Document Categorization in Multilingual Environment

Karel Ježek, Michal Toman

Department of Computer Science and Engineering, University of West Bohemia
Univerzitní 22, Pilsen, Czech Republic
{jezek_ka, mtoman}@kiv.zcu.cz

Abstract

This paper deals with various methods for multilingual document categorization and informs about the results of experiments in which EuroWordNet (EWN) plays the central role and serves as a fundamental problem solving tool. We describe both the algorithmic principles and the methodologies used in our classification system and consequently prove their functionality by experimental results. The aim of experiments was to verify the impact of multilingual collection on the quality of categorization and also find how thesaurus can be used to improve the classification and how the use of multilingual thesaurus can generalize monolingual version of categorization.

1 Introduction

Automatic processing of documents is one of the hottest areas of information science research. A useful and still progressing showing sub area of documents processing research is the task of text documents categorization. Particularly when documents are written in various languages, the way of their classification is not satisfactorily solved so far. Simultaneously this implies a number of additional tasks such language recognition, encoding recognition, word sense disambiguation etc, which have to be solved together with documents topics identification. It is worthy to note that documents topics identification is basically the same problem as the opposite one - documents retrieval and their multilingual solutions can share lots of algorithms. Multilingual tools become more important with respect to continuing integration of society. A convenient example could be Internet, regarded as the largest digital library of the world, which is accessed in 35% from English language zone (users are native English speakers), 14% Chinese, 9% Spanish, 8% Japanese 7% German etc. see [1].

We started our experiments with only two languages – Czech and English. Currently we are collecting materials for processing German and Slovak languages as well. It should be noted that the principles remain the same for any arbitrary number of processed languages.

2 Description of the modules

Our experimental system consists of three parts: 1. lemmatization, 2. indexing and 3. classification. The lemmatization part covers interconnection of words in the categorized document with their basic forms, which are generated from the ISPELL program [3]. Consequently these words can be mapped (for various languages) to EWN sets of synonyms (synsets). Lemmatization complexity depends on a morphological diversity of concrete language. It is relatively simple in one case of English, but much more complicated in the case of e.g. Czech language. Therefore the Czech language lemmatization comprises a morphological analysis [4] as well.

The obvious component of lemmatization is a language recognition module. We have used two independent approaches to language recognition. The first one is based on frequencies of characters in various languages, the second uses stop list (frequently used words not carrying any semantic information). Even though we have experimented only with four languages, the results are excellent for both methods. A combination of methods can scale up the precision if necessary.

The second part involves the indexing of classified documents. It means the document is converted into a sequence of EWN synsets indexes. At this moment the whole document is transferred into a language independent form and subsequent processing – categorization can be executed in the classical way. But at least one problem remains, the problem of words disambiguation.

Disambiguation is a relatively self contained task, which has to be done over the course of lemmatization. There exist words which have identical basic forms but their meanings differ. To decide the right meaning requires knowledge of the word's context. We implemented the disambiguation method based on a Bayes classifier [5]. It is a supervised method using a training corpus [6]. In our implementation we designed and verified some improvements as context constraints with a sliding window, use of syntax relations in the sentence and weight function utilization.

Concerning the categorization methods themselves, we experimented with the Naïve Bayes classifier, the Itemsets classifier, their combination and a classifier based on counting the frequency of terms in documents. All methods are described in cited literature [7], [8], [9]. Results of the methods mentioned (precision and recall of categorization) will be presented in the conclusion and will compare both multilingual and monolingual processing.

2.1 Language recognition

Document processing in the multilingual environment implies a need for a language recognition module which can distinguish different languages and recognize text coding. In some languages (e.g. Central European), there is used different coding in Linux (e.g. ISO-8859-2) and Windows (e.g. CP-1250) environment. A language recognition module is therefor an obvious module for computing system which deals with multilingual text processing.

We have used two different approaches to recognize the language. The first one is based on determination of letters frequency. Each letter in the text is contained in particular language with a certain (in general different) probability. We used these frequencies for language recognition. Letters are also used for text coding recognition. In the case of coding recognition, only a few letters differ in each coding page (scheme), but it is still possible to distinguish text coding according them.

The second approach is based on the use of a stoplist (list of words not carrying any semantic information). Such words are unique for each language and they are good clues for our purpose. Words contained in the stop-list are for example: a, an, the, of, from, at, is, etc. Some preliminary results, combination of both methods gives even better results than the frequency based method.

An example of differences in letters coding used in the frequency based method:

	138	141	142	154	157	158	169	171	174	185	187	190
CP1250	Š	Ť	Ž	š	ť	ž	©	«	®	ą	»	ľ
ISO-8859-2	□	□	□	□	□	□	Š	Ť	Ž	š	ť	ž

Tab. 1, Differences between Czech letters coding

For each coding and language one binary vector is defined with ones at positions which are characteristic for a particular language and coding. During the processing it is created a characteristic vector for the text. This vector is afterwards compared with patterns of all coding pages and languages. Hamming distance is used for comparison of characteristic vectors – how many bits differ between binary strings. The smaller the number the higher is the possibility of such coding. As the languages have different number of the specific characters (eg. 28 for Czech language, 34 for Slovak language, 7 for German language) we have to norm the value. Finally, is used the value computed as division of the hamming distance and the normative coefficient.

2.2 EWN

For language independent processing we have to find out a technique which could transform all the multilingual texts into the form that can be easily processed. For this task we have used the EuroWordNet (EWN) thesaurus (see [2]). In the EWN are sets of synonymous words stored in so called synsets. Each synset has assigned unique identification (index). These indexes are used for indexing in later phases of processing.

EWN is multilingual thesaurus (database of words and relations) for most of European languages (English, Danish, Italian, Spanish, German, French, Czech, Estonian language). Its structure is similar as the original Princeton WordNet. EWN contains sets of synonyms and relations between them. The synsets are connected (among languages) through inter-lingual-index in such way, that the same synset in one language has the same index in another one. Because of this feature we can consider this index as equivalent to the one from standard indexing process.

A disadvantage of the EWN is a too fine-grained synset structure. For synsets with a similar meaning are often defined different indexes. This problem can lead to improper indexing and difficulties during cross-language text processing. We are working on synsets clustering to avoid this drawback. The second hitch is quite small size of some languages contained in EWN. EWN could be anyway extended through bilingual dictionaries. Of course, it wouldn't be possible to build the relations between newly added words.

2.3 Lemmatization

We analyzed all the possibilities for lemmatization and afterwards we have chosen dictionary lemmatization because of its simplicity and generality. This causes some difficulties in the preparation phase – we had to create good lemmatization dictionaries. Algorithmic and dictionary based methods used till now were changed to associate lemmas with Eurowordnet synsets.

We create the lemmatization dictionary by extracting word forms from Ispell program (see [3]). Our main idea lies in the usage of the stems stored in the Ispell dictionary, and the generation of all existing word forms from them. We consider the stem as a basic form of the word which should appear in the thesaurus. It works perfectly in the case of simple languages, e.g. English, but fails for languages with richer flexes. For example in the Czech language, the stem isn't generally the same as the basic form. We can take a Czech word, lano (rope), as an example. The basic form lano differs from the stem lan.

We take into account the assumption that the basic form is an element of the set of all possible forms generated from Ispell. Taking the dictionary created from Ispell, it's possible to generate subsets of word forms for each stem and look for the corresponding lemma in EWN.

The algorithm can be described as follows:

```
-  For each set of forms generated from one stem do
     o  For each form do
          ▪  Search corresponding lemma in EWN
          ▪  If it's found assign the basic form (lemma) for each item from the set
          ▪  If not, continue with the next form
```

In the case that we cannot find any suitable lemma in the set of word forms, we take the stem as a lemma. This case is quite rare.

Further improvement was reached by using morphological analysis in the case of the Czech language. As a typical example of the morphological analysis we could take the word "was" and its basic form "to be". Both words are typically indexed with different indexes in most lemmatization systems. After usage of such analysis, both forms are converted into one correct basic form "to be".

The input for the stemming module is a stream of words from the text. The output is a sequence of indexes referencing the corresponding EWN synsets. Each index consists of three parts: an abbreviation of the team which included the word in the thesaurus (e.g. eng20 – English team, EWN version 2.0), the unique synset number (e.g. 06900919) and the suffix describing the word type (e.g. noun, verb, adjective, etc.).

Example of the lemmatization
Input: The table has four legs.
Output: eng20-04209815-n eng20-02139918-v eng20-02111607-a eng20-0318657-n

In the case of the English language, lemmatization is relatively simple, where it is possible to use algorithmic lemmatization (Porter's algorithm). The big advantage of this solution is the use of proven parts (EWN, Ispell), which can be freely downloaded from the Internet. The results quality of lemmatization can be expected for other languages somewhere between the English language (a relatively simple language), and the Czech language (a language with a wide morphological diversity). However, the use of language specific processing (e.g. morphological analysis) is required in some cases.

2.4 Indexing

In our system the indexing task is solved by the lemmatization module described in the previous part of the article. Using EWN ILI (EWN inter-lingual-index) as the index, it's trivial to transform ILIs to indexes which are used for further processing (mainly classification in our case). Corresponding synsets have the same indexes in all languages. Thus the cross-language classification can be easily done and all documents can be stored in a language independent form. We can for example train the classifier on the English training set and test it on Czech examples and vice versa.

2.5 Disambiguation

Word sense disambiguation (WSD; [13]) is a necessary module in most of the natural language processing systems. It enables computers to understand the meaning of a text or a message. Let us take the automatic language translation as an example. E.g. the English word *bank* can be translated to Czech language as *banka* (financial institute) or *břeh* (border of a river). The correct translation depends on the context where the word is used and it is obvious that the translations are not interchangeable.

We have solved the disambiguation as a classification task where each meaning of the word is represented by one class. Possible meanings are typically presented in a dictionary which can also include some other information useable for disambiguation (synonymic relationships, word type, part of speech, etc.). Classification of the word into one of the possible meanings is determined by its context, eventually relevant information obtained from a dictionary, thesaurus, encyclopedia or other lexical sources. The thesaurus EWN is very often used as such information source.

Our text analysis discovered that nearly 20% of words are ambiguous. It shows the importance of disambiguation in all of the natural language processing (NLP) tasks.

We experimented with the supervised methods, mainly with Bayesian disambiguation. Some heuristic modifications have been tested with the aim to refine disambiguation accuracy. They include:

- Excluding from the context the words with low probability of occurrence.
- The context restriction by a sliding window and its modification.
- Inclusion of distances between the context words – ambiguous word into the Bayesian formula.
- Inclusion of syntactic relations into the disambiguation process.

When the supervised method is used we have to provide the training data set (usually as a tagged text) to train the disambiguator. Each occurrence of an ambiguous word *w* is tagged with a corresponding sense s_i^w (represented as the EWN index). We denote it as a semantic tag. Such an approach converts the task of disambiguation to the classification. The word *w* is classified into *k* classes, each representing one sense s_i , where *i* is the number from *1* to *k* and *k* is the number of possible classes. The Naive Bayesian disambiguation takes into account a set of words *c* (enclosing the ambiguous word) as an unordered set without any relationships among them. We have considered the absence of relationships and inability to change method's parameters quite limiting. That's why we introduced some heuristic modifications giving the better results.

Our primary aim is to improve the method to minimize flaws of Bayesian approach and find representative attributes in context window. With regard to the lack of the space we can not include all results and more detailed description into this text; they will be introduced in our presentation.

2.6 Categorization

The process of text document categorization is a challenging task, especially in case of multilingual documents. Our tested corpus includes Czech and English texts, in particular – press articles ČTK and Reuters, which have been used in many of our previous monolingual tests. The corpus consists of total of 82000 Czech and 25000 English articles is split into 5 classes – weather (5%), sport (30%), politics (58%), agriculture (3%), health (5%). The main goal was to investigate the influence of the multilingual environment on the results of classification, to compare the results with the monolingual ones and to verify the usability of EWN for multilingual indexing. We used various multilingual processing levels – processing with/without the use of EWN thesaurus, EWN applied on monolingual text corpus, EWN applied on multilingual text corpus and finally cross-language text classification (training on Czech corpus, testing on English and vice versa).

We have tested these classification methods: Naïve Bayes (NB) – see [5] for more information, NBCI – see [8], TFIDF – see [9] and [12], Itemsets – see [7],[10] and [11].

3 Results

3.1 Coding Recognition

Preliminary results for recognition of Czech and Slovak language stored in CP1250 or ISO-8859-2 coding are shown in the table. It shows the error percentage of recognition module for both methods.

	CZ ISO [%]	**CZ WIN [%]**	**SK ISO [%]**	**SK WIN [%]**
Stop list based	2,25	2,25	1,30	1,52
Frequency based	0,01	0,01	15,97	20,09

Tab. 2, Language recognition error rate

Because of similarities in both languages this example presents one of the most difficult cases of language and coding recognition. We expect better results for other combinations of languages.

3.2 Categorization

As we discovered through the text analysis, both parts of our multilingual corpus (Czech and English) differs each other. The articles from ČTK (Czech Press Agency) are much more general than those from Reuters. The Reuters articles are aimed at business and economical news, which are spread among all the classes. This fact leads to relatively bad results of Naïve Bayes method (see below).

Tests were done with the methods used before on monolingual corpuses. In all tests we have used various levels of multilingual preprocessing. Firstly we have done a reference setup (see tab. 3, row 1, 2, 3, 4; the column monolingual) – the articles were classified separately (as two monolingual corpuses).

In the next step EWN based lemmatization was applied on monolingual corpus (see tab. 3, row 1, 2, 3, 4; the column multilingual). We proved that the primary idea of using EWN as main solving tool is right. As you can see in the table 3, the results are close to the reference case and sometimes they are even better.

In the third step was classified the multilingual text corpus. We have considered that documents of both languages (Czech, English) are mixed in a random ratio (see tab. 3, rows 5, 6, 10, 11). That stands for the situation when some articles from various languages are already classified and we want to add another ones. This is easy achievable. In other words, the same word has different indexes in different languages and there are not any relationships between the languages (inter lingual indexes are not used). The categorization method gives results the same quality as the reference case does. The test was repeated with the use of classical lemmatization and EWN-based lemmatization.

The last tested case was directed to cross-language classification with the use of EWN and EWN inter lingual indexes (see tab. 3, row 7, 8, 9). This is the case when we have monolingual articles already classified and the new ones written in different language should be added. In other words the testing and training language differs. In this case the similarity of both corpuses seems to be very important. For example, NB method is unstable in this case and it leads to systematic error (as seen in the table 3).

The typical precision of our categorization oscillates in the range from 80 % to 95 %.

			monolingual		multilinguaul	
	Metod	**data**	**Precision [%]**	**Recall [%]**	**Precision [%]**	**Recall [%]**
1	**NBCI**	**cz**	90.53	93.83	91.28	93.41
2	**NB**	**cz**	95.36	95.57	92.44	93.15
3	**NBCI**	**eng**	95.11	95.47	96.04	96.20
4	**NB**	**eng**	96.85	96.91	94.79	95.17
5	**NBCI**	**cz+eng**	86.75	92.06	86.05	89.52
6	**NB**	**cz+eng**	95.25	95.46	92.04	92.83
7	**NBCI cross**	**cz+eng**	-	-	80.93	89.42
8	**NB cross**	**cz+eng**	-	-	3.42	3.42
9	**Itemsets cross**	**cz+eng**	-	-	73.78	81.49
10	**Itemsets**	**cz+eng**	75.76	81.91	78.65	84.90
11	**TFIDF**	**cz+eng**	93.37	93.37	92.79	92.79

Tab. 3, Classification results

3.3 Disambiguation

In the disambiguation tests we were looking for influence of size and shape of the sliding window, stop list and lemmatizer parameters.

	Precision [%]	**Training context window size**	**Testing context window size**	**Processing methods enabled (*)**
Reference case	87.55	5	5	-
Best context window (measure words)	95.56	5	12	SD
Best cont. window (measure sentence)	95.95	1 sentence	17	S
Best cont. window (measure paragraph)	91.22	1 paragraph	40	S
Window restricted by sentence	96.37	5	10	SD
Window restriction by paragraph	95.79	5	12	SD
Use of weight function	95.78	-	-	SD
Syntactic analysis	97.00	-	-	SD

Tab. 4, Disambiguation results

(*) S – stop-list used, D – dynamic sliding window used

For disambiguator training was used Semcor (Semantic concordance) corpus [14]. Words were lemmatized in the same way as in case of classification task. For lemmatization was used English dictionary of 119486 words and the stop list with 48 words.

As you can read in the table, the correct parameter setup is crucial in this task. We have gained the increase of precision about 1% when we restricted the context window on one sentence. The biggest precision gain was received using the syntactic analysis – approximately 2 % against the best plain context window setup.

4 Further work

A remarkable improvement of classification can be reached by synsets clustering. As we discovered during testing, the EWN structure is too fine. As a result the classifiers work with a big number of classes and a relatively small corpus which leads to precision decrease. Another issue related to EWN structure is that there are missing word's equivalents in some languages. This problem appears mainly in cross language classification using the NB method. We hope to eliminate this defect by clustering and assigning new indexes to the newly created clusters. This ensures a smaller number of classes and thereby an improved final precision. Clusters of synsets can be created, e.g. according to the context in which they appear.

Another planned improvement of described classification is its modification using weighted indexes for each word. It means the word will be characterized by a set of indexes. The weight of the index will be determined by the use of relationships contained in EWN.

5 Conclusion

Presented results have proved that the EWN can be used as a fundamental tool for multilingual text classification and methods based on EWN give promising results. Although precision is slightly (1 or 2%) lower in case of the multilingual environment, its value at over 90 % is quite remarkable. After application of synsets clustering we expect further improvement. In the near feature we intend to incorporate multilingual processing into the task of information retrieval as well.

The multilingual classification module will be used as a part of academic staff support system. That's the project that should help researchers to maintain the big amount of information available today. The module could be used to extend existing digital libraries. As a result it will be possible to process multilingual information sources as well.

References

1. http://www.global-reach.biz/globstats/index.php3 (20. 4. 2005)
2. *EuroWordNet*, http://www.illc.uva.nl/EuroWordNet/ (20. 4. 2005)
3. *Ispell*, http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html (20. 4. 2005)
4. Hajic, J., *Morfologický analyzátor*, http://quest.ms.mff.cuni.cz/pdt/Morphology_and_Tagging/Morphology/index.html
5. Manning, C. D.., Hinrich S. (2000), *Foundations of Statistical Natural Language Processing*, The MIT Press
6. Brown, P.F., S. A. Della Pietra, V. J. Della Pietra a R. L. Mercer (1991), *Word-Sense Disambiguation Using Statistical Methods*, Berkeley
7. Hynek J., Jezek K. (2000), *Document Classification Using Itemsets*. Proc. of conf. ISM 2000, ISBN: 80-85988-45-3, pp. 97-102
8. Kučera, M., Ježek, K., Hynek, J. (2004), *Text Categorization Method NBCI* (in Czech), Znalosti 2004, ISBN 80-248-0456-5
9. Joachims, T. (1997), *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*, International Conference on Machine Learning (ICML)
10. Hynek J., Ježek K (2001), *Automatic dokument classification using Itemsets Metod, its modification and evalution*. Sborník konference Datacon 2001 Brno, Mária Bieliková (Ed.), ISBN 80-227-1597-2
11. Hynek J., Jezek K.(2000) *Dokument Classification Using Itemsets,* Sbornik konference MOSIS 2000, ISBN: 80-85988-45-3
12. Joachims, T. (1997), *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization*, International Conference on Machine Learning (ICML), 1997
13. Veronis, J., Ide, N. (1998), *Word Sense Disambiguation, State of the Art*, Computional Linguistics 1998.
14. Cognitive Science Laboratory Princeton, http://www.cogsci.princeton.edu/~wn/doc/man/semcor.htm