updated to handle ambiguity in the word assignment to classes:

$$p_{LM_C}(w_j \mid w_i) = \sum_{c \in C(w_j)} \breve{p}(w_j \mid c) \sum_{e \in C(w_i)} p(c \mid e)\breve{p}(e \mid w_i). \tag{11}$$

C(w) denotes here the set of classes the word *w* can belong to, i.e. the set of classes that were assigned by the tagger to various occurrences of the word *w* in the corpus. *p(c|e)* is the probability that the class *c* is the successor of the class *e*. It can be taken directly from class-based LM. $\breve{p}(e \mid w)$ is the probability that the occurrence of the word *w* is actually assigned to the class *e*. $\breve{p}(w \mid c)$ is the probability that the appearance of the word *w* is actually tagged by the class *c*. Both these probabilities can be easily estimated by counting word and class occurrences in the unambiguously tagged corpus. By using formulas (8), (9) and (11), the probability $p(w_j|w_i)$, for any pair of words from $\Re$ can be computed.

4. EXPERIMENTS

In order to investigate how the reduction of LM size affects the speech recognition accuracy, the experiment has been carried out. Five domain specific LMs related to typical areas of speech recognition in medicine have been tested. The tested domains are: computed tomography (CT), magnetic resonance (MR) and ultrasonography (USG) diagnostic image reporting, psychiatric episode describing and general medicine speech. CT, MR and USG domains are rather narrow, the dictionaries are small and typical utterances are used when describing diagnostic images. In result, ASR accuracy in these domains is high. On the opposite end, the LM for psychiatric episodes descriptions is located. The language used in this area is similar to common language. Due to great variety of situations appearing in episodes, large amount of words in the dictionary are necessary. The LM is therefore large and the ARS accuracy is relatively low. The detailed specification of tested LMs is presented in Table 1.

Table 1. Tested speech domains and related LMs specification.

Domain	Corpus size [MB]	Words count	Word bigrams count	Word model size [MB]	Class count	Class bigrams count	Reduced classes bigram count	Reduced class model size [MB]	Combined model size - 80% bigram reduction [MB]
General medicine	79	139567	1322151	13.7	932	51812	33751	0.17	4.3
Psychiatry	122	216491	1970153	19.8	1045	69765	37259	0.21	5.8
CT	58	46827	670714	7.1	599	15969	12320	0.07	2.3
MR	46	33066	420228	4.7	592	15197	13211	0.08	1.7
USG	14	7184	64181	0.7	524	13720	11494	0.06	0.4

The ASR recognizer used in the experiment is based on Large Vocabulary Speech Recognition System Julius ([6]). For each domain, the set of utterances recorded by 4 speakers (2 females, 2 males) was prepared. Each speaker recorded approximately 10 minutes of speech in each domain. The set of utterances for each domain consisted of approximately 4800 words. Speaker dependent approach was applied, i.e. for each speaker the personalized acoustic model was prepared by adapting the gender-specific generic model.

The experiment consisted in building the series of combined LMs using the procedure described in the section 3.1 for various rates of bigrams count reduction ranging from 0.0 to 1.0. 0.0 corresponds to pure word-based LM while 1.0 corresponds to pure class-based model. For each resultant model, the ASR accuracy was evaluated on the test set of utterances. Results are presented in Fig. 1.

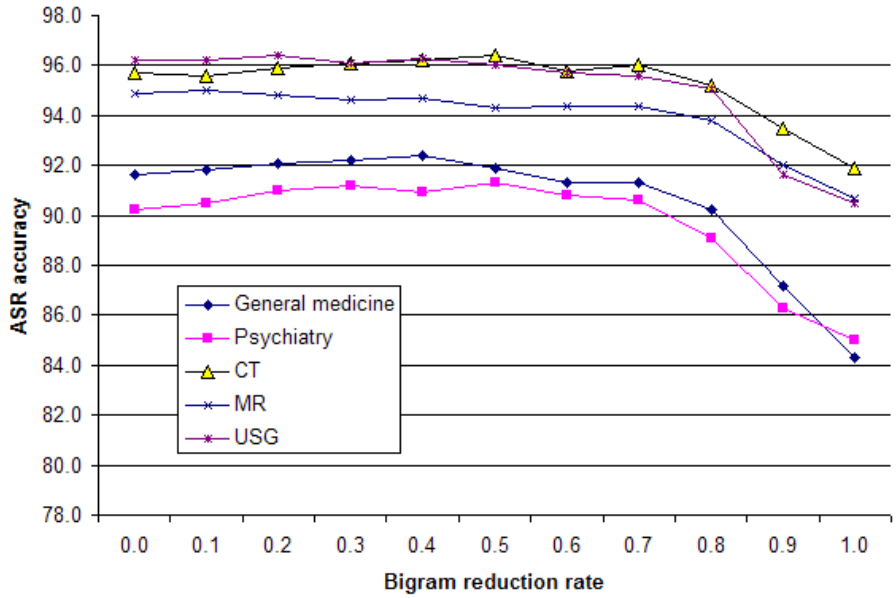

Fig. 1. ASR accuracy dependence on the bigram reduction rate.

It can be observed that elimination of up to approximately 50% of bigrams from word-based LM and replacing their explicit probabilities by the probabilities derived from class-based LM not only does not deteriorate ASR accuracy but even results in small increase of recognition quality. It can be explained by the fact that in the initial iterations of the model reduction procedure, these bigrams are removed, which appear very rarely in the corpus. Due to small number of occurrences, the related $p(w_i|w_{i-1})$ probability may be estimated inaccurately. The probability estimation provided by class-based model may be in such cases more accurate. The reduction of bigrams count up to about 75% results in ASR accuracy still close to the one obtained with pure word-based model. Further reduction of bigrams number leads to rapid deterioration of ASR accuracy.

5. CONCLUSIONS

Experimental results obtained for five LMs in domains related to typical applications of ASR in medicine indicate that the combined LM created using proposed method can be successfully applied without significant downgrade of ASR accuracy. The practical recommendation can be formulated that 80% of bigrams can be eliminated from the model. The removal of bigrams corresponds to almost proportional reduction of the amount of memory necessary to store LM during speech recognition. The final size of the memory necessary to store the combined model with the 80% reduction of bigrams is presented in Table 1. Reduction of LM size is important in the case of ASR implementation in mobile devices. The small pool of mobile device RAM (512 MB in middle class contemporary devices) needs to be shared with other applications, so the limitation of the memory allocated by the application is crucial.

BIBLIOGRAPHY

[1] BROWN P., DESOUZA P. V., MERCER R. L., PIETRA V. J. D., LAI J. C., Class-based n-gram models of natural language, Computational Linguistics, 1992, Vol. 18, No. 1, pp. 467–479.

[2] BRYCHCIN T., KONOPIK M., Morphological based language models for inflectional languages, Proceedings of 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems, 2011, pp. 560–563.

[3] CHEN S., GOODMAN S., An empirical study of smoothing techniques for language modeling, Computer Speech and Language, 1999, Vol. 13, No. 1, pp. 359–394.

[4] DEVINE E., GAEHDE S., CURTIS A., Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports, Journal of American Medical Informatics Association, 2007, Vol. 7, No. 1, pp. 462–468.

[5] JELINEK F., Statistical methods for speech recognition Speech and language processing, The MIT Press, Cambridge, 1998.
[6] LEE A., KAWAHARA T. SHIKANO K., Julius - an open source real-time large vocabulary recognition engine. Proceedings of European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691–1694.
[7] MIKOLOV T., DEORAS A., KOMBRINK S., BURGET L. CERNOCKY J., Empirical evaluation and combination of advanced language modeling techniques, INTERSPEECH, ISCA, 2011, pp. 605–608.
[8] BROWN L.D., CAI T., DASGUPTA A., Interval Estimation for a Binomial Proportion. Statistical Science, 2001, Vol. 16, No. 2, pp. 101–133.
[9] NIESLER T., WHITTAKER E.W.D., WOODLAND P., Comparison of part-of-speech and automatically derived category-based language models for speech recognition, Proceedings of ICASSP 98, 1998, pp. 177–180.
[10] NIESLER T., D., WOODLAND P., Word-to-category backoff language model, CUED/F-INFENG/TR.258, Cambridge University Technical Report, 1996.
[11] PIASECKI M., Polish tagger TaKIPI: Rule based construction and optimization, Task Quarterly, 2007, Vol. 11, No. 1, pp. 151–167.
[12] SAS J., Optimal spoken dialog control in hands-free medical information systems, Journal of Medical Informatics and Technologies, 2008, Vol. 13, pp. 113–120.
[13] SAS J., Application of local bidirectional language model to error correction in polish medical speech recognition, Journal of Medical Informatics and Technologies, 2010, Vol. 15, No. 1, pp. 127–134.
[14] SAS J., ZOLNIEREK A., Distant co-occurrence language model for ASR in loose word order languages, Advances in Intelligent and Soft Computing, Proceedings of International Conference on Computer Recognition Systems Cores, 2011, pp. 767–778.
[15] VAICIUNAS A., KAMINSKAS V., RASKINIS G., Statistical language models of Lithuanian based on word clustering and morphological decomposition, Informatica, 2004, Vol. 15, No. 4, pp. 565–580.
[16] WARD W, ISSAR S,. A class based language model for speech recognition, Proceedings of the Acoustics, Speech, and Signal Processing, ICASSP 96, 1996, pp. 416–418.
[17] WHITTAKER, E., WOODLAND, P., Language modeling for Russian and English using words and classes, Computer Speech and Language, 2003, Vol. 17, No. 1, pp. 87–104.
[18] WOLINSKI M., Morphosyntactic tag system in IPI PAN corpus, Polonica, 2003, No. 22, pp. 39-54.
[19] YOUNG S., EVERMAN G., HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge CB2 1PZ, United Kingdom, 2009.