EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Classification of heterogeneous text data for robust domain-specific language modeling

Ján Staš*, Jozef Juhár and Daniel Hládek

Abstract

The robustness of *n*-gram language models depends on the quality of text data on which they have been trained. The text corpora collected from various resources such as web pages or electronic documents are characterized by many possible topics. In order to build efficient and robust domain-specific language models, it is necessary to separate domain-oriented segments from the large amount of text data, and the remaining out-of-domain data can be used only for updating of existing in-domain *n*-gram probability estimates. In this paper, we describe the process of classification of heterogeneous text data into two classes, to the in-domain and out-of-domain data, mainly used for language modeling in the task-oriented speech recognition from judicial domain. The proposed algorithm for text classification is based on detection of theme in short text segments based on the most frequent key phrases. In the next step, each text segment is represented in vector space model as a feature vector with term weighting. For classification of these text segments to the in-domain and out-of domain area, document similarity with automatic thresholding are used. The experimental results of modeling the Slovak language and adaptation to the judicial domain show significant improvement in the model perplexity and increasing the performance of the Slovak transcription and dictation system.

Keywords: Document similarity; Language modeling; Speech recognition; Term weighting; Text classification; Topic detection

1 Introduction

With an increasing amount of the text data gathered from various web pages or electronic documents and growing need for more accurate and robust models of the Slovak language [1], a question of how to classify the text data according to their content arises even more than expected. This question is getting on importance with using heterogeneous text corpora, in which we do not have any knowledge about the document boundaries. In the case of the *task-oriented speech recognition* and *domain-specific language modeling* [2], these heterogeneous text data bring many ambiguities caused by the overestimating such *n*-gram probabilities that are typically unrelated with the area of speech recognition into the process of the training language models. Therefore, we were looking for a way of classification of the text data into predefined domains as good way as possible and adjustment of the parameters of language modeling for effective *large vocabulary continuous speech recognition* (LVCSR).

There are two ways existing for assigning text data into domains; using text classification or document clustering with topic detection. The difference between them is that the *text classification* is based on assigning the text data into two or more predefined classes, whereas *document clustering* tries to group similar documents into a number of classes and find some relationship between them. The similarity of two documents represented by their feature vectors is usually based on computing cosine of the angle between them [3]. After clustering, the topic detection for every cluster of documents is needed [4]. Unlike clustering, the classification is supervised learning technique and requires the training data for classifying new documents. Considering fact that we need to group text documents only into two classes, we focused our research on the text classification techniques.

A growing number of statistical methods have been applied to the problem of text classification in recent

*Correspondence: jan.stas@tuke.sk
Department of Electronics and Multimedia Communications, Technical University of Košice, Park Komenského 13, 041 20 Košice, Slovakia

years, including *naïve Bayes classifier* and *probabilistic language models* [5,6], similarity-based approaches using *k-nearest neighbor classifier* [5,7], *decision trees* and *neural networks* [8], *support vector machines* [5,9], or *semi-supervised clustering* [10]. When large amount of documents is used, these algorithms usually suffer from a very high computational complexity. Moreover, for correct estimation of parameters of these classification algorithms, a training corpus is needed. Therefore, we proposed an algorithm based on computing similarity between two documents and decision, which one will appertain to the domain and one which will not, using a threshold value calculated automatically on a development data set. This simple and effective algorithm classifies short text segments (such as paragraphs) from heterogeneous text corpora gathered from various resources to the in-domain and out-of-domain data. Classified text data are then used in statistical language modeling for enhancing its quality and robustness in the task-oriented speech recognition.

The rest of this paper is organized as follows. Section 2 starts with a short overview about the source data used either for text classification, training acoustic and language models, and testing the Slovak LVCSR system. Our proposed approach for text classification based on the key phrase identification, term weighting, measuring similarity between two documents, and automatic thresholding is introduced in the Section 3. Section 4 presents the speech recognition setup used for evaluating language models trained on classified text corpora. The experimental results with adapted models of the Slovak language into the selected domain are discussed in the Section 5. Finally, Section 6 summarizes the contribution of our work and concludes this article with future directions.

2 Source data

2.1 Acoustic database

For testing language models using speech recognition system, the Slovak acoustic database was created, on which acoustic models have been trained. Speech database consists of three subsets (see the Table 1):

- The first part is characterized by gender-balanced speakers, contains 250 h of speech recordings obtained from 250 speakers together and consists of two parts: APD1 and APD2 databases. The APD1 database includes 100 h of readings of real adjustments from the court with personal data changed, recorded in sound studio conditions. The APD2 database consists of 150 h of read phonetically rich sentences, web texts, newspaper articles, short phrases, and spelled items, recorded in conference rooms using table and close-talk headset microphones [2].
- The second PAR database includes 90 h of 90% male and 10% female speech recordings realized in the main conference hall of the Slovak Parliament using conference gooseneck condenser microphones [11].
- The mixture of Broadcast news (BN) databases consists of 145 h of speech recordings acquired from main and morning TV shows and 35 h from broadcast news and TV and radio shows, together realized with TV DVB-S PCI card [12].

All speech recordings were downsampled to 16-kHz 16-bit PCM mono format for training and testing. The whole acoustic database was manually annotated by our team of trained annotators using the Transcriber tool [12], double checked, and corrected.

2.2 Text corpora

The main part of text corpora used for text classification and statistical language modeling was created by using our proposed *system for gathering text data* from various web pages and electronic resources written in Slovak language [1]. From the retrieved text data, there was a large amount of numerals, symbols, or grammatically incorrect words filtered out and the rest of the data were normalized into their pronounced form by additional processing, such as word tokenization, sentence segmentation, numerals transcription, and abbreviations expanding. The processed text corpora were later divided into smaller domain-specific subcorpora ready for the training language models. Contemporary text corpora consists of following subsets:

Table 1 Acoustic database description

Acoustic database	Hours	Sampling (kHz)	Resolution (bit)	Microphone type	Sound environment and conditions
APD1 database	100	48	16	Close-talk headset	Sound studio conditions
APD2 database	150	48	16	Close-talk headset	Offices and conference rooms
PAR database	90	44	24	Gooseneck condenser	Main conference hall of the Slovak Parliament
BN1 database	145	48	16	TV DVB-S PCI card	Sound studio, telephone, and degraded speech
BN2 database	35	48	16	TV DVB-S PCI card	Sound studio, telephone, and degraded speech
Evaluation data set	5.25	48	16	Close-talk headset	Sound studio, offices, and conference rooms

- *Slovak web corpus* was collected by crawling whole web pages from various Slovak domains saved with information about date, title, URL, extracted text, and HTML source code.
- *Corpus of newspapers* is a collection of articles that have been gathered from the most popular online news portals, magazines, and journals in the Slovak Republic. This corpus was extended by a large amount of newspaper articles downloaded via RSS channels and collection of manually corrected speech transcriptions of four main TV broadcast news and five radio shows.
- *Corpus of legal texts* (judicial corpus) was obtained from the Ministry of Justice of the Slovak Republic in order to develop the automatic dictation system for their internal purpose [2].
- *Corpus of fiction texts* was created from 1,625 electronic books and other stories freely available on the Internet written in Slovak language.
- *Corpus of contemporary blogs* consists of web-extracted blog texts from main news portals in the Slovak Republic saved without contribution's comments.
- *Development data set* (held-out data) was created from 10% randomly selected sentences from (in-domain) corpus of legal texts that were not used in the process of training language models.
- *Speech annotations* (transcriptions) of data obtained from acoustic database are a special portion of the text corpus. Transcriptions also contain a large amount of filled pauses and additional disfluent speech events together with useful text. We have discovered that filled pauses have a positive effect on the quality of language modeling, both for dictated or spontaneous speech. Therefore, we decided to include these speech transcriptions into the process of language modeling.

The complete statistics on the total number of tokens and sentences for particular text subcorpus are summarized in the Table 2.

Table 2 Statistics on the text corpora

Text corpus	Tokens	Sentences	Documents
Slovak web corpus	748,854,697	50,694,708	2,803,412
Corpus of newspapers	554,593,113	36,326,920	2,022,483
Corpus of legal texts	565,140,401	18,524,094	1,503,271
Corpus of fiction texts	101,234,475	8,039,739	367,956
Corpus of contemporary blogs	55,711,674	4,071,165	211,533
Development data set	55,163,941	1,782,333	165,577
Speech annotations	4,434,217	485,800	5,520
Total	2,085,132,518	119,924,759	7,079,752

Moreover, each text corpus was annotated using our proposed *Slovak morphological classifier* [13] based on a hidden Markov model (HMM) together with suffix-based word clustering function and restricted by *manually morphologically annotated lexicon of words*. The HMM has been trained on trigram statistics generated from *morphologically annotated corpus* together with the lexicon delivered by the Slovak National Corpus [14]. Note that the morphologically annotated corpus were then used in the process of extraction of key phrases from development data set of the proposed algorithm for classification of heterogeneous text data.

3 Proposed text classification approach

As it was mentioned before, we proposed an effective approach for classification of heterogeneous text corpora into the two data sets, the in-domain and out-of-domain data, to increase the robustness of domain-oriented statistical language modeling in the Slovak LVCSR system. Our algorithm is based on identifying key phrases with their occurrences in short text segments. Each text document is represented as a vector of key phrases in a vector space (a key phrase/document matrix). For reducing the influence of frequent key phrases in documents, term weighting was applied. The next step includes measuring the similarity between reference and examined document to determine the closeness between them. Based on the automatic thresholding, the algorithm then decides which text document belongs or does not belong to the examined domain (in our case to the judicial one). The block scheme of the proposed text classification approach is depicted in Figure 1.

In the following sections, the proposed text classification approach is described in more detail.

3.1 Key phrase extraction

The first step in the process of classification of the text data is to propose an algorithm for extracting key phrases from examined domain (from development data). Based on morphologically annotated corpora, described in the Section 2.2, we created a set of 14 *morpho-syntactic patterns* for extracting bigrams, trigrams, and quadrigrams from this corpora, summarized in the Table 3. Morpho-syntactic patterns take into account part of speech of the corresponding words and syntactic dependency between them, unlike other statistical approaches based on computing pointwise mutual information, t score or χ^2 score between n words. In order to prevent any occurrence of key phrases from other domains in this list, we filtered out all key phrases from the other out-of-domain corpora, except corpus of legal texts. Using this approach, we created a list of 5,210 in-domain key phrases that are later used in the block key phrase identification and measuring similarity between two documents. More details and

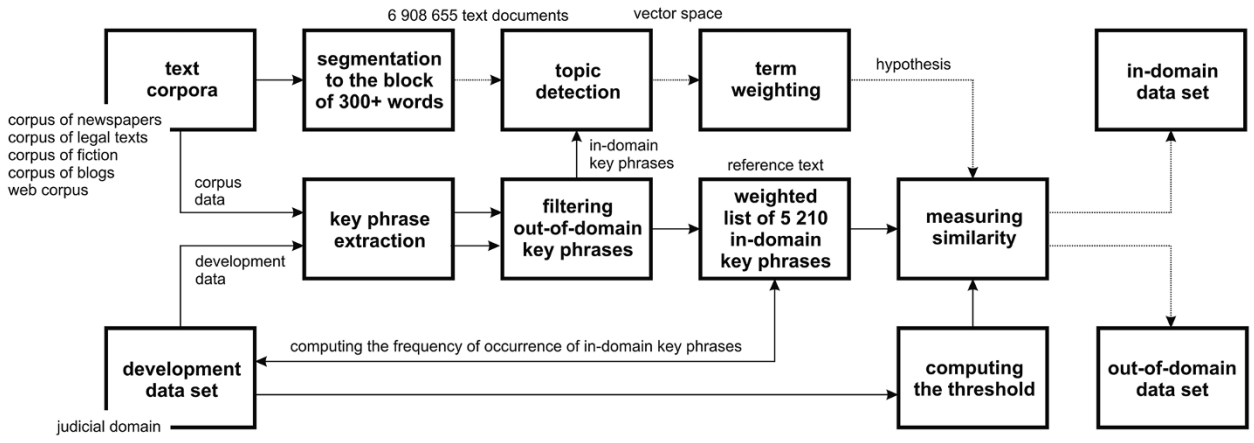

Figure 1 The block scheme of the proposed approach for text classification.

background on how the set of morpho-syntactic patterns were created can be found in [15].

3.2 Text segmentation

In general, text data gathered from the Internet are characterized by a large variety of domains or topics that are contained in the web articles, from which the text corpus is composed. Moreover, in case of large-scale text documents, they may also contain more than one theme within. As it was mentioned earlier, this problem is gaining on importance when using heterogeneous text corpora, in which we have no knowledge about the document boundaries. Therefore, the next step in the text classification process includes segmentation of the used text corpora into the *small segments* (paragraphs) *with at least 300 words*. This value was determined empirically from the statistical observation and expresses the average number of words contained in one paragraph of a web-based article. By application of segmentation rules, we obtained a total of 6,908,655 short (300+ words) text segments - documents - entering to the process of text classification. The statistics on the number of documents after text segmentation for particular subcorpus are resumed in the Table 2.

Table 3 Morpho-syntactic patterns

Type	Characterization	Scheme
2-gram	Adjective + noun	AS
	Numeral + noun	NS
	Noun + noun	SS
	Abbreviation + noun	WS
3-gram	Adjective + adjective + noun	AAS
	Adjective + noun + noun	ASS
	Adverb + adjective + noun	DAS
	Numeral + adjective + noun	NAS
	Noun + adjective + noun	SAS
	Noun + preposition + noun	SES
	Noun + numeral + noun	SNS
4-gram	Noun + preposition + adjective + noun	SEAS
	Noun + preposition + noun + noun	SESS
	Noun + noun + conjunction + noun	SSOS

3.3 Key phrase identification

In the next step, the key phrases were used in *computing the frequency of their occurrence* in examined text segments of 300+ words. The key phrase identification process is similar to any topic detection approach. However, in this process we have not considered removal of stop-words, because key phrases extracted using proposed morpho-syntactic patterns contained such part-of-speech classes as prepositions or conjunctions (see the Table 3). Also lemmatization (or stemming) is very time-consuming and would cause high memory requirements, therefore it has not been introduced into this process of text classification. Note that text segments that did not contain any key phrases were automatically classified as out-of-domain data.

3.4 Vector space modeling

One of the simplest way how to represent the occurrence of terms (key words or key phrases) in any text document is to use a *vector space model* (VSM). In each ith document, $\bar{d}_i$ is represented as a feature vector of the terms t_j that appear in this document as follows [5]:

$$\bar{d}_i = (t_{i,1}, t_{i,2}, \ldots, t_{i,N}). \quad (1)$$

Using this approach, each short text segment was represented by a vector of 5,210 key phrases. With respect to the number of documents in the collection (see the

Section 3.2), we have received the matrix with 5,210 columns and 6,908,655 rows. However, the main disadvantage of such representation is a very high dimension of this matrix and sparsity of values in the vector space, resulting in very high requirements on its storage.

3.5 Term weighting

As it was mentioned earlier, term weighting was applied as a feature selection algorithm for reducing the influence of frequently occurring terms in a collection of documents. In this research, we have tested three different weighting schema: (a) tf-idf, (b) Okapi BM25, and (c) Ltu factors.

The conventional term weighting came from the computing frequency of a term in a document using the *term frequency* and the frequency in the collection of documents in which the term appears, which is expressed as the *document frequency*. A number of term weighting schemes based on these two frequency functions exist such as idf - *inverse document frequency*, expressed by the negative reciprocal value of the document frequency; ridf - *residual* idf, defined as the difference between actual idf and logarithm of idf predicted by Poisson distribution in a term distribution model; tf-idf and tf-ridf that combines term frequency and document frequency into one algorithm, which can be scaled *logarithmically* or normalized by *augmented* version [5]. Moreover, term weighting does not have to be performed on the entire collection of documents. It can be calculated on a small training corpus and used in clustering dynamic data streams using tf-icf weighting [16].

Based on the previous research [17] focused on a comparative study of term weighting schemes, we observed that standard tf-idf achieved the best results in clustering of the Slovak text documents obtained from Wikipedia. Therefore, we used this weighting scheme in the proposed classification too.

The tf-idf is a standard term weighting scheme used in information retrieval or data mining and combines term frequency and inverse document frequency together. The importance of tf-idf increases proportionally to the occurrence of a word in the document and is offset by the frequency of the word in the collection of documents according to formula [5]

$$w_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i = \frac{f_{i,j}}{\sum_k f_{k,j}} \times \log \frac{N}{\mathrm{df}_i}, \tag{2}$$

where $f_{i,j}$ is the number of occurrence of a term t_i in a document d_j and sum in the denominator of $\mathrm{tf}_{i,j}$ component expresses the number of occurrences of all terms t_i in d_j. Then, N is the total number of documents, and the denominator of idf_i component expresses the total number of documents in a collection that t_i occurs in well-known as document frequency df_i.

Contemporary term weighting schemes take into account additional factors such as *maximum of term frequency* $\max(\mathrm{tf}_{i,j})$ in a document, *length of a document* dl_i, or *average document length* dl_{avg} in a collection of documents. Between these, we can fit a simple *automated text classification* (ATC), which uses the idf as the term importance factor and Euclidean vector length as the document length normalization factor, either Okapi BM25 or Ltu scoring [18] that were used in our experiments.

The Okapi BM25 score is defined as a bag-of-words retrieval function that ranks a collection of documents regardless of the inter-relationship between the terms within a document [5]. It is based on computing BM25-tf score and idf component derived from the binary independence model that is well-known from the probabilistic theory in the information retrieval [19]:

$$w_{i,j} = \mathrm{BM25} - \mathrm{tf}_{i,j} \times \mathrm{idf}_i^* = \frac{\mathrm{tf}_{i,j}}{0.5 + 1.5 \times \frac{\mathrm{dl}_i}{\mathrm{dl}_{\mathrm{avg}}} + \mathrm{tf}_{i,j}} \times \log \frac{N - \mathrm{df}_i + 0.5}{\mathrm{df}_i + 0.5}, \tag{3}$$

where $\mathrm{tf}_{i,j}$ means term frequency, N is a total number of documents in the collection, df_i presents the document frequency, dl_i document length and $\mathrm{dl}_{\mathrm{avg}}$ the average document length for the collection. In addition, we can put the Okapi BM25 scoring into the tf-idf scheme, which was presented in [20].

In Ltu term weighting scheme, L factor expresses the *logarithm of the term frequency*, t factor the *inverse document frequency*, and u the *length normalization* factor as follows [21]:

$$w_{i,j} = \mathrm{L} \times \mathrm{t} \times \mathrm{u} = (\log \mathrm{tf}_{i,j} + 1) \times \log \frac{N}{\mathrm{df}_i} \times \frac{1}{0.8 + 0.2 \times \frac{\mathrm{dl}_i}{\mathrm{dl}_{\mathrm{avg}}}}. \tag{4}$$

As we can see from these equations, both the Okapi BM25 and Ltu scores are only a certain variation of the conventional tf-idf weighting.

The problem of data sparsity and high dimension of VSM after term weighting can be efficiently eliminated using *latent semantic analysis/indexing* (LSA/LSI) or its *probabilistic* (pLSA) version that projects terms and documents into a space of co-occurring terms, also by *principal component analysis* (PCA), based on a *singular value* or *eigen-value decomposition* of a term/document matrices [22]. However, this space reduction is very time-consuming and computationally intensive considering a large amount of documents in our collection. Therefore, they were not implemented into the process of text classification.

3.6 Document similarity measurement

The next step involves measuring similarity of two documents. In this approach, we measured the document similarity between reference and examined texts, not between all documents in a collection, commonly used in the tasks oriented on the document clustering. The reference text contained weighted form of all key phrases which occurred in a development data set. Both reference and examined text documents were represented by the vector of 5,210 key phrases weighted according to the selected weighting scheme, described in the Section 3.5, so they could be compared.

By empirical study of numerous similarity measures described in [23], we have chosen three different measures: (a) Bhattacharyya coefficient, (b) Jaccard correlation index, and (c) Jensen-Shannon divergence, satisfying the conditions of *non-negativity*, *symmetricity*, *triangle inequality*, and *identity*, when distance is equal to 0.

For clustering phonemes in the process of training acoustic models, the *Bhattacharyya coefficient* is often used. In general, it can be used as a classification criterion in many other tasks oriented on clustering in information theory. Therefore, we used this coefficient as one classification criterion. Bhattacharyya coefficient comes from the *sum of geometric means* between two probability density functions and specifies the separability of two classes x and y as follows:

$$d_{\text{Bha}} = -\ln \sum_{i=1}^{N} \sqrt{x_i y_i}. \tag{5}$$

On the contrary, *Jaccard correlation index* is defined as a *harmonic mean* between two probability density functions and expresses a scalar sum of two vectors. It comes from equation on computing cosine similarity [5], normalized by absolute deviation of two distributions x and y according to the formula

$$d_{\text{Jac}} = \frac{\sum_{i=1}^{N} (x_i + y_i)^2}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} x_i y_i}. \tag{6}$$

Jensen-Shannon divergence comes from the principle of uncertainty. It is often used in information theory and natural language processing as a special case of *relative entropy approach* similar to the averaged Kullback-Leibler divergence, satisfying the condition of symmetry in the entire range of values. For two probability density functions x and y, it is computed as

$$d_{\text{JS}} = \frac{1}{2} \sum_{i=1}^{N} x_i \ln \left(\frac{2x_i}{x_i + y_i} \right) + \frac{1}{2} \sum_{i=1}^{N} y_i \ln \left(\frac{2y_i}{x_i + y_i} \right). \tag{7}$$

3.7 Automatic thresholding

The last step in the classification process is to correctly adjust the threshold that determines which documents will appertain to the in-domain and which to the out-of-domain area. In general, this value is usually determined empirically from long-term observation or can be adjusted automatically based on a set of statistic values derived from development data. There are many algorithms for automatic thresholding. A comprehensive study about those can be found in [24].

We used the *median* (centroid) of a sequence of coefficients derived from a set of values determining the similarity of two documents as a method of automatic thresholding (see the Section 3.6). The threshold value was calculated on a development data set and its acquisition shares the same process with classification of the text data described in the previous sections. This means that the development data were divided into short text segments consisting of at least 300 words, represented by VSM through the key phrases, and weighted, and each document was compared with the reference text (weighted list of key phrases) using one of the presented similarity measure. Using this process, we get a list of the coefficients (one coefficient for each document in development data set) expressing distance to the target domain. This list was sorted and the median value was selected as a threshold.

In the Table 4, we can find the statistics of the number of in-domain and out-of-domain documents after applying the proposed classification approach to the segmented text corpora for different term weighting scheme and distance measure used in the step of measuring similarity between the reference and examined documents with automatic thresholding.

The performance between in-domain and out-of-domain language models is summarized in the Table 5. Model perplexity evaluated on a development data set was used for testing the quality of the language models. Its calculation will be introduced in the next section.

4 Speech recognition setup

4.1 Decoding

For evaluation of the quality of language modeling after text classification and performance of the Slovak LVCSR,

Table 4 The number of documents after text classification

Similarity/weighting	tf-idf	Okapi	Ltu
In-domain data set			
Bhattacharyya coefficient	1,166,806	607,004	698,061
Jaccard correlation index	1,258,169	537,729	699,033
Jensen-Shannon divergence	2,305,230	956,243	698,062
Out-of-domain data set			
Bhattacharyya coefficient	5,741,849	6,301,651	6,210,594
Jaccard correlation index	5,650,486	6,370,926	6,209,622
Jensen-Shannon divergence	4,603,425	5,952,412	6,210,593

Table 5 Model perplexity for particular language models computed on development data

Similarity/weighting	tf-idf	Okapi	Ltu
In-domain data set			
Bhattacharyya coefficient	14.1223	15.7542	17.2876
Jaccard correlation index	14.0815	14.8402	17.2872
Jensen-Shannon divergence	15.0343	15.4863	17.2878
Out-of-domain data set			
Bhattacharyya coefficient	90.6770	25.7417	183.670
Jaccard correlation index	75.0398	20.7094	162.901
Jensen-Shannon divergence	99.8450	24.3595	187.167

we configured a speech recognition setup based on Julius, an open-source continuous speech recognition engine [25]. Julius uses *two-level Viterbi search algorithm*, when input speech is processed in the forward search with bigram model, and the final backward search is performed again using the result obtained from the first search to narrow the search space with reverse language model of the highest order (in our case with trigram model). Proposed speech recognition setup is depicted in the Figure 2.

4.2 Acoustic modeling

The speech recognition setup involves a set of *triphone context-dependent acoustic models* based on HMMs. All models have been generated from feature vectors containing 39 *mel-frequency cepstral* (MFC) *coefficients*, where each of four states had been modeled by 32 Gaussian mixtures. Acoustic models have been trained on four databases of annotated speech recordings, described in the Section 2.1, using HTK Toolkit. The training set also involves model of silence, short pause, and additional noise events. Rare triphones have been modeled by the *effective triphone mapping algorithm* [11].

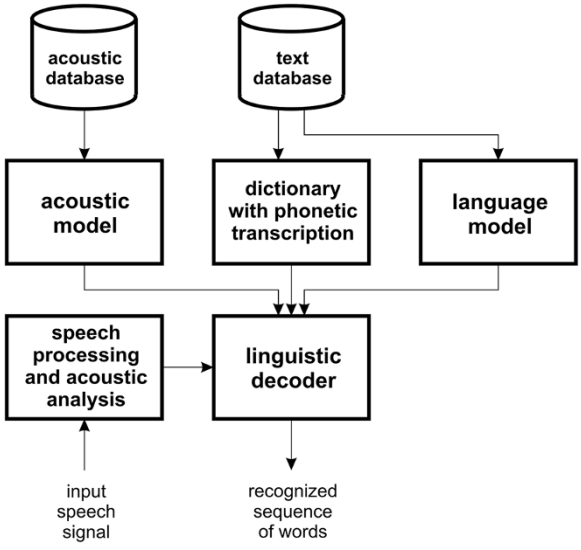

Figure 2 The Slovak LVCSR system.

4.3 Language modeling

The experimental results have been performed taking an advantage of *trigram models* created using the SRI LM Toolkit [26], restricted by the vocabulary size of 325,555 unique words and smoothed by the *Witten-Bell back-off algorithm*. All models have been trained on the processed text corpora size of about 2 billion of tokens in 120 million of sentences (see the Table 2) and divided into two parts, to the in-domain and out-of-domain data, after text classification (see the Table 4). Particular models trained on in-domain and out-of-domain data were combined with a model trained on the small portion of text data obtained from speech transcriptions (see the Table 2). Finally, the resulting trigram model was composed from three independent models and *adapted to the judicial domain* using *linear interpolation* with computing interpolation weights by our proposed algorithm based on the *minimization of perplexity* on a development data set. The complete process of building the Slovak language models is depicted on the Figure 3 and described in [1].

In this article we have compared the contribution of changes performed in the vocabulary, also using better text preprocessing steps, adding new text data, or introducing new principles into the Slovak language modeling during the recent time periods. These contributions and differences between language models are summarized in the Table 6.

During this period, the named entities such as people names, surnames, and geographical items were assigned into the word classes in recognition dictionary. The vocabulary has been continually updated with the new words, checked, and corrected. We have introduced filled pauses into the language modeling as transparent words and model some geographically named entities as multiwords. We have also tested a number of methods for language model adaptation to the ted domain and algorithms for text classification and clustering.

4.4 Evaluation

For evaluation of the Slovak language models after text classification, three standard measures have been used.

Accuracy (Acc) and *Correctness* (Corr) are the standard extrinsic measures for evaluating the performance of the LVCSR system. If N is the total number of words in an evaluation data set (reference), S, I, and D reflect the total number of substituted, inserted, and deleted words in recognized hypothesis, respectively, and $H = N - (S + D)$ is the total number of words in hypothesis, then

$$\text{Acc} = \frac{H - I}{N} \quad \text{and} \quad \text{Corr} = \frac{H}{N}. \tag{8}$$

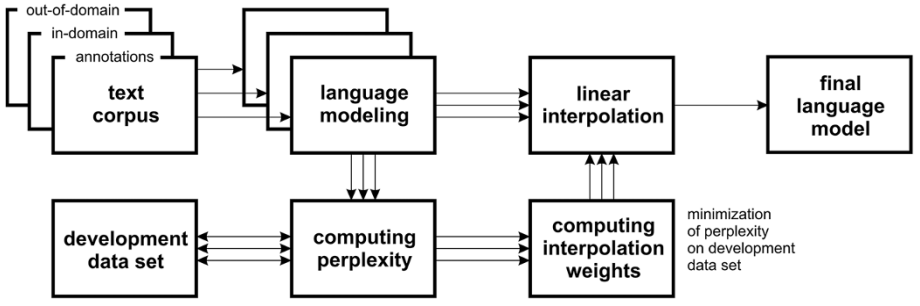

Figure 3 The block scheme of the process of building the Slovak language models.

For intrinsic evaluation of the quality of language modeling, the model perplexity has been used. *Model perplexity* (PPL) is defined as the reciprocal value of the weighted (geometric) probability assigned by the language model to each word in the test set and is related to cross-entropy $H(W)$ by the equation

$$\mathrm{PPL} = 2^{H(W)} = \frac{1}{\sqrt[n]{P(W)}}, \tag{9}$$

where $P(W)$ is the probability of sequence of n words in a language model.

The *evaluation data set* used for testing the performance of the LVCSR system and the quality of the Slovak language modeling after text classification were represented by randomly selected segments from the APD databases (see the Section 2.1, Table 1) containing 1,950 male and 1,476 female speech utterances with total length of about 5.25 h. These speech segments were not used in the training of acoustic models and contain 41,868 words in 3,426 sentences and short phrases. We have decided to include also short phrases in the test set because people make pauses in real conditions not only on the sentence boundaries, but also on phrase boundaries, usually before conjunctions.

5 Experimental results

The experiments have been oriented on the evaluation of the model perplexity and performance of the Slovak LVCSR system on the evaluation (test) data after text classification and statistical modeling of the Slovak language from judicial domain. The selection of this domain was intentional concerning our research oriented on development of the Slovak automatic dictation and transcription system for the Ministry of Justice of the Slovak Republic in recent years [2]. The same approach for text classification and statistical language modeling can be also used for several other domains, in the task of broadcast news transcription, meeting speech recognition, etc.

As it was mentioned in the Section 3, the statistics on the numbers of in-domain and out-of-domain documents after text classification regarding the used term weighting scheme in combination with selected similarity measure are resumed in the Table 4.

Table 6 Differences in the text processing and language modeling during the recent time periods

	Period				
	Dec 2011	**Jul 2012**	**Dec 2012**	**Apr 2013**	**May 2013**
No. of pronunciation variants	475,156	475,357	474,456	474,453	474,453
No. of unique word forms	326,299	326,295	325,555	325,555	325,555
No. of words under classes	97,471	97,680	97,678	97,678	97,678
No. of classes of words	20	22	22	22	22
No. of transparent words	4	5	5	5	5
Vocabulary extension	•	•	•	•	-
Word classes extension	•	•	-	-	-
Adding new text data	•	-	-	•	•
Additional text processing	•	-	•	•	•
Filled pause modeling	-	•	•	•	•
New text classification	•	-	-	-	•

• Change was performed.

Table 7 Language model perplexity and performance of Slovak LVCSR system with different acoustic models

PPL	Text classification		APD1+APD2 250 h (table mic.) sp. adapt.: no eval. set: gender-bal.		APD1+APD2 250 h (close-talk mic.) sp. adapt.: no eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: no eval. set: gender-bal.		APD1+APD2 +PAR+BN 520 h sp. adapt.: no eval. set: gender-bal.	
	Weighting	Similarity	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %
40.4302	Reference language model		91.84	93.08	93.61	94.51	94.36	95.13	94.06	94.89
36.0428	tf-idf	Bhattacharyya	92.44	93.64	93.99	94.85	94.70	95.46	94.36	95.13
35.9444		Jaccard index	92.46	93.65	93.97	94.85	94.72	95.47	94.37	95.16
38.1756		Jensen-Shannon	92.23	93.39	93.78	94.70	94.50	95.25	94.21	94.99
38.1289	Okapi	Bhattacharyya	92.17	93.34	93.77	94.65	94.61	95.34	94.27	95.02
39.9782		Jaccard index	92.10	93.31	93.60	94.54	94.48	95.21	94.11	94.89
39.2267		Jensen-Shannon	92.27	93.42	93.77	94.67	94.61	95.36	94.18	94.95
40.1325	Ltu	Bhattacharyya	91.86	93.12	93.57	94.51	94.42	95.16	94.05	94.87
40.1439		Jaccard index	91.87	93.12	93.56	94.50	94.40	95.16	94.04	94.87
40.1319		Jensen-Shannon	91.87	93.12	93.57	94.51	94.42	95.16	94.05	94.87

As we can see from this results, we achieved the best class separation of in-domain and out-of-domain data in combination of Okapi BM25 weighting with similarity based on computing Jaccard correlation index. Using this combination, we yielded the in-domain data with the best possible concentration of key phrases in it. On the contrary, the worst separation of classes was observed when using tf-idf weighting and Jensen-Shannon divergence. Although this combination gives the largest number of text documents in the in-domain corpus, it has a much weaker concentration of key phrases in it. If we review the Ltu weighting, similar results of class separation were noticed for any similarity measure we have chosen. It would be interesting to discover the overlap between classes for the same term weighting and different distance/similarity measure. Their intersection or union could produce more interesting results in the future.

However, if we look at the performance between in-domain and out-of-domain language models using perplexity evaluated on development data summarized in the Table 5, the text classification using tf-idf weighting with measuring similarity based on computing Jaccard correlation index or Bhattacharyya coefficient predetermines the

Table 8 Language model perplexity and performance of the Slovak LVCSR system with gender-dependent acoustic models

PPL	Text classification		APD1+APD2 +PAR 340 h sp. adapt.: female eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: male eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: female eval. set: female sp.		APD1+APD2 +PAR 340 h sp. adapt.: male eval. set: male sp.	
	Weighting	Similarity	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %
40.4302	Reference language model		90.15	91.68	92.72	93.80	95.72	96.48	94.10	94.87
36.0428	tf-idf	Bhattacharyya	91.23	92.50	93.23	94.18	95.97	96.68	94.34	95.06
35.9444		Jaccard index	91.26	92.55	93.24	94.22	95.98	96.68	94.73	95.11
38.1756		Jensen-Shannon	90.71	92.10	92.92	93.94	95.81	96.54	94.23	94.94
38.1289	Okapi	Bhattacharyya	90.95	92.23	93.03	94.01	95.88	96.59	94.25	94.96
39.9782		Jaccard index	90.59	91.99	92.82	93.84	95.81	96.53	94.17	94.90
39.2267		Jensen-Shannon	90.93	92.27	93.00	93.97	95.94	96.65	94.17	94.89
40.1325	Ltu	Bhattacharyya	90.19	91.70	92.72	93.78	95.73	96.49	94.10	94.85
40.1439		Jaccard index	90.18	91.70	92.73	93.78	95.76	96.51	94.11	94.86
40.1319		Jensen-Shannon	90.18	91.70	92.72	93.78	95.73	96.49	94.10	94.85

Table 9 Model perplexity and performance of Slovak LVCSR system with different language and acoustic models

PPL	Language model (period)	APD1+APD2 250 h (table mic.) sp. adapt.: no eval. set: gender-bal.		APD1+APD2 250 h (close-talk mic.) sp. adapt.: no eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: no eval. set: gender-bal.		APD1+APD2 +PAR+BN 520 h sp. adapt.: no eval. set: gender-bal.	
		Acc %	Corr %	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %
44.9254	Dec. 2011	91.89	93.09	93.44	94.39	94.21	94.98	93.90	94.68
38.9688	Jul. 2012	92.33	93.55	93.78	94.69	94.46	95.26	94.30	95.11
40.2543	Dec. 2012	92.47	93.66	93.86	94.77	94.65	95.43	94.38	95.19
44.3262	Apr. 2013	92.35	93.56	93.76	94.69	94.53	95.33	94.30	95.12
35.9444	May 2013	92.46	93.65	93.97	94.85	*94.72*	*95.47*	94.37	95.16

optimal combination in terms of the quality the text segments used only for in-domain language modeling. Using Ltu factor, we observed significant degradation in the perplexity of language models trained not only on in-domain, but also on out-of-domain text data for each selected similarity measure. This is probably caused by inappropriate setting of a threshold in the last step of the proposed algorithm.

As regards the overall results performed on the randomly selected speech utterances from judicial domain, the first part of the experiments presented in the Tables 7 and 8 were oriented on the computing of model perplexity and performance of the Slovak LVCSR system after language modeling trained on classified text corpora using proposed approach.

The first table summarizes the performance of the Slovak language modeling using acoustic models trained on different speech databases, described in the Section 2.1. The results have shown that increasing the amount of acoustic data that were close to the examined domain with similar recording environment improved the recognition accuracy. On the other hand, the BN database degraded the results because the recording environment was quite different to the evaluation data selected from the APD databases.

The second table presents the quality of language modeling using gender-dependent acoustic models (optimized to male and female speech) trained on the APD1, APD2 and PAR databases, giving the best results in previous experiment.

In the first two columns of the Table 8, the experimental results with acoustic models adapted to the male and female gender of speaker evaluated on the whole test data set are presented. The next two columns show the performance of language models in combination of gender-dependent acoustic models evaluated on the test speech utterances per gender.

As we can see from these results, gender-dependent acoustic modeling can significantly improve the recognition accuracy. If we look at the language model perplexity, we have achieved significant reduction about 11% relatively in comparison with the reference model trained on unclassified text corpora, if we applied combination of tf-idf weighting with similarity based on Jaccard correlation index in the text classification process. Similar results were obtained in the accuracy and correctness evaluated by the LVCSR system. Slightly worse results were noticed when using the Okapi BM25 and Ltu weighting in combination with one of the selected similarity measure. However, we can say that the proposed text classification approach had a significant impact on the overall robustness of the Slovak language modeling.

The second part of the experiments presented in the Tables 9 and 10 show the progress of acoustic and

Table 10 Model perplexity and performance of the Slovak LVCSR system with different language and gender-dependent acoustic models

PPL	Language model (period)	APD1+APD2 +PAR 340 h sp. adapt.: female eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: male eval. set: gender-bal.		APD1+APD2 +PAR 340 h sp. adapt.: female eval. set: female sp.		APD1+APD2 +PAR 340 h sp. adapt.: male eval. set: male sp.	
		Acc %	Corr %	Acc %	Corr %	Acc %	Corr %	Acc %	Corr %
44.9254	12/2011	90.34	91.70	92.68	93.72	95.77	96.48	93.93	94.72
38.9688	07/2012	91.23	92.53	93.18	94.24	95.85	96.61	94.21	95.00
40.2543	12/2012	*91.28*	*92.60*	93.22	94.25	95.93	*96.70*	94.30	95.05
44.3262	04/2013	91.26	92.58	93.24	94.22	95.92	96.67	94.21	94.99
35.9444	05/2013	91.26	92.55	*93.25*	*94.27*	*95.97*	96.68	*94.73*	*95.11*

language modeling in development of the Slovak transcription and dictation system from the judicial domain, observed during the recent time periods.

With increasing amount of acoustic and linguistic data from the judicial domain, using gender-dependent acoustic modeling and speaker adaptation based on maximum likelihood linear regression (MLLR) as well as much better text preprocessing and classification for robust domain-specific language modeling, we achieved the speech recognition accuracy nearly 95% with a significant decrease in language model perplexity. Besides the better text processing and classification of training data, this result was achieved either by introducing classes of names, surnames, and other named entities into the recognition dictionary; representation of geographically named entities and technical terms by multiword expressions; by modeling of filled pauses in a language; or by effective adaptation of language models to the ted domain (see the Table 6).

In the future, we want to build also a new evaluation data set containing different acoustic environments to compare the performance of the Slovak LVCSR system for mixed end-user environments.

6 Conclusions

This paper proposed an algorithm for classification of heterogeneous text corpora to the in-domain and out-of-domain data with the aim of increasing robustness and quality of the statistical language modeling in task-oriented continuous speech recognition. By combining straightforward and effective methods used for text classification and document clustering based on topic detection with key phrases in short text segments, term weighting, measuring similarity between documents and automatic thresholding, we have achieved significant improvement in the quality of modeling of the Slovak language and performance of the Slovak automatic transcription and dictation system. The proposed algorithm can also be used in classification of heterogeneous text corpora into the other domains depending on the used development data.

Further research should be also focused on a better key phrase extraction in fully unsupervised manner without using morphologically annotated corpora or application of dimensionality reduction based on singular value decomposition and using latent semantic indexing or principal component analysis for better representation of text documents in the vector space despite of very high time and memory requirements of this process. Based on the initial tests with document clustering using the latent Dirichlet allocation, our proposed classification approach gives the similar results in the model perplexity as well as the recognition accuracy of the Slovak LVCSR system.

Besides the better text preprocessing and classification of the training data, the robustness and quality of modeling of the Slovak language can be enhanced by addition of large amount of text data from transcripts of real speech recordings, introducing modeling of disfluent speech in a language, or by adaptation of language models to a specific user, group of users, or conversation, depending on the speech recognition task in which they will be used, for example, broadcast news transcription or meeting speech recognition.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The research presented in this paper was partially supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research projects MS SR 3928/2010-11 (20%) and VEGA 1/0386/12 (30%) and the Research and Development Operational Program funded by the ERDF under the project ITMS-26220220141 (50%).

Received: 20 November 2013 Accepted: 26 March 2014
Published: 15 April 2014

References

1. J Juhár, J Staš, D Hládek, in *New Technologies - Trends, Innovations and Research*, ed. by C Volosencu. Recent progress in development language model for Slovak large vocabulary continuous speech recognition (InTech Open Access, Rijeka, 2012), pp. 261–276
2. Rusko M, J Juhár, M Trnka, Staš J, S Darjaa, D Hládek, R Sabo, M Pleva, Ritomský M, Ondáš S, in *Proceedings of the 6th Language and Technology Conference on HLT*. Recent advances in the Slovak dictation system for the judicial domain (Poznań, LTC, 2013), pp. 555–560
3. A Huang, in *Proceedings of the 6th New Zealand Computer Science Research Student Conference*. Similarity measures for text document clustering (Christchurch, NZCSRSC, 2008), pp. 49–56
4. L Yue, S Xiao, X Lv, T Wang, in *Proceedings of 2011 International Conference on Mechatronic Science, Electric Engineering and Computer*. Topic detection based on keyword (Jilin, MEC, 2011), pp. 464–467
5. CD Manning, P Raghavan, H Schütze, *Introduction to Information Retrieval*. (Cambridge University Press, Cambridge, 2009)
6. F Peng, D Schuurmans, S Wang, Augmenting naïve Bayes classifiers with statistical language models. Inf. Retr. **7**(3–4), 317–345 (2004)
7. S Tan, An effective refinement strategy for KNN text classifier. Expert Syst. Appl. **30**(2), 290–298 (2006)
8. N Remeikis, I Skučas, Melninkaité V, Text categorization using neural networks initialized with decision trees. Informatica. **15**(4), 551–564 (2004)
9. T Joachims, in *Proceedings of the 10th European Conference on ML*. Text categorization with support vector machines: learning with many relevant features (Chemnitz, ECML, 1998), pp. 137–142
10. W Zhang, T Yoshida, X Tang, in *Proceedings of the 2nd International Conference on Business Intelligence and Financial Engineering*. Text classification using semi-supervised clustering (Beijing, BIFE, 2009), pp. 197–200
11. S Darjaa, M Cerňak, M Trnka, M Rusko, in *Proceeding of INTERSPEECH 2011*. Effective triphone mapping for acoustic modeling in speech recognition (Florence, INTERSPEECH, 2011), pp. 1717–1720
12. M Pleva, J Juhár, Building of broadcast news database for evaluation of the automated subtitling service. Communications. **15**(2A), 124–128 (2013)
13. D Hládek, Staš J, J Juhár, Dagger, in *Proceedings of the 54th International Symposium ELMAR 2012*. the Slovak morphological classifier (Zadar, ELMAR, 2012), pp. 195–198
14. R Garabík, in *Proceedings of the 1st Workshop on Intelligent and Knowledge Oriented Technologies*. Slovak morphology analyzer based on Levenshtein edit operations (Bratislava, WIKT, 2006), pp. 2–5
15. Staš J, D Hládek, J Juhár, M Ološtiak, in *Proceedings of the 7th International Conference on Natural Language Processing, Corpus Linguistics and E-learning*. Automatic extraction of multiword units from Slovak text corpora (Bratislava, SLOVKO, 2013), pp. 228–237
16. JW Reed, Y Jiao, TE Potok, BA Klump, MT Elmore, AR Hurson, TF-ICF, in *Proceedings of the 5th International Conference on Machine Learning and*

Applications. a new term weighting scheme for clustering dynamic data sets (ICMLA Orlando, 2006), pp. 258–263
17. Zlacký D, Staš J, J Juhár, A Čižmár, *Term weighting schemes for Slovak text document clustering*. (J. Electr. Electron. Eng, ed.), vol. 6, (2013), pp. 163–166
18. R Jin, C Falusos, AG Hauptmann, in *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval*. Meta-scoring: automatically evaluating term weighting schemes in IR without precision-recall (New Orleans, USA, SIGIR ACM, New York, 2001), pp. 83–89
19. SE Robertson, S Walker, S Jones, MM Hancock-Beaulieu, M Gatford, in *Proceedings of the 3rd Text Retrieval Conference*. Okapi at TREC-3 (Gaithersburg, TREC-3, 1996), pp. 109–126
20. JS Whissell, Clarke ChLA, Improving document clustering using Okapi BM25 feature weighting. Inf. Retr. **14**(5), 466–487 (2011)
21. A Singhal, in *Proceedings of the 6th Text Retrieval Conference*. AT&T at TREC-6 (Gaithersburg, TREC-6, 1998), pp. 215–226
22. S Lee, J Song, Y Kim, An empirical comparison of four text mining methods. J. Comp. Inf. Sys. **51**(1), 1–10 (2010)
23. SH Cha, Comprehensive survey on distance/similarity measures between probability density functions. Intl. J. Math. Model. Methods Appl. Sci. **1**(4), 300–307 (2007)
24. PL Rosin, Edges: saliency measures and automatic thresholding. Technical Note No. I.95.58: Institute for Remote Sensing Applications (1995)
25. A Lee, T Kawahara, in *em Proceedings of the 2009 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. Recent development of open-source speech recognition engine Julius (Sapporo, APSIPA ASC, 2009), pp. 131–137
26. A Stolcke, J Zheng, W Wang, V Abrash, in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*. SRILM at sixteen: update and outlook (Waikoloa, ASRU, 2011), p. 5 pages

doi:10.1186/1687-4722-2014-14
Cite this article as: Staš *et al.*: **Classification of heterogeneous text data for robust domain-specific language modeling.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:14.