Language model adaptation for highly-inflected Slovenian language in comparison to English language

Mirjam Sepesy Maučec and Zdravko Kačič
University of Maribor, Faculty of Electrical Engineering and Computer Science
Smetanova 17, SI-2000 Maribor, Slovenia
mirjam.sepesy@uni-mb.si

Abstract

The language model in this article is meant as an information source of a speech recogniser. The environment to which the recogniser will be put is topic-specific. The idea is to try to adapt the model of general language to the target domain of discourse. We concentrate on feature extraction process devoted to highly-inflected languages. Results of experiments on English and Slovenian corpora are reported.

1. Introduction

A language model aims to provide a representation of language. We concentrate on statistical aspects of modelling and not on grammatical ones. Statistical models rely on the assumption that the future use of a language will follow similar linguistic patterns to those used in the past.

Few years ago the problem of statistical language model was defined as the sparse data problem. Almost all language model research has adopted "bigger is better" approach where enormous volumes of training text are analysed in order to derive more reliable statistics. However, improvements with size did not yield much better language models for speech recognition. Training corpus should be used in more advanced way.

2. Basic language modelling

The task of a language model is to assign the probability $P(W)$ to every conceivable word string W. N-gram language models are most widely used (Jelinek, 1998). An N-gram is a model that uses the last N-1 words of the history as its sole information source. Although they are very simple, the experiments have shown that they are surprisingly difficult to improve on. We use trigram models which restrict the history to two immediately preceding words

$$P(W) = \prod_{i=1}^{n} P(w_i | w_1 ... w_{i-1}) \approx \prod_{i=1}^{n} P(w_i | w_{i-2} w_{i-1}) \qquad (1)$$

We should point out that the word refers to a word form defined by its spelling. Two differently spelled inflections or derivations of the same stem are considered different words. This fact don't lead to a problem in modelling English language but in modelling highly-inflected languages. Great number of different word forms derived from one lemma cause the enormous vocabulary growth. This problem can be solved by choosing another basic unit instead of word (for example, morpheme) (Byrne et al., 2000). In this article we were not concerned with the basic language modelling. Deriving trigram and bigram probabilities is always a sparse estimation problem, probability smoothing was performed by Katz backing-off algorithm (Katz, 1987).

3. Topic adaptation

N-gram techniques seem to capture well short-term dependencies. They lack any ability to exploit the linguistic nuances between domains. The environment to which the recogniser will be put is topic-specific. If we could effectively identify the domain of discourse, a model appropriate for the current domain could be used. We do not assume the target domain to be equivalent to one predefined topic. Target domain can be seen as a combination of several elemental topics.

The goal of the adaptation is to lower the language model perplexity by providing a higher probability of words and word-sequences, which are characteristic of the domain of discourse. Adopted models were built on three semantic levels:

- general language model (G). It was built by using all available training text.
- topic model (T). It was built by using only the text of a predefined topic most similar to the target domain.
- general topic model (T_{10}). It was built by using the text of 10 predefined topics most similar to the target domain.

We built two interpolated models:

- combined model (C)

$$P_C(w) = \lambda P(w) + (1-\lambda)P_G(w)\,. \tag{2}$$

- novel model (N)

$$P_N(w) = \lambda_1 P_T(w) + \lambda_2 P_{T_{10}}(w) + \lambda_3 P_G(w) \tag{3}$$

Figure 1 shows the adaptation scheme.

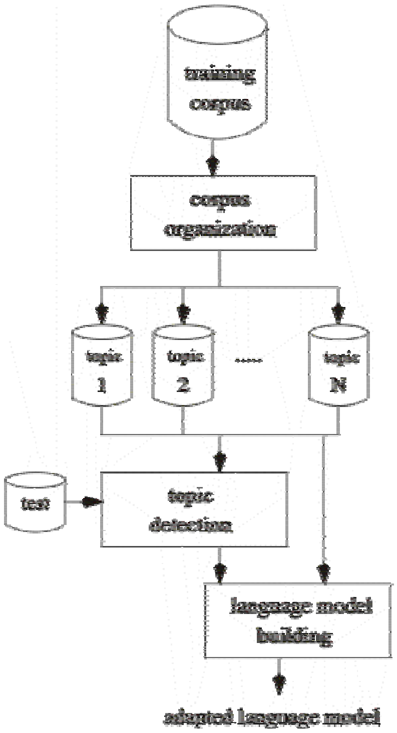

Figure 1: The adaptation scheme.

The adaptation scheme consists of the following three steps:

- *corpus organisation*. With the growing availability of textual data in electronic form large topically diverse corpora are constructed. Stories that share similar topics are gathered together into a set of clusters.
- *topic classification*. A classifier is used to find the clusters that are most similar in topic to the sample from the target environment.
- *language model building*. Language models at different semantic levels are built. The models are interpolated at the word level.

The first step is corpus organisation. We need the corpus organisation, which enable us to extract target-topic-similar parts of the whole collection and treat them as more representative. Given a corpus with keywords assigned to each story, topic clusters are simply created by defining each keyword as a label for a cluster. Unfortunately for minority languages such corpora are often not available. Automatic generation of clusters of documents need to be used. The success of automatic clustering is conditioned with the quality of document and topic cluster representation. Before introducing the right representation we will describe main characteristics of Slovenian.

4. Inflectional morphology of Slovenian

Slovenian is a South Slavic language with a speech area, wedging into the Croatian, Italian, German, and Hungarian linguistic territories. The morphological complexity of Slovenian in comparison to English will be described. It is Slovenian inflectional morphology which formally distinguishes both languages.

Using a highly simplified notion, the word formation in Slovenian does not differ much from those languages where new word forms are created using a stem with the addition of derivational suffixes.

Slovenian morphology introduces three main concepts: word classes, inflection, and grammatical categories. The main feature of word classes is their division into

- *inflectional classes*: substantive words, adjective words, and verb, and
- *non-inflectional classes*: adverb, predicate, preposition, conjunction, copula, and interjection.

The basic grammatical categories of Slovenian are: gender, number, case, degree, person, tense, mood and aspect. Slovenian shares its grammatical categories with other Slavic languages, except the category of dual in addition to singular and plural, and the Slovenian nominal system does not possess a category to express an appeal. A match of the grammatical categories and word classes results in different inflectional patterns. We will describe each grammatical category with examples.

4.1 The category of gender

Slovenian distinguishes three genders: the masculine, the feminine and the neuter. In the majority of the Slavic languages, gender is inherent in substantives, inflected in adjectives, and not expressed in pronouns. Slovenian has extended gender to personal pronouns, and marginally to verbal inflections. Table 1 shows some examples.

	Masculine		*Feminine*		*Neuter*	
	Slovenian	English	Slovenian	English	Slovenian	English
Noun	brat	brother	sestra	sister	dete	baby
Adjective	lep	pretty	lep-a	pretty	lep-o	pretty
Verb	dela-l	worked	dela-la	worked	dela-lo	worked
Pronoun	moj	my	moj-a	my	moj-e	my

Table 1: The examples of the application of gender.

4.2 The category of number

Slovenian morphology uses besides the singular and plural also the dual, when referring to two persons or objects. The category of number is applied to nouns, adjectives and pronouns. Table 2 shows some examples.

Singular		*Dual*		*Plural*	
Slovenian	English	Slovenian	English	Slovenian	English
en-a miz-a	one table	dv-e mizi	two tables	tr-i miz-e	three tables
en-o mest-o	a town	dv-e mest-i	two towns	tr-i mesta	three towns
lep-a	pretty	lep-i	pretty	lep-e	pretty
on	he	onadva	they two	oni	they

Table 2: The examples of the applications of number.

4.3 The category of case

The most striking differences between Slovenian and English morphology is the use of six cases in Slovenian, which denotes the relationship of individual words in sentence. In Slovenian we use different patterns for the following word classes: nouns, adjectives, and pronouns. The sentences in Table 3 illustrate the use of cases of the word *mesto* (Eng. "the town") and shows the main differences between Slovenian and English.

Case	Slovenian	English
Nominative	To je mest-o.	This is a town.
Genitive	Ne vidim nobenega mest-a.	I can't see any town.
Dative	Približujem se mest-u.	I'm walking towards this town.
Accusative	Kako bi opisal to mest-o?	How would you describe this town?
Locative	Kdo živi v tem mest-u?	Who lives in this town?
Instrumental	Pod tem mest-om teče reka.	There is a river beneath the town.

Table 3: The examples of the use of cases.

4.4 The category of degree

The gradation of adjectives and adverbs in Slovenian is quite similar to that in English. As in English, there are three degrees of comparison in Slovenian. Table 4 shows some examples.

Positive		*Comparative*		*Superlative*	
Slovenian	English	Slovenian	English	Slovenian	English
bel	white	bolj bel	more white	najbolj bel	most white
star	old	star-ejši	older	naj-star-ejši	the oldest

Table 4: The examples of the use of gradation.

4.5 The category of person

Verbal forms are related to the three types of the category of person. Table 5 shows the example of the conjugation of the verb "to work".

	Singular		Dual	Plural	
	Slovenian	English	Slovenian	Slovenian	English
1st person	dela-m	I work	dela-va	dela-mo	we work
2nd person	dela-š	you work	dela-ta	dela-te	you work
3rd person	Dela	he work	dela-ta	dela-jo	they work

Table 5: The examples of the applications of category of person.

4.6 The category of tense

There are four tenses in Slovenian language. Table 6 shows the use of them.

Tense	Slovenian	English
Present	del-am	I work
Past	dela-l sem	I worked
Future	dela-l bom	I shall work
Plusperfect	dela-l sem bil	I had worked

Table 6: The examples of the use of tenses.

4.7 The category of mood

There are three moods in Slovenian: indicative, imperative and conditional. Table 7 shows their use.

	Slovenian	English
Indicative	dala-m	I work
Imperative	del-aj	work
Conditional	dela-l bi	I would work

Table 7: The examples of the use of mood.

4.8 The category of aspect

Every verb obligatorily belongs to one of two classes of aspect: perfective or imperfective. The contrast between them is expressed not only by different suffixes, but also by a radical alternation of the stem. Table 8 gives some examples.

Perfective		*Imperfective*	
Slovenian	English	Slovenian	English
dvig-n-iti	to lift	dvig-a-ti	to be lifting
se-č-i	to reach	se-ga-ti	to be reaching
pri-ti	to come	pri-h-ajati	to be coming

Table 8: The examples of the use of aspect.

4.9 Morphemic alternations

There is one additional feature of Slovenian. Besides the extensive set of suffixes, words are also subject to a process of alternation. Two types of alternation are relevant to the written form of Slovenian: vocalic and consonantal. See Table 9.

Alternation in Slovenian		Meaning in English
vet-e-r	vetra	wind
bolez-e-n	bolezni	illnes
jo-k-ati	jo-č-em	to cry
zgu-b-iti	zgu-blj-en	lost

Table 9: The examples of morphemic alternation.

On the basis of the above description of the morphological structure of Slovenian in comparison to English, two main points can be emphasized:

- Slovenian displays features of the extremely rich inflectional morphology;
- Slovenian is characterized by various types of morphemic alternations in both stems and suffixes during inflection.

5. Document and topic cluster representation

Given a corpus with keywords assigned to each story, topic clusters are simply created by defining each keyword as a label for a cluster. Unfortunately for Slovenian such a corpora is not yet available. Automated generation of clusters of documents based on some similarity measure need to be used.

First we have to find the suitable representation. Documents and clusters are represented as a set of features. In most applications words are used as features. It has been argued that maximum performance is often not achieved by using all available features, but using a good subset of those only. Having features which do not help to discriminate between topics add noise. We want to show that it makes sense to group features into clusters, at least for languages with rich morphology. We want to group all words with the same meaning, but different grammatical form, in one cluster and represent them as one feature. We propose a novel approach for feature extraction based on soft comparison of words.

To avoid the use of an additional knowledge source like lexicon we define a set of membership functions. Each cluster defines its own membership function. The membership function associates to each word from the vocabulary a number representing the grade of membership of this word in that cluster. Membership function $\mu_{\tilde{c}}$ of cluster c is defined as

$$\tilde{c} = \{(w, \mu_{\tilde{c}}) \mid w \in V\}. \tag{4}$$

$\tilde{c}$ denotes a fuzzy set of cluster c. Cluster membership functions are based on fuzzy comparison function. Each word defines its own fuzzy set

$$\tilde{w} = \{(w, \mu_{\tilde{w}}) \mid w \in V\}. \tag{5}$$

The function sees the word as a sequence of characters. It returns value 1 if compared words are the same and 0 for extremely different words. In other cases it returns the value between 0 and 1. The comparison function is created by using fuzzy rules, which provide a natural way of dealing with partial matching. We define three sets of rules:

- language independent rules,
- rules describing English language and
- rules describing Slovenian language.

The rules are expressed as fuzzy implications, which use linguistic variables to express the grade of similarity (for example: not very similar, quite similar).

To get the impression of language independent rules we present two examples. a denotes the word with n characters and b denotes the word with m characters. The fuzzy implication

$$\textit{characters of words are different} \Rightarrow \textit{words are not very similar} \tag{6}$$

is transformed into the predicate

$$p_1(i,a,b) = \begin{cases} 0 & \exists\, j{:}\ a(i) = b(j) \\ 1 & otherwise \end{cases} \tag{7}$$

$$e_1(a,b) = \sum_{i=1}^{n} \frac{p_1(i,a,b)}{n+m} + \sum_{i=1}^{m} \frac{p_1(i,b,a)}{n+m}. \tag{8}$$

The fuzzy implication

$$\textit{two character sequences of words are the same} \Rightarrow \textit{words are quite similar} \tag{9}$$

is transformed into the predicate

$$p_2(i,a,b) = \begin{cases} 1 & \exists\, j{:}\ a(i) = b(j) \wedge a(i+1) = b(j+1) \\ 0 & otherwise \end{cases} \tag{10}$$

$$e_2(a,b) = \sum_{i=1}^{n} \frac{p_2(i,a,b)}{n+m-2} + \sum_{i=1}^{m} \frac{p_2(i,b,a)}{n+m-2}. \tag{11}$$

The predicates are scaled by linguistic variables. Their values are empirically chosen. The final value of comparison function is computed using scaling

$$\mu_{\tilde{a}}(b) = \frac{\max(e_{Similar}(a,b))}{\max(e_{Similar}(a,b)) + \max(e_{NotSimilar}(a,b))}. \tag{12}$$

$e_{Similar}$ denotes the set of predicates which describe similarity and $e_{NotSimilar}$ denotes the set of predicate which denote the distinction.

Language dependent rules for English are taken from the suffix stripping set of rules provided by (Porter, 1980). Language dependent rules for Slovenian describe in simplified form the inflectional morphology, described in Section 4 (Popovič, 1991).

The membership function of word w_i in cluster c_j is computed by using a modified single link agglomerative clustering (Voorhees, 1986). Similarity values of word pairs can be represented as a weighted, undirected graph where nodes represent words and weights represent the similarity of words connected by the edge. To save space, we keep only edges with weights greater then a prespecified threshold. The result of the single link hierarchy are locally coherent clusters. To avoid a chaining effect and consequently elongated clusters, we modify the merging criterion. A word is added to the cluster if its average similarity with all words in the cluster is the largest among all the words not yet clustered. Clusters are made one at the time. We start building a new cluster as soon as the largest similarity value does not exceed a prespecified threshold. Each cluster defines one feature. The number of clusters represents a feature vector length.

6. Topic detection

Once we have training documents, topic clusters and test samples represented as feature vectors, we use topic detection to determine the similarity between two feature vectors. Topic detection is performed by the use of TFIDF classifier (Joachims, 1996). It is used to determine the similarity between two documents or clusters.

7. Experiments

In our experiments we were using the broadcast news corpus (1996 CSR Hub-4 Language Model) for English and newspaper news corpus (Večer) for Slovenian due to their semantic richness.

The English broadcast news corpus contains 100 mio words. It was organised into topic-specific clusters of documents based on manually-assigned keywords. We were experimenting with topic clusters that have at least 300 articles. 244 clusters satisfy this constrain. Language model adaptation was performed on 20 randomly chosen topics. 80% of each topic cluster text was used for language model training, 10% of text for interpolation parameter estimation and 10% of text was used as test sample.

All words from the corpus were used for feature extraction. Before word clustering was performed, words from stop word list were removed. Using language independent word clustering feature vector size was reduced from 170,000 to 36,000. A sample of clusters is shown in Table 10. It shows that also misspelled words are correctly clustered.

aadmirable admirable admirably admira admirally admire admired admirer admires admir admirers
bbecause becau becaue
chinasports sports sport sporto sporty sported spotrer sportin sporting sportscar sportsman sportsmen sportcoat sportless
cilton clnton cinton
conferenced conferences conferencing teleconference teleconferenced teleconferences videoconferences

Table 10: Sample of English clusters.

For each test sample, we want to model, all topic clusters were ranked by the similarity value. If we used language dependent rules in feature extraction process the top 10 topics didn't change. Four types of language models were built: general language model (G), topic model (T), combined model (C) and novel model (N). All language models were trigram models with the vocabulary of 64,000 most frequent words. Results of 5 topics are shown in Table 11. Averaging over all 20 experimental topics the perplexity of adopted language model was reduced by 15%.

Topic	PP_G	PP_T	PP_C	PP_N
Automobiles	58	247	55	53
Middle East	55	145	27	27
Clinton, Bill	48	90	46	41
Holidays	62	331	49	48
Simpson, O. J.	45	59	25	23

Table 11: Test set perplexities for English language.

Unfortunately, the Slovenian newspaper corpus of 20 mio words is not yet annotated with keywords. Clustering was done automatically. Documents were merged into 100 clusters iteratively by the use of agglomerative clustering(Voorhees, 1986), TFIDF classifier and feature vectors built in feature extraction process. Using language independent word clustering the feature vector size was reduced from 200,000 to 21,000. A sample of clusters in shown in Table 12. Words in italic are from semantic point of view not correct clustered. Adding language dependent rules feature vector size was reduced to 18,000.

afer afera aferah afere aferi afero aferami *fer*
bančna bančne bančnem bančni bančno bančnih bančnik bančnim bančnikom bančnikov bančniški
cestnemu mestnemu mestnem mestne mestnega mestna mestni mestno
nihanje nihanj nihanja nihanju *ihan*
dobojevati izbojevati izbojeval

Table 12: Sample of Slovenian clusters.

Three test samples were manually created. All of them consist of 5 documents similar in topic. Language models were built in the same way as in the previous experiment with English corpus. There was only one difference. All words from test samples were added to the vocabulary to avoid the problem of out-of-vocabulary words. By using language independent feature extraction we have got 14% perplexity reduction. By adding language dependent rules the perplexity was reduced up to 30%. Results are given in Table 13.

Language independent feature extraction				
Topic	PP_G	PP_T	PP_C	PP_N
Sport	200	621	199	197
Weather forecast	154	201	153	141
Politics	220	598	210	215

Language dependent feature extraction			
Topic	PP_T	PP_C	PP_N
Sport	455	183	171
Weather forecast	168	150	115
Politics	598	201	189

Table 13: Test set perplexities for Slovenian language.

8. Conclusion

In our experiments we have shown that topic adaptation does result in a decrease in perplexity. To train a language model it does not make sense to use only a small portion of topic specialised text.

The results have shown that word clustering delivers significant topic detection improvement for highly-inflected languages and almost no improvement for English language.

The main drawbacks of experiments on Slovenian corpus were corpus size and absence of keyword labels.

9. References

Jelinek F 1998 *Statistical methods for speech recognition* .Cambridge, MIT Press.

Byrne W, Hajič J, Ircing P, Krbec P, Psutka J 2000 Morpheme based language models for speech recognition of Czech. In *Proceedings of the Third International Workshop: Text, Speech and Dialogue,* pp 211-216.

Katz M S 1987 Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transaction on Acoustics, Speech, and Signal Processing* 35(3):400-401.

Porter M F 1980 An algorithm for suffix stripping. *Program* 14(3).130-137.

Voorhees E M 1986 *Implementing agglomerative hierarchic clustering algorithms for use in document retrieval.* Unpublished TR 86-765, Cornell University.

Joachims T 1996 *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization.* Unpublished CMU-CS-96-118, Carnegie Mellon University.

Popovič M 1991 *Implementation of a Slovene language-based free-text retrieval system.* Unpublished PhD thesis, University of Sheffield.