INFORMATICA, 2004, Vol. 15, No. 4, 565–580

Statistical Language Models of Lithuanian Based on Word Clustering and Morphological Decomposition

Airenas VAIČIŪNAS, Vytautas KAMINSKAS

Department of Applied Informatics, Vytautas Magnus University
Vileikos 8, LT-3035 Kaunas, Lithuania
e-mail: airenas@freemail.lt, V.Kaminskas@if.vdu.lt

Gailius RAŠKINIS

Center of Computational Linguistics, Vytautas Magnus University
Donelaičio 52, LT-3000 Kaunas, Lithuania
e-mail: idgara@vdu.lt

Received: March 2004

Abstract. This paper describes our research on statistical language modeling of Lithuanian. The idea of improving sparse n-gram models of highly inflected Lithuanian language by interpolating them with complex n-gram models based on word clustering and morphological word decomposition was investigated. Words, word base forms and part-of-speech tags were clustered into 50 to 5000 automatically generated classes. Multiple 3-gram and 4-gram class-based language models were built and evaluated on Lithuanian text corpus, which contained 85 million words. Class-based models linearly interpolated with the 3-gram model led up to a 13% reduction in the perplexity compared with the baseline 3-gram model. Morphological models decreased out-of-vocabulary word rate from 1.5% to 1.02%.

Key words: language models, n-grams, class-based models, morphology, inflections, interpolation, perplexity reduction, out-of-vocabulary words.

1. Introduction

Statistical language modeling attempts to capture and exploit regularities in natural language. Statistical language models (LM) have become key components for large vocabulary continuous speech recognition (LVCSR) systems. These models provide prior probabilities that are used to rate hypothesized sentences and to disambiguate their acoustical similarities.

During the last few decades, much experimental work has been done in the field of statistical language modeling covering widespread world languages such as English, French, and German. Unfortunately, statistical language modeling of Lithuanian is still at its infancy stage due to two primary reasons:

1. There was no demand for Lithuanian LM as there were no attempts to build Lithuanian LVCSR systems. Lithuanian speech recognition research was basically limited to solving isolated-word ASR tasks of small to medium vocabulary (Lipeika *et al.*, 2002; Filipovič, 2003; Laurinčiukaitė, 2003; Raškinis and Raškinienė, 2003).
2. Until very recently, there were no Lithuanian text corpora large enough to serve as a basis for building statistical LM.

Lithuanian language is highly inflected, i.e., new words are easily formed by inflectional affixation. This property of a language results in difficulties of statistical modeling known as huge vocabulary size, model sparseness, high perplexity, and a high out-of-vocabulary (OOV) word rate. An attempt to overcome the abovementioned difficulties by applying word parsing into stems and endings was described in our previous paper (Vaičiūnas and Raškinis, 2003). In this paper, we investigate alternative approaches of modeling highly inflected languages, such as conventional class-based modeling and morphological modeling based on word decomposition into word base form and part-of-speech tag.

1.1. *n-gram Language Models*

The aim of the statistical LM is to return the a priori probability for every word sequence $w_1, \ldots, w_n$. The definition of joint probability states that:

$$\begin{aligned} P(w_1 \ldots w_n) &= P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\ldots P(w_n|w_1 \ldots w_{n-1}) \\ &= \prod_{i=1}^{n} P(w_i|w_1 \ldots w_{i-1}). \end{aligned} \tag{1}$$

Though there are no effective methods for calculating the probability (1) accurately, as it would require too much data, it can be approximated by a series of probabilities based on a limited number of previous words. LMs built using this approach are called n-gram LMs. The most frequently used n-gram LM is 3-gram LM which is based on the conditional probability of seeing one word given the two preceding words:

$$P(w_1 \ldots w_N) \approx \prod_{i=1}^{N} P(w_i|w_{i-2}, w_{i-1}). \tag{2}$$

Conditional probabilities of a 3-gram LM can be estimated by formula

$$\widehat{P}_{W^3}(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}, \tag{3}$$

where $C(\cdot\,)$ denotes the count function in training corpus.

Because any particular training corpus is finite, the n-gram LMs have a very large number of zero probability n-grams that should really have some non-zero probability. This is known as model sparseness problem. Good-Turing, Witten-Bell (Jelinek, 2001) and Knesser-Ney (Chen and Goodman, 1999) smoothing techniques are often applied for re-evaluating the zero-probability n-grams. These techniques operate by subtracting probability mass from the observed n-grams and redistributing it to the zero-probability n-grams. Backing-off (Katz, 1987) is also used and refers to using a probability estimate from a more general n'-gram distribution when the estimate from the specific n-gram distribution ($n' < n$) is non-existent or unreliable.

1.2. *Class-Based n-gram Language Models*

Class-based modeling is another approach for dealing with model sparseness problem. Class-based LM requires all words of a given language to be clustered into equivalence classes. Class-based n-gram LM can be thought of as a generalization of word-based n-gram model. It estimates the conditional probability of seeing one word given the two preceding words as follows:

$$\widehat{P}_{C(W^3)}(w_i|w_{i-2},w_{i-1}) = \widehat{P}_{WC(W)}(w_i|c_i) \cdot \widehat{P}_{C^3(W)}(c_i|c_{i-2},c_{i-1}), \quad w_i \in c_i. \tag{4}$$

Here, $\widehat{P}_{C^3(W)}(c_i|c_{i-2},c_{i-1})$ is the estimate of the conditional probability of seeing class c_i given the two previous classes c_{i-2}, c_{i-1}, and $\widehat{P}_{WC(W)}(w_i|c_i)$ is obtained from the 1-gram distribution of words assigned to the class c_i. The formula (4) describes the case when every word w_i is assigned just to one class c_i. Conditional probability can be also defined for the case when words are assigned to multiple classes (Huang *et al.*, 2001).

1.3. *Combining Multiple n-gram Language Models*

Standalone class-based models of type (4) usually perform poorly and must be combined with some other LMs. Linear interpolation is the simplest way of combining statistical LMs. For instance, standalone class-based 3-gram model $P_{C(W^3)}(w_i|w_{i-2},w_{i-1})$ and a word-based 3-gram model $P_{W^3}(w_i|w_{i-2},w_{i-1})$ can be interpolated in the following way:

$$\begin{aligned}&\widehat{P}_{W^3+C(W^3)}(w_i|w_{i-2},w_{i-1})\\&\quad = \lambda\widehat{P}_{W^3}(w_i|w_{i-2},w_{i-1}) + (1-\lambda)\widehat{P}_{C(W^3)}(w_i|w_{i-2},w_{i-1}),\end{aligned} \tag{5}$$

here $0 \leqslant \lambda \leqslant 1$ is the interpolation parameter optimized on the validation corpus.

1.4. *Evaluating and Comparing Language Models*

Statistical language models are evaluated on the homogenous corpus partition (called test corpus) that was excluded from model training and optimization. Models are evaluated by measuring their perplexity and out-of-vocabulary (OOV) word rate.

Perplexity *PP* refers to how many different equally probable words a statistical LM expects to appear in average for a particular type of a context:

$$PP = 2^{LP}, \text{ where } LP = \frac{1}{n}\sum_{i=1}^{n} \log P(w_i|w_{i-2}, w_{i-1}), \tag{6}$$

and n is a count of words in the test corpus.

OOV rate refers to the percentage of words that are not covered by the language model and found within test corpus.

2. Related Work

Class-based modeling is among the most popular techniques for reducing huge-vocabulary-related sparseness of statistical language models. Class-based LM requires all words of a given vocabulary to be clustered into equivalence classes. Both manual and automatic word clustering techniques are used. Automatic word clustering is reported to outperform manual (Maltese *et al.*, 2001) but is computationally very expensive. Automatic word clustering is often based on iterative hill climbing (Whittaker, 2000) or simulated annealing (Maltese *et al.*, 2001) search techniques. Standalone class-based models usually perform poorly[1] and are combined with some other LMs. Many researchers demonstrated that linear interpolation (5) of a standalone class-based LM and a standard word 3-gram model reduced model perplexity. Perplexity reduction ranging from 4.48% to 49.6% was reported for English, French, Greek, Italian and Spanish (Maltese *et al.*, 2001), up to 19% reduction was reported for Russian (Whittaker, 2000). There exist more sophisticated techniques for combining two statistical LMs, such as log-linear interpolation (Klakow, 1998) and interpolation by dynamically updating model weights λ (Kobayashi *et al.*, 1999). The maximum entropy approach (Jelinek, 2001) for combining LMs is also attractive as it combines features from different and sometimes disparate models into one model instead of combining models themselves.

Word decomposition into a sequence of particles is another popular technique for reducing vocabulary size, extending coverage of a language model and reducing out-of-vocabulary word rate. Words are decomposed either by means of some string matching strategy, such as longest suffix match principle (Sepesy Maucec *et al.*, 2001; Vaičiūnas and Raškinis, 2003), or by means of a morphological analyzer (Siivola *et al.*, 2001; Ircing *et al.*, 2001; Martins et al, 1999). Iterative search techniques are often used in a hope to discover an optimal word decomposition (Whittaker, 2000). Particles themselves may represent either sub-strings of the original word: morphemes, stems, inflectional suffixes (Martins *et al.*, 1999, Ircing *et al.*, 2001) or some derivative information about the word, such as word base form[2] and its part of speech (Siivola *et al.*, 2001). Sometimes a single

[1] Russian standalone class-based model outperforming standard word 3-gram (Whittaker, 2000) is an exception to this rule.

[2] For instance, the infinitive is the base form of the verb.

particle-based model is built out of the sequence of word particles (Martins *et al.*, 1999). As an alternative, two separate models for word stems and for inflectional suffixes can be built and then combined into one model (Siivola *et al.*, 2001; Ircing *et al.*, 2001).

The abovementioned word decomposition techniques were reported to reduce the vocabulary by 29% for Portuguese LM (Martins *et al.*, 1999). The vocabulary of Czech LM was reduced by 58% and OOV rate dropped from 8.56% to 4.62% (Ircing *et al.*, 2001). However, the perplexity of particle-based models is greater in comparison to word-based models. Particle-based methods are often paired with complementary techniques such as automatic topic detection and topic adaptation. Topic adaptation can be achieved by combining multiple topic-specific models as reported for Slovenian (Sepesy Maucec *et al.*, 2001) and Finnish (Siivola *et al.*, 2001).

Though it was shown that in many cases perplexity of a statistical LM embedded within a LVCSR system correlates with the word error rate of the LVCSR system (Rosenfeld, 2000) it is not always the case for highly inflected languages. For instance, changes in WER dues to LM of Czech LVCSR system (Ircing *et al.*, 2001) and that of Portuguese (Martins *et al.*, 1999) were insignificant.

3. Investigation of Statistical Language Models of Lithuanian

3.1. *Modeling Data and Tools*

Our experiments were based on a 84,202,576 word Lithuanian text corpus[3] which had vocabulary of $V_W = 1422746$ distinct words[4]. This corpus represented a great variety of genres and topics of the present day written Lithuanian. It included texts from local and national newspapers and journals, law and administrative documents, novels, and books on such specific subjects as history and philosophy.

The whole corpus was manually divided into three parts: 98% were used for training, 1% for validation (optimization), and 1% for testing of our models. While subdividing the corpus we tried to keep the same proportions of text genres within training, validation and testing parts. We used some text clearing. All punctuation was removed and all numbers where replaced by the same tag <num>.

Majority of our investigations were carried out using CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson and Rosenfeld, 1997). We have extended this toolkit to handle vocabularies $V_W > 65$k words and developed our own word clustering tools. Morphological analysis was performed by morphological lemmatizer of Lithuanian (Zinkevičius, 2000). We have extended its functionality by adding support of the most frequent family names of Lithuanian and foreign origin.

[3] The corpus was compiled by the Center of Computational Linguistics at Vytautas Magnus University, Kaunas, Lithuania (Marcinkevičienė, 2000). At the time we were finishing this article, the size of this corpus reached 100 million words.

[4] Words correspond to character strings and include misspellings, names, non-Lithuanian words.

3.2. *Baseline Word n-gram Language Model*

Our baseline language model was defined to be the conventional word-based 3-gram $\widehat{P}_{W^3}(w_i|w_{i-2},w_{i-1})$ including all singleton n-grams and smoothed using Katz back off and Good-Turing discounting techniques (Jelinek, 2001).

It can be seen from the Tables 1 and 2 that word 3-gram language models of inflected free word order languages (Lithuanian, Russian) have much greater perplexities and OOV rates in comparison to less inflected fixed word order languages (English). Though LM based on 65k word vocabulary is sufficient for English LVCSR applications (1% OOV rate) it cannot be applied to Lithuanian LVCSR (11% OOV rate). Models having lower perplexity and OOV rates must be developed for using them within Lithuanian LVCSR systems.

3.3. *Class-based n-gram Language Models*

Prior to building class-based model $\widehat{P}_{C(W^3)}(w_i|w_{i-2},w_{i-1})(4)$, we clustered all words into equivalence classes. Let V be the vocabulary size, K be the desired number of classes and *IT* be the number of iterations. An iterative hill climbing clustering technique described in (Whittaker, 2000) was used:

Table 1

Perplexities and OOV rates of the baseline Lithuanian LM for various vocabulary sizes V_W

Vocabulary, V_W	Perplexity, PP_{W^3}	OOV, %
65k	414.30	10.92
100k	449.76	8.25
200k	512.75	5.05
400k	570.32	3.07
500k	588.03	2.58
800k	621.69	1.82
1000k	631.09	1.62

Table 2

Perplexity and OOV rate of baseline Lithuanian LM for $V_W = 65$k in comparison with corresponding English and Russian (Whitaker, 2000) perplexities and OOV rates

Language	**Perplexity,** PP_{W^3}	**OOV, %**
Lithuanian	414.30	10.92
Russian	387.40	7.60
English	208.40	1.10

```
Cluster (V, K, IT)
    sort words $w_i$, i=1,...,V in order of decreasing frequency
    move word $w_i$ to class $c_i$ i=1, ..., K-1
    move all remaining words $w_i$, i=K-1,..., V to the class $c_K$
    for a fixed number of iterations IT
        for each word $w_i$, i=1, ..., V
            for each class $c_j$ j=1, ..., K
                move word $w_i$ to class $c_j$
                calculate mutual information criteria I
            move word $w_i$ to the class $c_{min}$ giving minimum I
```

Here average mutual information criterion I is defined:

$$I = \sum_{i=1}^{K} \sum_{j=1}^{K} P(c_i, c_j) \log \frac{P(c_j|c_i)}{P(c_j)}. \tag{7}$$

It can be rewritten using 2-gram estimates by

$$\widehat{I} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{K} C(c_i, c_j) \log \frac{C(c_i, c_j)n}{C(c_i)C(c_j)}, \tag{8}$$

where $C(\cdot)$ denotes count function, and n is the total number of words in training corpus.

The experiments were conducted using two fixed vocabulary sizes $V_W = 65$k, 1000k, multiple sets $K_W = 102, 202, 502, 1002, 2002, 3002, 4002, 5002$[5] of classes and run for *IT* = 2 iterations. Class 3-gram $\widehat{P}_{C^3(W)}(c_i|c_{i-2}, c_{i-1})$ estimates were smoothed using Katz back off and Witten-Bell method (Jurafsky, 2000). Word probability given its class was estimated by

$$\widehat{P}_{WC(W)}(w|c) = \frac{C(w)}{C(c)}. \tag{9}$$

Linear interpolation (5) was used to join class-based and word-based 3-gram models together. The optimum value λ was obtained by minimizing perplexity on validation data. Clustering results are illustrated and the performance of class-based n-gram models is shown in Tables 3 to 5.

We run the same clustering algorithm for $IT = 15$ iterations and found that perplexity still improves with every iteration but the improvement becomes insignificant for $IT > 2$. Thus $IT = 2$ iterations can be considered enough to reach the local minimum of the clustering algorithm (see Fig. 1).

[5] Two classes are reserved for unknown words and numbers.

Table 3

10 words of the first 4 classes in alphabetical order ($V_W = 65$k, $K_W = 5002$, $IT = 2$)

Class	10 first words of the class in alphabetical order
1.	aiškiu, amžinu, auksiniu, aukštu, aukštuoju, aukščiausiuoju, aštriu, baisiu, begaliniu, blogu, …
2.	adatą, akmenį, aplanką, apsiaustą, automatą, bandelę, batą, bokalą, bulvę, buteliuką, …
3.	akmeniniai, anoniminiai, antriniai, apatiniai, apsauginiai, augaliniai, biologiniai, branduoliniai, buitiniai, cheminiai, …
4.	adomą, albertui, albertą, albinui, albiną, aldonai, aldoną, aleksandrui, aleksandrą, alfonsui, …

Table 4

Perplexities of class-based and interpolated LMs for $V_W = 65$k

Number of classes, K_W	Class model perplexity, $PP_{C(W^3)}$	Interpolated perplexity, $PP_{W^3+C(W^3)}$	Improvement[6] over the baseline (631.09), %	λ^*
102	2509.28	405.13	2.21	0.19
202	1957.58	400.73	3.28	0.23
502	1406.09	393.66	4.98	0.27
1002	1061.08	388.55	6.22	0.30
2002	807.93	383.87	7.34	0.30
3002	696.43	382.20	7.75	0.30
4002	627.36	381.27	7.97	0.31
5002	585.09	380.90	8.06	0.31

[6] Improvement is calculated: $improvement = \frac{baseline - perplexity}{baseline} \cdot 100\%$.

[*] See formula (5) for details.

Table 5

Perplexities of class-based and interpolated LMs for $V_W = 1000$k

Number of classes, K_W	Class model perplexity, $PP_{C(W^3)}$	Interpolated perplexity, $PP_{W^3+C(W^3)}$	Improvement over the baseline (414.30), %	λ
102	4659.78	607.57	3.73	0.23
202	3702.17	596.86	5.42	0.26
502	2637.05	579.23	8.22	0.31
1002	1987.87	567.06	10.14	0.33
2002	1499.88	555.53	11.97	0.33
3002	1276.30	551.26	12.65	0.33
4002	1144.45	549.10	12.99	0.33
5002	1058.86	548.81	13.04	0.33

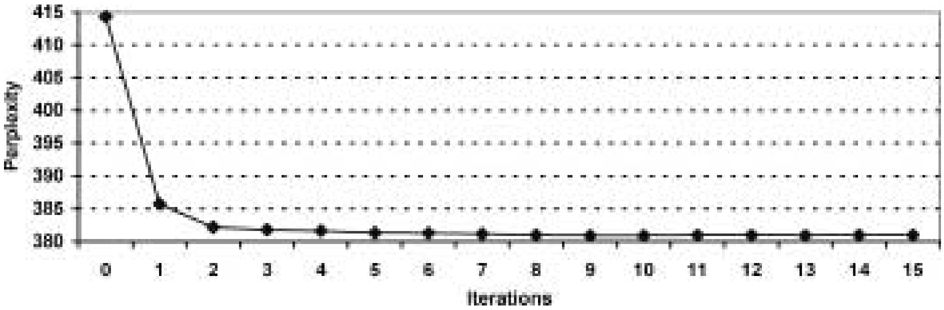

Fig. 1. Impact of number of iterations on the perplexity of interpolated model ($V_W = 65$k, $K_W = 3002$).

3.4. *Language Models Based on Morphological Word Decomposition*

3.4.1. *Word Decomposition*

Morphological analyzer was at the basis of our word decomposition procedure. Ideal morphological analyzer can be thought of as a procedure M which outputs a pair $\{s, g\}$ for each input word w, where s and g are the base form and the part-of-speech (POS) tag of the word w respectively.

$$M\colon\ w \to M(w) =< s, g > . \tag{10}$$

Unfortunately, many Lithuanian words are morphologically ambiguous, and morphological analyzer outputs a set of possible base form/POS tag decompositions $\{s'_i, g'_{ij}\}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$.

$$\begin{aligned} M'\colon\ \ & w \to M'(w) \\ & = \big\{< s'_1, g'_{11} >, \ldots, < s'_1, g'_{1n_1} >, < s'_2, g'_{21} >, \ldots, < s'_m, g'_{mn_m} >\big\}. \end{aligned} \tag{11}$$

As the ambiguity could not be solved without some context-aware morphological disambiguator we had to make some arbitrary choices. Base form s was assigned the first base form out of the list of base forms returned by morphological analyzer, i.e., $s = s'_1$. POS tag g(henceforth called generalized POS tag or GPOS tag) was assigned the set of all POS tags returned by morphological analyzer, i.e., $g = \{g'_{ij}\},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i$. In addition, certain non-inflected prepositions and conjunctions had $g = s$ as we thought such words may influence the inflections of the following words (see line 5 in Table 6). The words rejected by the morphological analyzer had $s = w$ and $g =$ unrecognized (see line 6 in Table 6).

3.4.2. *Pure Morphological Language Models*

We have investigated two language models $\widehat{P}_{S^3 \times G^3 S}$ and $\widehat{P}_{S^3 \times (GS + G^3)}$ based on morphological word decomposition. Model $\widehat{P}_{S^3 \times (GS + G^3)}$ can be thought of as the simplification of$\widehat{P}_{S^3 \times G^3 S}$, because of $\widehat{P}_{G^3 S}$ is replaced by the linear interpolation of two simpler

Table 6

Word decomposition examples

Word w	Base forms	POS tags	Decomposition M
vaikas (child)	s_1 =vaikas	g_{11} =<noun masc. sg. nom.>	<$s_1, \{g_{11}\} >$
vaikai (children, dissipate [you])	s_1 =vaikas s_2 =vaikyti	g_{11} =<noun masc. pl. nom.> g_{12} =<noun masc. pl. voc.> g_{21} =<verb ind. pres. t. sg. 2 pers.>	<$s_1, \{g_{11}, g_{12}, g_{21}\} >$
bėgau (ran [I])	s_1 =bėgti	g_{11} =<verb ind. past t. sg. 1 pers.>	<$s_1, \{g_{11}\} >$
bėgo (ran [he/they])	s_1 =bėgti	g_{11} =<verb ind. past t. sg. 3 pers.> g_{12} =<verb ind. past t. pl. 3 pers.>	<$s_1, \{g_{11}, g_{12}\} >$
po (after)	s_1 =po	g_{11} =<conj.>	$< s_1, \{s_1\} >$
Aminorūgštys (amino acids)			<w, {unrecognized}>

models:

$$\begin{aligned}
&\widehat{P}_{S^3\times G^3S}(w_i|w_{i-2}\ldots w_{i-1})\\
&\quad = \widehat{P}_{S^3\times G^3S}(s_ig_i|s_{i-2}g_{i-2}s_{i-1}g_{i-1})\\
&\quad \approx \widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\widehat{P}_{G^3S}(g_i|g_{i-2}g_{i-1}s_i), \qquad (12)\\
&\widehat{P}_{S^3\times(GS+G^3)}(w_i|w_{i-2}\ldots w_{i-1})\\
&\quad \approx \widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\big(\lambda_1\widehat{P}_{GS}(g_i|s_i)+(1-\lambda_1)\widehat{P}_{G^3}(g_i|g_{i-2}g_{i-1})\big), \qquad (13)
\end{aligned}$$

where

$\widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})$ is the probability estimate of seeing word base form s_i given the two preceding word base forms s_{i-1} and s_{i-2},

$P_{G^3}(g_i|g_{i-2}g_{i-1})$ is the probability estimate of seeing GPOS tag g_i given the two preceding GPOS tags g_{i-1} and g_{i-2},

$\widehat{P}_{G^3S}(g_i|g_{i-2}g_{i-1}s_i)$ is the probability estimate of seeing GPOS tag g_i given the word base form s_i and the two preceding GPOS tags g_{i-1} and g_{i-2},

$\widehat{P}_{GS}(g_i|s_i)$ is the probability estimate of seeing GPOS tag g_i given the word base form s_i.

Models$\widehat{P}_{S^3}, \widehat{P}_{G^3S}, \widehat{P}_{GS}$ and $\widehat{P}_{G^3}$ were smoothed using Witten-Bell smoothing technique.

Morphological word decomposition resulted in $V_S = 371251$ distinct base forms and $V_G = 3164$ distinct GPOS tags (Table 7).

Both morphological models decreased OOV rate from 1.5% to 1.02% at the expense of increased perplexity. Model $\widehat{P}_{S^3\times G^3S}$ achieved lower perplexity than model$\widehat{P}_{S^3\times(GS+G^3)}$. Within $\widehat{P}_{S^3\times(GS+G^3)}$ model, the 3-gram $\widehat{P}_{G^3}$ was practically ignored ($\lambda_1 = 0.99$). This means that the relative frequency of GPOS tag given word's base form is of much greater importance.

Table 7

Perplexities and OOV rates of pure morphological language models

Model	**Vocabulary size**		**Perplexity, PP**	**OOV %**
	V_S	V_G		
$\widehat{P}_{S^3\times G^3S}$	371251	3164	1336.77	1.02
$\widehat{P}_{S^3\times (GS+G^3)}$	371251	3164	1363.16	1.02
$\widehat{P}_{W^3}$ (baseline)	$V_W{}^7 = 1157911$		644.46	1.50

[7] Baseline LM had $V_W = 1157911$ instead of $V_W = 1422746$ words. Vocabulary reduction resulted from corpus clearing performed before morphology-based modeling: misspelled words, foreign words and words that were both rejected by morphological analyzer and found just once in the training corpus were removed from this new vocabulary.

3.4.3. *Class-based Morphological Language Models*

The model $\widehat{P}_{S^3\times G^3S}(12)$ was selected for further investigations as it gave lower perplexity than$\widehat{P}_{S^3\times (GS+G^3)}$. We investigated 5 ways of introducing class-based modeling within model$\widehat{P}_{S^3\times G^3S}$. We clustered word base forms and GPOS tags into classes and replaced components$\widehat{P}_{S^3}$, $\widehat{P}_{G^3S}$ of $\widehat{P}_{S^3\times G^3S}$ by class-based models:

$$\begin{aligned}
&\widehat{P}_{(S^3+C^3(S))\times G^3S}(w_i|w_{i-2}\dots w_{i-1}) \approx \big[\lambda_2\widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\\
&\quad + (1-\lambda_2)\widehat{P}_{SC(S)}(s_i|cs_i)\widehat{P}_{C(S^3)}(cs_i|cs_{i-2}cs_{i-1})\big]\cdot\widehat{P}_{G^3S}(g_i|g_{i-2}g_{i-1}s_i), &(14)\\
&\widehat{P}_{S^3\times C^3(G)S}(w_i|w_{i-2}\dots w_{i-1}) \approx \widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\widehat{P}_{GC(G)}(g_i|cg_i)\\
&\quad \times \widehat{P}_{C(G^3)S}(cg_i|cg_{i-2}cg_{i-1}s_i), &(15)\\
&\widehat{P}_{S^3\times (G^3S+C^3(G)S)}(w_i|w_{i-2}\dots w_{i-1}) \approx \widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\\
&\quad \times \big[\lambda_3\widehat{P}_{G^3S}(g_i|g_{i-2}g_{i-1}s_i) + (1-\lambda_3)\widehat{P}_{GC(G)}(g_i|cg_i)\\
&\quad \times \widehat{P}_{C(G^3)S}(cg_i|cg_{i-2}cg_{i-1}s_i)\big], &(16)\\
&\widehat{P}_{S^3\times (G^3S+G^3C(S))}(w_i|w_{i-2}\dots w_{i-1}) \approx \widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\\
&\quad \times \big[\lambda_4\widehat{P}_{G^3S}(g_i|g_{i-2}g_{i-1}s_i) + (1-\lambda_4)\widehat{P}_{G^3C(S)}(g_i|g_{i-2}g_{i-1}cs_i)\big], &(17)\\
&\widehat{P}_{(S^3+C^3(S))\times C^3(G)S}(w_i|w_{i-2}\dots w_{i-1}) \approx \big[\lambda_5\widehat{P}_{S^3}(s_i|s_{i-2}s_{i-1})\\
&\quad + (1-\lambda_5)\widehat{P}_{SC(S)}(s_i|cs_i)\widehat{P}_{C(S^3)}(cs_i|cs_{i-2}cs_{i-1})\big]\cdot\widehat{P}_{GC(G)}(g_i|cg_i)\\
&\quad \times \widehat{P}_{C(G^3)S}(cg_i|cg_{i-2}cg_{i-1}s_i), &(18)
\end{aligned}$$

where

cs_i is the class to which the base form s_i of the word w_i is assigned,

cg_i is the class of a GPOS tag g_i,

$\widehat{P}_{SC(S)}(s_i|cs_i)$ is a relative frequency of the base form s_i given its class cs_i,

$\widehat{P}_{GC(G)}(g_i|cg_i)$ is a relative frequency of the GPOS tag g_i given its class cg_i,

$\widehat{P}_{C(S^3)}(cs_i|cs_{i-2}cs_{i-1})$ is the probability estimate of seeing word base form class cs_i given the two preceding word base form classes cs_{i-1} and cs_{i-2},

$\widehat{P}_{C(G^3)S}(cg_i|cg_{i-2}cg_{i-1}s_i)$ is the probability estimate of seeing class of GPOS tags g_i given the word base form s_i and the two preceding classes of GPOS tags g_{i-1} and g_{i-2},

$\widehat{P}_{G^3C(S)}(g_i|g_{i-2}g_{i-1}cs_i)$ is the probability estimate of seeing GPOS tag g_i given the word base form class cs_i and the two preceding GPOS tags g_{i-1} and g_{i-2},

We clustered $V_S = 371251$ base forms into $K_S = 102, 502, 1002, 3002, 5002$ classes and $V_G = 3164$ GPOS tags into $K_G = 52, 102, 202, 502, 1002, 2002$ classes using the method described in Subsection 3.3. The results were as follows.

Models $\widehat{P}_{(S^3+C^3(S))\times G^3S}(14)$ and $\widehat{P}_{(S^3+C^3(S))\times C^3(G)S}(18)$ gave the improvement in perplexity as shown in Fig. 2.

The greatest improvement in perplexity $PP_{(S^3+C^3(S))\times G^3S} = 1236.73$ (or 7.48% with respect to $PP_{S^3\times G^3S} = 1336.77$) was obtained with $K_S = 3002$ classes of word baseforms. Complex model $\widehat{P}_{(S^3+C^3(S))\times C^3(G)S}$ was slightly worse than$\widehat{P}_{(S^3+C^3(S))\times G^3S}$. The lowest perplexity obtained with this model was $PP_{(S^3+C^3(S))\times C^3(G)S} = 1239.88$ for $K_G = 2002$ and $K_S = 3002$.

Models $\widehat{P}_{S^3\times C^3(G)S}(15)$ and $\widehat{P}_{S^3\times(G^3S+C^3(G)S)}(16)$ based on the use of GPOS tag classes did not give the improvement in perplexity as shown in Fig. 3.

The perplexity of model $\widehat{P}_{S^3\times C^3(G)S}$ reached the perplexity of a non-class-based model $\widehat{P}_{S^3\times G^3S}$ for $K_G = 2002$ GPOS tag classes.

Model $\widehat{P}_{S^3\times(G^3S+G^3C(S))}(17)$ resulted in no significant improvement in perplexity ($PP_{S^3\times(G^3S+G^3C(S))} > PP_{S^3\times G^3S}$). The component $\widehat{P}_{G^3C(S)}$ of $\widehat{P}_{S^3\times(G^3S+G^3C(S))}$ was practically ignored during the interpolation.

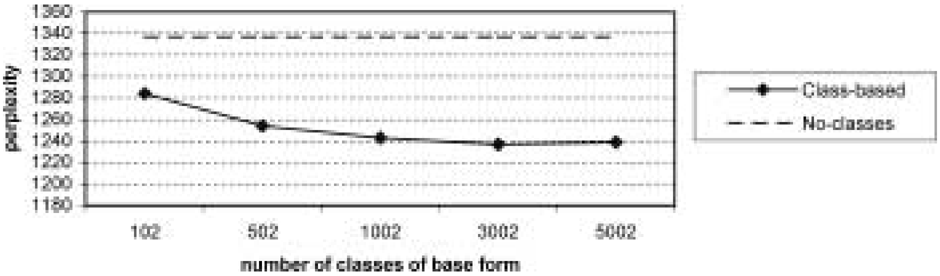

Fig. 2. Impact of number of classes of word base forms on the perplexity of interpolated model (14).

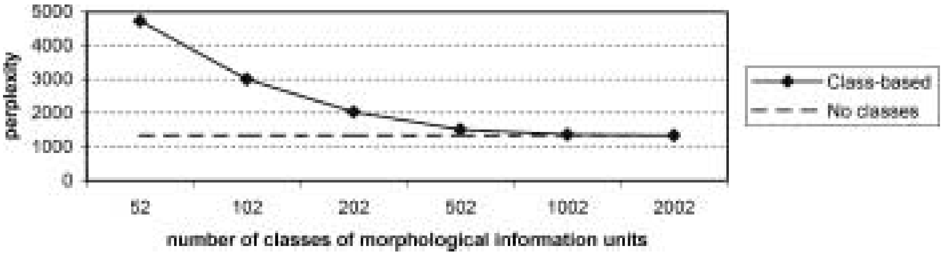

Fig. 3. Impact of number of classes of morphological information units on perplexity for model (15).

4. Discussion

Our work described in this paper affirmed the difficulties of modeling highly inflected languages, such as huge vocabulary size necessary to achieve moderate OOV word rates, model sparseness, and high model perplexity. We were not able to achieve 1% OOV rate even with $V_W = 1000$k Lithuanian vocabulary.

Our investigations confirmed that class-based language modeling is helpful in coping with data sparseness problem. Class-based models improved perplexity estimate by 13.04% and 8.06% for 1000k and 65k vocabularies respectively. We found that the optimum number of Lithuanian word classes is somewhere about 3000. Word 3-gram built over 1000k word vocabulary is sparser and has worse probability estimates than 3-gram built over 65k vocabulary. This explains why class based modeling gives more significant improvement in perplexity estimate for bigger vocabularies.

The time complexity of one clustering iteration of the algorithm described in 3.3 grew as fast as $\mathrm{O}(V \cdot K^2)$, where V was the number of distinct words and K was the number of classes. For instance, the average time required for one clustering iteration with $K = 102$, $K = 4002$ and $K = 5002$ classes was 0.75h, 150h and 220h respectively ($V = 1000$k; Pentium III 935Mhz operating Windows 2000). Fortunately, it appeared that clustering algorithm reached the local maximum and converged almost after 2 iterations. More sophisticated initialization methods or some less greedy clustering techniques may lead to perplexity improvements surpassing 13% given we can find a way of overcoming the limitations of the computational resources (parallel computing).

Language models based on word decomposition into its base form and its POS tag decreased vocabulary by 67,7% (from 1157911 to 371251+3164) at the expense of increased perplexity. We see three possible ways of improving perplexity estimates:

The first possibility is to decrease the vocabulary of base forms by cleaning text corpus (removing garbage, fixing typesetting errors, marking-up sentence boundaries, further extending the possibilities of the morphological analyzer).

The second possibility is to improve the morphological language model by solving the problem of morphological ambiguity on the basis of contextual information. It is clear that our present set of 3164 generalized POS tags is redundant and morphological disambiguation can reduce this set.

The third possibility is to use adaptation to text corpus, i.e., to complement language models described in this paper by adaptive language models such as cache-based models topic-specific language models, etc.

5. Conclusions

In this paper, we presented our ongoing research on statistical language modeling of Lithuanian. We investigated the idea of improving sparse n-gram models of highly inflected Lithuanian language by interpolating them with complex n-gram models based on word clustering and morphological word decomposition. Words, word base forms and part-of-speech tags were clustered into 50 to 5000 automatically generated classes. Complex 3-gram and 4-gram class-based language models were built and evaluated on Lithuanian text corpus, which contained 85 million words. Our investigations confirmed that class-based language modeling is helpful in coping with data sparseness problem. Class-based models improved perplexity estimate by 13.04% and 8.06% in comparison with the baseline 3-gram model for 1000k and 65k vocabularies respectively. We found that the optimum number of Lithuanian word classes is somewhere about 3000. Language models based on word decomposition into its base form and its POS tag decreased vocabulary by 67,7% and out-of-vocabulary word rate from 1,5% to 1.02% at the expense of increased perplexity.

References

Chen, S., and J.T. Goodman (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, **13**, 359–394.

Clarkson, P., and R. Rosenfeld (1997). Statistical language modeling using the CMU – Cambridge Toolkit. In *Proceedings of 5th European Conference on Speech Communication and Technology, Eurospeech'97*.

Filipovič, M. (2003). Isolated word recognition using neural networks. In *Proceedings of the Conference "Informacinės technologijos 2003"*, IX, KTU. pp. 10–20 (in Lithuanian).

Huang, X., A. Acero and H.W. Hon (2001). *Spoken Language Processing*. Prentice–Hall, New Jersey.

Ircing, P., P. Krbec, J. Hajic, J. Psutka, S. Khudanpur, F. Jelinek and W. Byrne (2001). On large vocabulary continuous speech recognition of highly inflectional language – Czech. In *Proceedings of 7th European Conference on Speech Communication and Technology, Eurospeech'2001*, B14. pp. 487–491.

Jelinek, F. (2001). *Statistical Methods for Speech Recognition*. Massachusetts Institute of Technology, Cambridge.

Jurafsky, D., and J.H. Martin (2000). *Speech and Language Processing*. Prentice–Hall, New Jersey.

Klakow, D. (1998). Log-linear interpolation of language models. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia.

Kobayashi, N., and T. Kobayashi (1999). Class-combined word N-gram for robust language modeling. In *Proceedings of 6th European Conference on Speech Communication and Technology, Eurospeech'99*. pp. 1599–1602.

Laurinčiukaitė, S. (2003). Isolated Lithuanian word recognition based on hidden Markov models. In *Proceedings of the Conference "Informacinės technologijos 2003"*, IX, KTU. pp. 21–24 (in Lithuanian).

Lipeika, A., J. Lipeikienė and L. Telksnys (2002). Development of isolated word speech recognition system. *Informatica*, **13**(1), 37–46.

Maltese, G., P. Bravetti, H. Crépy, B.J. Grainger, M. Herzog and F. Palou (2001). Combining word- and class-based language models: a comparative study in several languages using automatic and manual word-clustering techniques. In *Proceedings of 7th European Conference on Speech Communication and Technology, Eurospeech'2001*, A32. pp. 21–24.

Marcinkevičienė, R. (2000). Corpus linguistics in theory and practice. *Darbai ir Dienos*, VDU, **24**, 7–64 (in Lithuanian). `http://donelaitis.vdu.lt/`

Martins, C., J.P. Neto and L.B. Almeida (1999). Using partial morphological analysis in language modeling estimation for large vocabulary portuguese speech recognition. In *Proceedings of 6th European Conference on Speech Communication and Technology, Eurospeech'99*. pp. 1603–1606.

Raškinis, G., and D. Raškinienė (2003). Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system. *Informatica*, **14**(1), 75–84.

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here. In *Proceedings of the IEEE*, **88**(8).

Sepesy Maucec, M., and Z. Kacic (2001). Topic detection for language model adaptation of highly-inflected languages by using a fuzzy comparison function. In *Proceedings of 7th European Conference on Speech Communication and Technology, Eurospeech'2001*, A42. pp. 243–247.

Siivola, V., M. Kurimo and K. Lagus (2001). Large vocabulary statistical language modeling for continuous speech recognition in Finnish. In *Proceedings of 7th European Conference on Speech Communication and Technology, Eurospeech'2001,* B25. pp. 737–741.

Vaičiūnas, A., and G. Raškinis (2003). Statistical modeling of Lithuanian language. In *Proceedings of the Conference "Informacinės technologijos 2003"*, KTU, IX. pp. 35–40 (in Lithuanian).

Whittaker, E.W.D. (2000). *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*. PhD thesis. Cambridge University, Cambridge.

Zinkevičius, V. (2000). Lemuoklis – tool for morphological analysis. *Darbai ir Dienos,* VDU, **24**, 245–274 (in Lithuanian).

A. Vaičiūnas (born in 1976) received his MSc degree in computer science from the Vytautas Magnus University in Kaunas in 2000. Presently, he is a PhD student at the same university. His research interests are natural language modelling and speech recognition.

G. Raškinis (born in 1972) received his MSc degree in artificial intelligence and pattern recognition from the University of Pierre et Marie Curie in Paris in 1995. He received doctor's degree in the field of informatics (physical sciences) in 2000. Presently, he works at the Center of Computational Linguistics and teaches at the Department of Applied Informatics of VMU. His research interests include application of machine learning techniques to human language processing.

V. Kaminskas (born in 1946) graduated from Kaunas Polytechnic Institute in 1968. He received doctor's (1972) and habilitated doctor's (1983) degrees in the field of control systems and theory of information. He is an Expert-Member (1991) and a Corresponding Member (1998) of the Lithuanian Academy of Sciences. Prof. V. Kaminskas has published 4 books and over 200 scientific papers. Presently, he is a rector of Vytautas Magnus University. His research interests include computer aided simulation, identification, control and diagnostic systems.

Statistiniai lietuvių kalbos modeliai, pagrįsti žodžių klasterizacija ir žodžių morfologiniu išskaidymu

Airenas VAIČIŪNAS, Gailius RAŠKINIS, Vytautas KAMINSKAS

Šiame straipsnyje pateikti lietuvių kalbos statistinio modeliavimo tyrimai. Darbe išnagrinėti du būdai, kuriais įmanoma pagerinti smarkiai kaitomos Lietuvių kalbos n-gramos tipo statistinius modelius: kalbos žodžių grupavimas į klasterius ir morfologinis žodžių skaidymas į sudedamasias dalis. Tyrimo metu žodžiai, žodžių pagrindinės formos, ir žodžių kalbos dalies žymės buvo automatiškai grupuojamios į 50–5000 klasterių. Panaudojant 85 mln. žodžių apimties lietuvių kalbos tekstyną, buvo sukurti ir įvertinti keletas skirtingų 3-gramos ir 4-gramos tipo statistinių modelių, panaudojančių informaciją apie žodžių klasterius. Modeliai, panaudojantys žodžių klasterius tiesiškai interpoliuoti su įprastu 3-gramos tipo modeliu sumažino lietuvių kalbos modelio maišatį 13%. Morfologiniai modeliai sumažino neaprėpto žodyno dydį nuo 1,5% iki 1,02%.