The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
15-17 September 2011, Prague, Czech Republic

Morphological Based Language Models for Inflectional Languages

Tomáš Brychcín and Miloslav Konopík
Department of Computer Science and Engineering
University of West Bohemia in Pilsen, Univerzitní 8, 306 14 Pilsen, Czech Republic
E-mail: brychcin@kiv.zcu.cz, konopik@kiv.zcu.cz

***Abstract* – This paper shows a method to improve the language modeling for inflectional languages such as the Czech and Slovak language. Methods are based upon the principle of class-based language models, where word classes are derived from morphological information. Our experiments show that the linear interpolation with the class-based language models outperforms the stand-alone word N-gram language model about 10-30%.**

***Keywords* – class-based language models; inflectional languages; morphology; linear interpolation**

I. Introduction

Language modeling is a crucial task in many areas of NLP (Natural Language Processing). Speech recognition, optical character recognition and many other areas strongly depend on the performance of the language model that is being used. Every improvement in language modeling may also improve the particular job where the language model is used.

Similarly to other Slavic languages, Czech and Slovak are highly inflectional. Czech language has seven cases and three genders. Slovak language has six cases and also three genders. Many properties of both languages are very similar because of historical similarities and mutual interaction. Both languages have a relatively free word order (from the purely syntactic point of view): words in a sentence can usually be ordered in several ways which carry a slightly different meaning.

These properties of Czech and Slovak language complicate the language modeling task. High number of word forms and more sequences of words that are possible in the language lead to a higher number of n-grams. The data sparsity is a common problem of language models. In Czech, Slovak and other Slavic languages this problem is more evident.

We show in this article a way how to use the morphological knowledge of the language to decrease the problem of data sparsity.

II. State of the Art

Class-based modeling is among the most popular techniques for reducing huge vocabulary related sparseness of statistical language models. Individual words are clustered into a much smaller number of classes. As the result, less data are required to train a robust class-based language model. Both manual and automatic word clustering techniques are being used. Standalone class-based models usually perform poorly and that is the reason why they are usually combined with other models.

Many researchers demonstrated that linear interpolation of a standalone class-based language model and a standard word n-gram model reduced the model perplexity [6] or [10]. The log-linear interpolation or maximum entropy approach [5] for combining language models is also attractive as it combines features from different and sometimes disparate models into one model instead of combining models themselves.

The effective solution described in several works is to use information about the morphology of a language. In [7] the experiments with morphological random forests on Czech and Russian language are shown with conclusion that they can be used effectively for inflectional languages.

In [9] authors described the usage of morphology-based language models in a speech recognition system for conversational Arabic. Class based and single-stream factored language models using morphological word representations are applied into recognizer. Their results show word error rating improvement by 2%.

Speech recognition of Czech language using morphological class based language models was investigated in [8] however the tests were conducted only on small corpora. The work showed an improvement of recognition accuracy up to 2% by using language model based on log-linear interpolation of word model and morphological tag based model. We present experiments with more language models that were carried out on much bigger corpora.

III. Class-based N-gram Language Models

Class based language models belong among the state-of-the-art approaches for language modeling. The main task of the approach is to replace the statistical dependencies between words with dependencies between much lower number of word classes thus reducing the data sparsity problem.

This work was supported by grant no. SGS-2010-028.

Let W denote the set of possible words (word vocabulary) and C denote a class vocabulary. Then we can define a mapping function $m : W \rightarrow C$, which maps every word $w_i \in W$ to some $c_i \in C$. The mapping depends on the surrounding words because of the word ambiguity. So, the mapping function is defined as

$$c_i = m\left(w_i, w_a^b\right), \qquad a \leq i \leq b, \tag{1}$$

where the w_a^b means the word sequence from the position a to the position b and w_i is a single word on the position i.

The probability estimation of word w_i conditioned by its history w_1^{i-1} is given by the following formula

$$P\left(w_i|w_{i-n+1}^{i-1}\right) = P\left(w_i|c_i\right) \cdot P\left(c_i|c_{i-n+1}^{i-1}\right). \tag{2}$$

We are using the Modified Kneser-Ney interpolation (introduced in [1]) which is in present the state-of-the-art approach for smoothing methods. The main formula for smoothing is

$$P(w_i|w_{i-n+1}^{i-1}) = \cdots \\ \frac{c(w_{i-n+1}^{i}) - D(c(w_{i-n+1}^{i}))}{\sum_{w_i} c(w_{i-n+1}^{i})} + \gamma\left(w_{i-n+1}^{i-1}\right) P\left(w_i|w_{i-n+2}^{i-1}\right), \tag{3}$$

where $P()$ is the probability given by the Modified Kneser-Ney interpolation model and $c()$ is a count of n-gram. The smoothing of classes is analogical to this formula. The goal of discounting function $D(c)$ is to save some probability mass for lower-order models. The normalization function $\gamma(w_{i-n+1}^{i}) \in (0, 1)$ makes the probability distribution sum up to 1. Definitions and derivations of this functions should be founded in original paper.

The main advantage of Modified Kneser-Ney smoothing is different way to compute unigram probability distribution

$$P\left(w_i\right) = \frac{N_{1+}\left(\bullet w_i\right)}{N_{1+}\left(\bullet\bullet\right)}, \tag{4}$$

where symbol $\bullet$ means an arbitrary word (class) and $N_r(w_{i-n+1}^{i})$ is number of N-grams with frequency r. In different words, the unigram probability of w_i is given by the number of different bigrams ended with w_i divided by total number of different bigrams.

IV. Combining Language Models

We are using a simple but a very effective linear interpolation for combining different language models

$$P^{LI}\left(w_i|w_{i-n+1}^{i-1}\right) = \sum_{k=1}^{K} \lambda_k \cdot P_k\left(w_i|w_{i-n+1}^{i-1}\right), \tag{5}$$

where λ_k is the weight of the k-th language model $P_k()$. We are using *Expectation Maximization (EM)* algorithm described in [2] to compute optimal weights λ_k by the way of maximization of probability on held-out data.

V. Text Corpora

Language models in our experiments were trained on four morphologically annotated corpora in Czech and Slovak language. In each language there is one manually annotated corpus and one much larger corpus that is annotated automatically.

- **CZ-man**: The data in this corpus consists of manually annotated articles from several newspapers and journals in Czech language [13].
- **CZ-auto**: Contains news in many topics such as political, business, sports, international and other news gathered from one year in Czech language. Data in this corpus are provided by Czech News Agency (CNA) and they are annotated automatically.
- **SK-man**: Manually annotated part of Slovak National Corpus mainly consist of artistic, publicistic and professional oriented texts [11].
- **SK-auto**: Huge number of automatically annotated texts from Slovak National Corpus oriented also in artistic, publicistic and professional area [12].

Data from all corpora were divided into the training, testing and held-out set in proportions of 70%, 10% and 20%. Parameters of these corpora are shown in table I.

Table I. Corpora Used for Experiments.

	Corpora parameters			
	CZ-man	**CZ-auto**	**SK-man**	**SK-auto**
words in corpus	2M	36.2M	1.2M	76.8M
word vocabulary	168.4k	577.2k	129.3k	1M
lemma vocabulary	71.7k	288.6k	52k	515.6k
tag vocabulary	1.6k	1.7k	1.7k	1.3k

The lemma and morphological tag is assigned to each word in the corpora.

Morphological tags in Czech corpora are strings composed from 15 characters [4]. Every position encodes one morphological category such as part-of-speech, gender, number, case, tense and so on.

Every tag in Slovak corpus is composed of two parts [3]. First one defines morphological and grammatical properties of a word form. This string begins with a character encoding part-of-speech, followed by characters for other categories. The second (facultative) part specifies token as a part of specific word classes (proper names, defective forms).

Lemma and tag together should uniquely identify the word form.

VI. Experimental Results

In this section we try to show results of morphological based language models trained on both manually and automatically annotated corpora. There is also a comparison of modeling of both Czech and Slovak language.

We have tested the performance of several models where the rank (N) of all stand-alone N-gram language models was 3.

- **W**: simple word N-gram language model.
- **L**: stand-alone class based model, where words are mapped into classes by their lemma. Relation is M:N.
- **T**: stand-alone class based model with grouping by morphological tags. The relation is also M:N.
- **WL**: linear interpolation of the word based and lemma based model.
- **WT**: linear interpolation of the word based and tag based model.
- **WLT**: linear interpolation of the word based, lemma based and tag based model.

A. Perplexity Test

The perplexity results are shown in table II.

Table II. LANGUAGE MODELS PERPLEXITIES ON MANUALLY AND AUTOMATICALLY ANNOTATED CORPORA IN CZECH AND SLOVAK LANGUAGE.

	Perplexity					
	W	**L**	**T**	**WL**	**WT**	**WLT**
CZ-man	1400	2455	2308	1266	1038	995
SK-man	1354	2303	1847	1204	915	870
CZ-auto	274	461	1661	253	250	238
SK-auto	305	602	1777	288	279	270

As it can be seen, stand-alone class based models (L, T) give worse results than the word model (W) while their linear interpolations with word model (WL, WT, WLT) performs very well. It is that on automatically annotated corpora the perplexities are much lower because of much more training data available. On manually annotated corpora the reduction of perplexity is more significant (reduction about 30%) than on automatically annotated corpora (reduction slightly over 10%).

B. Word Estimation Test

The perplexity sometimes does not correspond with results from real-word applications and so we are introducing another test of our language models.

During the test our models try to estimate the missing word in the sentence. Since the selection of a correct word from the full vocabulary is computationally very expensive, we are using the list of most frequent words from which the language model tries to find the correct word. Of course the list is constructed in the way that it always contains the correct word (the word that was originally on the current place in the sentence).

During the m-th test, let O_m denote the order of the correct answer in the list of possibilities which is sorted according to the language model. The accuracy $ACC\%$ of word estimation is than defined as follows:

$$ACC\% = \frac{1}{M} \sum_{m:O_m=1} 1, \qquad 1 \leq m \leq M, \qquad (6)$$

where M is the total number of estimations. Results are shown in tables III and IV, where L denotes the size of list of possibilities.

Table III. ACCURACY OF WORD ESTIMATION TEST ON MANUALLY ANNOTATED CORPORA *CZ-man* AND *SK-man*.

CZ-man						
	ACC%					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	46.04	46.96	39.19	47.35	47.92	48.81
100	42.91	42.51	29.40	44.08	43.80	44.67
1000	41.16	39.99	28.17	42.25	42.26	43.06
10000	40.37	39.24	28.06	41.49	41.85	42.67
SK-man						
	ACC%					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	44.34	45.34	38.37	45.98	46.16	47.25
100	41.99	41.12	31.02	43.57	43.02	44.16
1000	40.59	40.22	29.48	41.93	41.74	42.79
10000	39.56	39.32	29.36	40.98	41.12	42.10

Table IV. ACCURACY OF WORD ESTIMATION TEST ON AUTOMATICALLY ANNOTATED CORPORA *CZ-auto* AND *SK-auto*.

CZ-auto						
	ACC%					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	72.97	69.19	41.88	73.61	72.50	73.79
100	67.30	61.17	32.41	67.76	67.34	67.88
1000	63.87	56.61	31.22	64.29	64.11	64.33
SK-auto						
	ACC%					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	73.66	67.67	41.96	74.00	72.93	74.11
100	68.24	59.48	32.29	68.48	68.68	68.96
1000	64.50	54.52	30.34	64.79	64.93	65.14

Linear interpolation of word model and morphological based models gives the best results on all corpora. The word model have a lower accuracy and the stand-alone class based models performs much worse. We can see, that improvement of accuracy on manually annotated corpora is more evident. This is caused by the fact, that stand-alone word model is poorly trained (too small peace of manually annotated data), that is evident from perplexity results, so the morphological knowledge is more needed. And, of course, mistakes are generated during the process of automatic tagging of morphology causing a degradation of morphological based language models.

The expectation value of order $E(O)$ is also a very important metric. Together with $ACC\%$ it helps us to make a reasonable idea about the probability distribution of the language model. It is defined as

$$E\left(O\right) = \frac{1}{M} \sum_{m=1}^{M} O_m, \qquad 1 \leq m \leq M. \qquad (7)$$

The expectation values of order in lists of different sizes are shown in tables V and VI.

Table V. THE EXPECTATION VALUES OF ORDER ON MANUALLY ANNOTATED CORPORA *CZ-man* AND *SK-man*.

CZ-man						
	$E(O)$					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	4.67	4.39	3.88	4.52	4.14	4.07
100	29.63	26.35	20.42	27.34	21.97	21.32
1000	200.18	178.15	115.45	179.60	120.96	120.96
10000	628.39	853.48	442.50	566.46	340.67	326.18
SK-man						
	$E(O)$					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	5.00	4.52	4.20	4.78	4.35	4.25
100	31.13	26.98	19.81	28.67	20.74	20.10
1000	216.38	184.83	103.78	191.60	105.24	100.72
10000	618.40	910.86	429.95	545.60	305.94	286.06

Table VI. THE EXPECTATION VALUES OF ORDER ON AUTOMATICALLY ANNOTATED CORPORA *CZ-auto* AND *SK-auto*.

CZ-auto						
	$E(O)$					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	2.39	2.42	3.38	2.31	2.37	2.31
100	11.26	10.91	17.81	10.26	10.48	8.83
1000	73.50	66.65	99.70	63.19	61.27	55.81
SK-auto						
	$E(O)$					
L	**W**	**L**	**T**	**WL**	**WT**	**WLT**
10	2.28	2.44	3.47	2.24	2.28	2.25
100	9.81	10.52	16.64	9.27	9.13	8.28
1000	64.68	63.14	88.35	57.31	52.49	48.95

We can see (in tables) that the linear interpolation of models produces the lowest (the best) order of a correct word on all corpora. The average order given by the stand-alone word model is one of the worst. The improvement of expectation value of order against the baseline word model is significantly better on manually annotated corpora as in accuracy results. It is also interesting that on manually annotated corpora the tag based language model estimates the correct answer in one of the lowest orders from all models but on automatic annotated corpora is by far the worst. This is again evidently caused by the fact that in the automatic tagging there can be mistakes.

The linear interpolation of all stand-alone models decreases the expectation value of order $E(O)$ of the correct answer about 10-50% and increases the accuracy $ACC\%$ about 0.5-2.5%.

The word estimation test verified the results of perplexity test. Both test showed that the linear interpolation of word based language model, lemma based model and tag based model can significantly improve the performance of a stand-alone word N-gram language model.

VII. CONCLUSION

In this article we have experimented with the technique for building language models based on morphological information. We have focused on working with manually annotated and much larger automatically annotated corpora.

We have created two class based language models, one of them used mapping of words to their lemmas and the second one to their morphological categories. During our experiments we have shown that the linear interpolation of the word based N-gram language model and both morphological class based models is very effective for Czech and Slovak language. These languages are representatives of inflective languages. As we have expected, the modeling of Czech and Slovak language is very similar.

The results presented in this article clearly show a very significant reduction in the perplexity and increase in the accuracy. We can state that the morphological class based language models are suitable for building large corpora language models.

ACKNOWLEDGEMENT

The access to the MetaCentrum computing facilities provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic is appreciated. We also appreciate Czech News Agency (CNA) for providing huge number of texts in Czech language.

REFERENCES

[1] F.S. Chen and J.T. Goodman: An Empirical Study of Smoothing Techniques for Language Modeling, *Technical Report TR-10-98B*, Computer Science Group, Harvard University, 1998.

[2] A.P. Dempster and N.M. Laird and D.B. Rubin: Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, SERIES B*, vol. 39, no. 1, New York, p. 1-38, 1977.

[3] L. Giantisová: Morphological Analysis of the Slovak National Corpus, *Insight into Slovak and Czech Corpus Linguistics*, Bratislava, p. 166 – 178, 2005.

[4] J. Hajic: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Prague, 2004.

[5] F. Jelinek: Statistical Methods for Speech Recognition. Massachusetts Institute of Technology, Cambridge, 2001.

[6] G. Maltese and P. Bravetti and H. Crepy and B.J. Grainger and M. Herzog and F. Palou: Combining word and class-based language models: a comparative study in several languages using automatic and manual wordclustering techniques. *In Proceedings of 7th European Conference on Speech Communication and Technology*, Eurospeech 2001, A32. p. 21-24, 2001.

[7] I. Oparin: Language Models for Automatic Speech Recognition of Inflectional Languages. *PhD thesis*, University of West Bohemia, Pilsen, 2008.

[8] A. Pražák and P. Ircing and L. Müller: Language Model Adaptation Using Different Class-Based Models. *SPECOM 2007 Proceedings*, p. 449-454, Moscow State Linguistic University, Moscow, 2007.

[9] D. Vergyri and K. Kirchhoff and K. Duh and A. Stolcke: Morphology-based language modeling for Arabic speech recognition. *In Proc. ICSLP 2004*, Jeju Island, Korea, Oct. 2004.

[10] E.W.D. Whittaker: Statistical Language Modelling for Automatic Speech Recognition of Russian and English. *PhD thesis*, Cambridge University, Cambridge, 2000.

[11] Slovak National Corpus – r-mak-3.0. Bratislava: Ludovit Stur Institute of Linguistics Slovak Academy of Sciences, 2009. http://korpus.juls.savba.sk

[12] Slovak National Corpus – prim-5.0-public-inf. Bratislava: Ludovit Stur Institute of Linguistics Slovak Academy of Sciences, 2011. http://korpus.juls.savba.sk

[13] The Prague Dependency Treebank 2.0: Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Prague, 2006. http://ufal.mff.cuni.cz/pdt2.0/