speech recognition,
language models,
medical information systems

Jerzy SAS[1]

APPLICATION OF LOCAL BIDIRECTIONAL LANGUAGE MODEL TO ERROR CORRECTION IN POLISH MEDICAL SPEECH RECOGNITION

In the paper, the method of short word deletion errors correction in automatic speech recognition is described. Short word deletion errors appear to be a frequent error type in Polish speech recognition. The proposed speech recognition process consists of two stages. At the first stage the utterance is recognized by a typical speech recognizer based on forward bigram language model. At the second stage the word sequence recognized by the first stage recognizer is analyzed and such pairs of adjacent words in the recognized sequence are localized, which are likely to be separated by a short word like conjunction or preposition. The probability of short word appearance in context of found words is evaluated using centered trigrams and backward bigram language model for short words prone to deletion. The set of probabilistic language properties used to correct deletions is called here Local Bidirectional Language Model (in contrast to purely forward or backward model used typically in speech recognition). The decision of short word insertion is based on comparison of deletion error probability of the first stage recognizer and the error probability of the decision based only on centered trigrams and backward model. Despite its simplicity, the method proved to be effective in correcting deletion errors of most frequently appearing Polish prepositions. The method was tested in application to medical spoken reports recognition, where the overall short word deletion error rate was reduced by almost 45%.

1. INTRODUCTION

Automatic speech recognition (ASR) has proved to be useful technology increasing the comfort of medical information systems usage. Although research in speech recognition continues, since early sixties of XX-th century, the recognition accuracy problem is still an issue. The speech recognition of Polish is particularly difficult in comparison to English due to reach inflexion of the language and loose order of words in grammatically correct sentences [9]. In case of application to medical information systems ASR can be often tailored to relatively narrow and specific language domains like diagnostic image reporting for determined imaging modalities, specific areas of electronic patient record etc. Restricting to a narrow language domain results in reducing the size of the dictionary what significantly simplifies speech recognition. What is even more important, typical medical domain specific language contains many typical and frequently repeated phrases or subphrases (word sequences). In result, the language perplexity [2] is reduced. It has been experimentally shown in [3] that word error rate (WER) in ASR almost linearly depends on the perplexity. For this reason, reaching acceptable WER for medical Polish speech at the level of 8-12% is feasible, even with limited training [6, 7].

The errors made by speech recognizer can be assigned to various categories. Some of them are especially annoying to the end user because they appear repeatedly and seem to be quite trivial to correct. One typical kind of such errors consists in erroneous deletion of *short words* (SW). Short word deletion (SWD) error most frequently occurs if the word consists just of single phoneme and a) preceding word ends with acoustically similar phoneme or b) the next word begins with such a phoneme. The SW occurrence where the word is likely to be deleted will be called *deleting contexts*. Examples of such word sequences in Polish are: *w wyniku* (*in result*), *z zebranych badań* (*of collected examinations), w prawym płucu* (*in right lung*). Our experiments with hidden Markov model speech recognizer revealed strong tendency to merge the short word with its successor or predecessor, resulting in SWD error.

The most deletion prone parts of speech are conjunctions and prepositions. In case of domain specific medical texts corpora, the frequency of short prepositions is relatively high. In Polish language two most frequently occurring prepositions that are prone to deletion are *w* (corresponding to English *in*)

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

and *z* (corresponding to English *with*, *from*, *out of* depending on the context). Table 1 shows the relative frequency of their occurrence in deleting contexts in five domain specific medical language corpora: X-ray radiography reporting (XR), computed tomography (CT), magnetic resonance (MRI), ultrasonography (USG), general medicine texts (GM), patient record elements (PR). Expected SWD error rate was calculated with the assumption that the average SWD probability in deleting context is 0.5. Such error rate was observed for *w/*z preposition in uncorrected ASR of medical texts.

Table 1. Short prepositions occurrence statistics for selected domains of medical language.

Domain	Corpus size [MB]	Dictionary size [words]	Number of words in corpus	Number of *w/z* preposition occurrences	Number of *w/z* in deleting contexts	Expected *w/z* deletion error rate [%]
XR	3.7	6814	360076	25750	11463	1.59
CT	7.8	17481	818471	52188	24201	1.48
MRI	10.8	18357	1093134	64604	27773	1.27
USG	10.0	7263	1067547	92047	42054	1.97
GM	80.5	13075	8512143	591203	291926	1.71
PR	12.6	28052	1301503	67864	32778	1.26

The relative frequency of *w/z* preposition occurrence in deleting contexts is about 3%. With the average SW deletion probability in deleting context 0.5 it introduces the overall deletion error rate 1.5%. With the overall error rate close to 8%, short word deletion errors constitute 20% of all errors. Reducing the rate of this kind of error would observably improve the overall ASR accuracy for analyzed areas of medical speech.

In most of ARS toolkits based on HMM paradigm the tendency to false merge of adjacent words can be to some extend controlled by artificial language model parameter - penalty for word insertion [5, 8]. Decreasing the penalty for word insertion leads to the tendency to words over-segmentation. Setting high penalty value on the other hand causes that the adjacent words are often replaced by longer word pronounced similarly or – as in the case being considered in this work – that short word is deleted at all. Unfortunately, setting optimal penalty, that minimizes the overall WER does not minimize SWD error rate. Therefore, another technique is required to reduce this specific type of error.

It can be observed in analyzed corpora of medical texts that there are many words that in most cases are preceded by the one of short prepositions, while the successors of the prepositions are not so specific. It leads to the conclusion that predicting the preposition occurrence preceding its successor can be more accurate that predicting preposition that succeeds another word. In other words: backward bigram language model seems to be more effective than the forward model. The problem of boosting ASR accuracy by improvements in language models is typically approached by tuning the language model smoothing method [3] or by applying forward language model in the first stage of ASR which results in selecting many recognition hypotheses and then by applying backward language model to select or modify the hypotheses found at the first stage [5]. When applied to medical domains listed in Table 1, these methods also increased the overall ASR accuracy, but SWD error rate remained still relatively high.

In this paper a novel two-stage method is proposed. At the first stage the utterance is recognized by typical ASR recognizer based on forward language model. At the second (correction) stage, specific correction rules are applied locally to characteristic fragments of recognized text, where the probability of short preposition is high. In result, missing prepositions are inserted. Experiments conducted with samples of speech form medical language domains showed that overall WER is observably decreased by the correction stage.

2. PROBLEM STATEMENT

Let us consider two-stage speech recognizer. At the first stage the typical speech recognizer Ψ is applied which takes as an input the sequence of acoustic observations O and produces the sequence of recognized words $W=(w_1,w_2,...w_n)$. Each word in actually spoken phrase comes from the finite dictionary *D*. The recognizer utilizes the forward bigram language model and the acoustic model. The recognition procedure is assumed to be erroneous, i.e. the recognized sequence of words does not necessary correspond to actual spoken phrase. We however consider here only SWD errors. Deletion of long words, substitution and insertion errors are not being considered here. The short words subject to erroneous deletion constitute the subset $\sigma \subset D$. The remaining words not subject to deletion errors will constitute the set $\lambda = D \setminus \sigma$. Typically, words in σ are prepositions and conjunctions. Only very small set of words exhibit strong tendency to deletion due to acoustic similarity to trailing/beginning fragments of adjacent words, so σ usually contains just a few words. We additionally assume that words from the set σ do not appear in the spoken phrase one after another, i.e. words from σ are always separated by words belonging to λ. Let

$$V=(v_1,v_2,...v_l),\ \ v_i \in D \tag{1}$$

denotes the actually spoken word sequence. The result of spoken phrase recognition is the word sequence that can be shorter. For the notation simplicity we assume here that the length of recognized sequence will be equal to the actual sequence length but we will introduce the empty word ε. The deletion error consists hence by replacing the actual short word v_i by the empty word ε. If the count of short words in the sequence *V* is *k* then the length *n* of the recognized sequence is not less than *l-k*.

The deletion errors at various positions in the spoken phrase are assumed to be dependant only on the surrounding words. The deletion probability $p_{DE}(v_i,v_{i-1},v_{i+1})$ in the context of surrounding words is assumed to be known as a specific property of first stage the recognizer Ψ:

$$p_{DE}(v_i,v_{i-1},v_{i+1}) = p(\Psi(v_i)=\varepsilon \mid v_{i-1},v_{i+1}),\ \ v_i \in \sigma,\ v_{i-1},v_{i+1} \in \lambda\,. \tag{2}$$

In practice, estimating of this probability individually for all triples of words (v_{i-1},v_i,v_{i+1}) is difficult or infeasible. Therefore words from the dictionary subset λ, which define the deletion context, can be assigned to subsets and deletion probabilities can be estimated rather for groups than for individual context words. Details of grouping will be discussed in the next section.

Because the only possible recognizer error consists in short word deletion, the actually spoken phrase is the one that belongs to the word sequence set defined by the pattern:

$$(X_1,w_1,X_2,w_2,X_3,...,X_n,w_n)\,. \tag{3}$$

X_i in the pattern represents one of words from the set $\sigma \cup \{\varepsilon\}$ if both w_{i-1} and w_i do not belong to σ and just ε if at least one of w_{i-1} and w_i belongs to σ.

Our aim is to select the word sequence form the set defined by the pattern (1) which most likely is the actually spoken phrase. It corresponds to insertion of short words on certain positions of the recognized sequence $(w_1,w_2,...w_n)$. This correction is expected to decrease word error rate of the recognizer. The correction procedure will be based on *Local not smoothed BiDirectional Language Model* (LBLM for short). LBLM consists of two parts: the set of *backward bigram* probability estimates and the set of *centered trigram* probability estimates.

The backward bigram probability is the estimated conditional probability $p(w_{i-1} \mid w_i)$ of preceding word w_{i-1} occurrence conditioned on the succeeding word w_i. By centered trigram probability we mean here the conditional probability $p(w_i \mid w_{i-1}w_{i+1})$ of the word w_i occurrence in two sided context of adjacent words w_{i-1},w_{i+1}. The forward bigram language model is not used at the correction stage because it is already exploited at the first stage. We call the language model used at this stage *local* because only such probabilities need to be represented in LBLM which are related to elements of σ. In case of bigrams we only need $p(w_{i-1} \mid w_i)$ for $w_{i-1} \in \sigma$ and for centered trigrams only such $p(w_i \mid w_{i-1}w_{i+1})$ are required, where $w_i \in \sigma$. Backward bigrams and centered trigrams can be easily calculated using sufficiently large corpus of domain specific texts.

3. APPLICATION OF LOCAL LANGUAGE MODEL TO SWD ERROR CORRECTION

The speech recognition procedure proposed here consists of two stages. At the first stage a typical ASR recognizer based on forward bigram language model is applied. The typical approach applied at this stage is described in many fundamental articles and books on ASR, e.g. in [4]. The recognized word sequence is then passed through the correction procedure aimed on short word deletion error elimination. Because typical approach is applied on the first stage, we describe here only the correction procedure. More detailed description of ASR procedure applied at the first stage can be found in [6].

3.1. LOCAL SWD CORRECTION RULE

In order to find the most likely sequence matching the pattern (3), sources of errors should be considered. The errors can appear only at such positions of symbols X_i in (3) that can be substituted by the symbols from σ, i.e. that are surrounded by words w_{i-1} and w_i that are elements of λ. The correction will consist in possible insertion of the short word at the corresponding position. According to the assumption on independency of deletion errors, each local correction can be considered independently.

Consider the subsequence $w_{i-1}X_i w_i$ where $X_i \in \sigma \cup \{\varepsilon\}$ and $w_{i-1}, w_i \in \lambda$. The correction consists in selecting a word which substitutes X_i from its domain i.e. from the set $\sigma \cup \{\varepsilon\}$. We can leave the first stage recognizer decision (ε) or insert a word from the set σ. The decision should be made so as to minimize the probability of error. If we trust Ψ recognizer and decide not to insert any word from the set σ then the probability of error at this position can be calculated as:

$$p_{E\Psi}(w_{i-1}, w_{i+1}) = \sum_{w\in\sigma} p_{DE}(w, w_{i-1}, v_{i+1})\, p(w \mid w_{i-1}, w_{i+1})\,. \tag{4}$$

The probabilities $p(w \mid w_{i-1}w_{i+1})$ are elements of LBLM, $p_{DE}(w, w_{i-1}, w_{i+1})$ are estimated properties of the first stage recognizer Ψ. Note that according to the assumption that the only possible errors are SWD, the summation in (4) is done only over the elements of σ; the case of ε does not have to be taken into account.

Alternately, the recognizer Ψ can discredited as not sufficiently reliable and the insertion decision can be made merely using LBLM. In order to minimize the error probability, such word $w \in \sigma \cup \{\varepsilon\}$ should be selected that maximizes the trigram probability $p(w \mid w_{i-1}, w_{i+1})$. The conditional SWD error probability for this context is:

$$p_{ELM} = 1 - \max_{w\in\sigma\cup\{\varepsilon\}} (p(w \mid w_{i-1}, w_{i+1})\,. \tag{5}$$

The obvious correction rule for substitution of X_{i} in the context w_{i-1}, w_{i+1} is:

```
if   $p_{E\Psi}(w_{i-1}, w_{i+1}) \le p_{ELM}(w_{i-1}, w_{i+1})$
     $X_i = \varepsilon$
else
     $X_i = \arg\max_{w\in\sigma\cup\{\varepsilon\}} p(w \mid w_{i-1}w_{i+1})$
```

The complete two-stage ASR algorithm can be formulated as follows:

```
apply the recognizer $\Psi$ to the acoustic observations $O$;
```

```
let $\Psi(O)=W=(w_1,_2,...,w_n)$ ;
for each pair of adjacent words in W such that $w_{i-1},w_i \in \lambda$ :
    if $p_{E\Psi}(w_{i-1},w_{i+1}) > p_{ELM}(w_{i-1},w_{i+1})$
        insert the word $w^* = \arg\max_{w\in\sigma\cup\{\varepsilon\}} p(w\,|\,w_{i-1}w_{i+1})$ between $w_{i-1},w_i$ .
```

3.2. PRACTICAL ISSUES

The first practical question that needs to be answering when applying proposed method in practice is which words from the language dictionary should be assigned to the set of short words σ. We observed that words consisting of more than two phonemes in their phonetic translation are very rarely subject to deletion errors. The practical recommendation is therefore to restrict *σ* set to words which phonetic translation is not longer than two phonemes. We also observed that the words consisting just of single phone corresponding to fricative consonants (*s*, *c*, *h, ch, f*) are most frequently deleted by ASR recognizer. In our works aimed on domain specific medical speech recognition we found that the most frequently occurring short words of this type are Polish prepositions *w* (*in*) and *z* (*from*). Detailed analysis of domain specific text corpora will be given in the next section. Finally, the set *σ* consists just of these two prepositions.

The next problem is how to estimate the probabilities used in the formula (4). Probabilities $p_{DE}(w,w_{i-1},w_{i+1})$ characterize the recognizer *Ψ* tendency to delete word *w* in the context w_{i-1},w_{i+1}. In practical applications where the dictionary size is of the order of a few thousands, the number of different context reaches millions and individual estimation of contextual deletion probability is infeasible. Contexts can be however merged in groups, where due to phonetic similarities, the deletion probability can be assumed similar. The context words can be grouped according to the phone directly adjacent to the short word *w*. The following situations can be distinguished:

- A - trailing/leading phone of the context word is the same as directly adjacent phone in *w*,
- B - adjacent phone in context word is a vowel,
- C - adjacent phone in context word is a fricative consonant,
- D - adjacent phone in context word is non-fricative consonant,
- E - adjacent word in silence (beginning or end of the phrase).

Equivalence classes of two-sided contexts based on combinations of left and right contexts are presented in Table 2. The numbers in cells of the table define context groups. Observe that group 1 is not disjoint with remaining groups. If the context found in the phrase being corrected belongs to the class 1 and to any other one then class 1 should be used. In case of limited amount of training data the groups can be further merged.

Table 2. Context classes based on phonetic similarities of adjacent phones (columns represent left context, rows represent right context).

	A	B	C	D	E
A	**1**	**2**	**2**	**2**	1
B	**2**	**3**	**4**	**3**	3
C	**2**	**4**	**5**	**5**	5
D	**2**	**3**	**5**	**6**	6
E	1	3	5	6	X

The most accurate method for p_{DE} estimation for contexts in 6 distinguished classes is to extract utterances containing *σ* member occurrences in each of seven distinguished class contexts form the training set used to build acoustic model for a speaker. The deletion probabilities can be then estimated by recognizing held-out utterances with the model created using remaining elements of the training set. The method is however ineffective if limited amount of training utterances is available.

Alternately, the method can be proposed which uses the hidden Markov model as a generator of artificial utterances. The training of the acoustic model consists in estimating parameters (state transition probabilities, observation emission probability distribution function parameters) of the hidden Markov model of individual phonemes. In order to simulate spoken utterance, which phonetic translation is know, the utterance model is constructed by concatenating models of subsequent phonemes. Then the model is randomly run started from its initial state. The model traverses from state to state until the terminal state is reached. The observations emitted when entering each emitting state constitute the observation sequence of the simulated utterance. The procedure of artificial utterance simulation is described in detail in [7]. In this way, practically unlimited set of testing utterances can be gathered without engaging a human speaker. For each context class and for each element of σ, necessary number of artificial utterances can be created. The utterances are then recognized by the recognizer *Ψ*, the frequency of deletions is counted and finally maximum likelihood estimation of p_{DE} can be calculated. The simulated subphrase always consists of words $w_{i-1}, w_{i+1} \in \lambda$ and $w \in \sigma$. If the recognizer incorrectly recognizes context words from the set λ then the recognition is not taken into account and the random creation of artificial utterance is repeated.

The last element needed to evaluate $p_{E\Psi}$ and p_{ELM} according to (4) and (5) is the centered trigram conditional probability $p(w \mid w_{i-1}, w_{i+1})$. Because we consider here only cases where *w* is either element of σ or the empty word ε, then the maximum likelihood estimate for it can be calculated as:

$$\overline{p}(w \mid w_{i-1}, w_{i+1}) = \frac{c(w_{i-1}, w, w_{i+1})}{\sum_{v \in \sigma} c(w_{i-1}, v, w_{i+1}) + c(w_{i-1}, w_{i+1})}, \tag{6}$$

where $c(w_{i-1}, w, w_{i+1})$ denotes the count of trigram w_{i-1}, w, w_{i+1} occurrences and $c(w_{i-1}, w_{i+1})$ is the count of bigram w_{i-1}, w_{i+1} occurrences in the text corpus used to build LBLM. If the number of trigram and bigram occurrences used in denominator of (6) is not large enough to reliably estimate the probability the backing-off technique is applied [1, 3]. According to Katz recommendation [1], the minimal number of n-gram occurrences used in estimation language model probabilities is less than certain number *m* (Katz suggests *m*=5) , then lower level *(n-1)*-grams should be used. In our case we are considering trigrams and bigrams based on w_{i-1}, w_{i+1}. If $\sum_{v \in \sigma} c(w_{i-1}, v, w_{i+1}) + c(w_{i-1}, w_{i+1})$ is less than *m* then right context bigrams are used, i.e. $\overline{p}(w \mid w_{i-1}, w_{i+1})$ is approximated as:

$$p(w \mid w_{i-1}, w_{i+1}) \approx \overline{p}(w \mid w_{i+1}) = \frac{c(w, w_{i+1})}{c(w_{i+1})}. \tag{7}$$

If the count of w_i, w_{i+1} bigram occurrences is also not sufficient, i.e. the conditional probability cannot be estimated accurately then the correction rule is not applied and the primary recognition of *Ψ* remains uncorrected. Only the right context information is used here because left context was already exploited by the first stage recognizer *Ψ* based on the forward language model.

4. EXPERIMENTS

The performance of the correction method described in this article was test using domain specific language models related to typical areas of medical reporting, where ASR application is justifiable: for diagnostic imaging in CT MRI, X-Ray, USG and MRI and for patient record related to medical history and treatment. The characteristics of domain corpora used to create language models and testing sets for these domains are presented in Table 1 in the introduction.

The testing sets were created by recording of approximately 10 minutes of speech for each of tested domains by each of speaker participating in an experiment. Only such test utterances were selected from the domain specific corpora, which contain at least one occurrence of SW. The test utterances extracted from the domain text corpora were recorded by three speakers: two males and one female. Speaker dependent acoustic models were used in the experiment. The models were created individually for each speaker with approximately 20 minutes of speech. Both language models and acoustic models were created using HTK toolkit [8]. The decoder described in [5] was used as the first stage recognizer. The set of short words subject to corrections was limited just to two Polish prepositions (*w* and *z*). This restriction is motivated by our observation in practical application of ASR to domains of medical language, where these prepositions exhibited most strong tendency to SWD errors. The deletion probability of the first stage recognizer $p_{DE}(w, w_{i-1}, w_{i+1})$ was estimated using the set of recorded test utterances created by selecting occurrences of *w* and *z* in phrases found in domain corpora. Due to the limited amount of recorded utterances only three categories of deleting contexts were distinguished. They were obtained by merging groups defined in Table 2. The group *I* corresponds to sum of classes 1 and 2 (the SW is adjacent to the same phoneme that constitutes SW), group *II* is a sum of classes 4 and 5 (at least one fricative consonant phoneme is adjacent to SW), group *III* corresponds to all remaining classes. The deletion probabilities for both SW prepositions were estimated as: $p_{DE}(\Psi(SW) = \varepsilon \mid I) = 0.85$, $p_{DE}(\Psi(SW) = \varepsilon \mid II) = 0.35$ and $p_{DE}(\Psi(SW) = \varepsilon \mid III) = 0.15$.

The aim of the experiment was to evaluate the relative SWD error reduction resulting from application of the correction stage. The method effectiveness can be measured by the gain factor *q* calculated as:

$$q = \frac{n_{CI} - n_{FI}}{n_{D\Psi}}, \tag{8}$$

where $n_{D\Psi}$ is the count of SWD errors introduced by the first stage recognizer, n_{CI} is the number of correctly inserted short words at the correction stage and n_{FI} is the number of false insertions (i.e. the short word was inserted between words, where actually there were no element of σ or incorrect word was inserted at the place of missing one). The results of experiments are presented in Table 3. Data related to SWD error frequencies are averaged over all speakers participating in the experiment. It should be stressed that absolute SWD error rate obtained in the experiment is much higher than in the case of practical ASR application in corresponding domain. It is because the utterances were selected so as to obtain sufficient number of SW occurrences assuring reliable estimation of SWD error rates. In result, the frequency of SW occurrence was much higher in the test set in comparison to the original domain corpus specified in Table 1.

Table 3. SWD error reduction rate for selected domains of medical speech.

Domain	Number of words in the test set	Number of *w/z* words in the test set	Uncorrected case (after 1st stage)		Corrected case (after 2nd stage)		Gain factor *q*
			Number of SWD errors	SWD error rate [%]	Number of SWD errors [2]	SWD error rate [%]	
XR	1473	205	117	7.94	61	4.14	0.48
CT	1264	179	97	7.67	51	4.03	0.47
MRI	1320	180	95	7.19	58	4.39	0.39
USG	1477	195	110	7.44	53	3.59	0.52
GM	1363	212	127	9.31	73	5.36	0.43
PR	1248	181	85	4.69	51	4.09	0.40
Avg:	8145	1152	631	7.74	347	4.26	0.45

[2] The error rate specified here includes false insertion errors

5. CONCLUSIONS

The SWD error correction method proposed in this article observably decreased deletion error rate in applications of ASR in specific areas of medical spoken language. The correction procedure utilizes centered trigrams and backward bigram model in order to find potential deletions of short prepositions. Despite the simplicity of method based merely on language model properties, the relative reduction of SWD errors is in the range 40-50% depending on the language domain. It is probably caused by the fact that in domain languages being considered here the appearance of short words is specific for two-sided contexts, which is not exploited by typical speech recognizer based on the forward bigram language model. The preposition occurrences can be also better predicted with its right context than with the left one. Therefore backward language model is effective if finding short preposition deletions.

The proposed method is based merely on language properties contained in LBLM. The acoustic evidence is not used by the correction procedure. The performance of the correction stage could be probably improved by rescoring the utterance fragment comprising context words with the recognizer that uses both LBLM and the acoustic observations. It however complicates the correction procedure logic and may slow down the recognition process significantly.

BIBLIOGRAPHY

[1] KATZ S.M., Estimation of Probabilities for Sparse Data for the Language Model Component for the Speech Recognition, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 35(3), 1987, pp. 400–401.

[2] JELINEK F., Statistical Methods for Speech Recognition, MIT Press, Cambridge, Massachusetts, 1997.

[3] CHEN S.F., GOODMAN J., An Empirical Study of Smoothing Techniques for Language Modeling, Proc. of the 34th Annual Meeting on Association for Computational Linguistics, 1996, pp. 310–318.

[4] JURAFSKY D., MARTIN J., Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, New Jersey, 2000.

[5] LEE A., KAWAHARA T., SHIKANO K., Julius – an Open Source Real–Time Large Vocabulary Recognition Engine, Proc. of European Conference on Speech Communication and Technology (EUROSPEECH), 2001, pp. 1691–1694.

[6] HNATKOWSKA B., SAS J., Application of Automatic Speech Recognition to Medical Reports Spoken in Polish, Journal of Medical Informatics & Technologies, Vol. 12, 2008, pp. 223–230.

[7] SAS J., Optimal Spoken Dialog Control in Hands–Free Medical Information Systems, Journal of Medical Informatics & Technologies, Vol. 13, 2009, pp. 113–120.

[8] YOUNG S., EVERMAN G., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2009.

[9] ZIOLKO B., SKURZOK D., ZIOLKO M., Word n–Grams for Polish, Proc. of 10th IASTED Int. Conf. on Artificial Intelligence and Applications (AIA 2010), 2010, pp. 197–201.