

# EXTRACTION OF TRAINING SETS FOR EXPERIMENTATION WITH CROSS LANGUAGE INFORMATION RETRIEVAL SYSTEMS

Nikitas N. Karanikolas<sup>1</sup>, Christos Skourlas<sup>1</sup>, John Bratos<sup>2</sup>

<sup>1</sup> Department of Informatics, Technological Educational Institute of Athens, Ag. Spyridonos Street, 12210 Aegaleo, Greece  
[nnk@teiath.gr](mailto:nnk@teiath.gr), [cskourlas@hol.gr](mailto:cskourlas@hol.gr)

<sup>2</sup> Technological Educational Institute of Kalamata, 24100 Kalamata, Greece  
[john@teikal.gr](mailto:john@teikal.gr)

## Abstract

In this paper we focus on methods, models and tools for the extraction of bilingual training / test sets useful for the (semi) automatic classification of textual documents. Such documents could be tutorials, technical specifications, articles, personal notes, etc. Another motivation for our research is the need for managing corpus of classified texts and especially parallel corpora (texts). We discuss the usage of pre-selected key-phrases as attributes for classification, and methods for classifying new documents. These methods could be applied to training data and produce (infer) the corresponding models. We also describe and discuss the classification of various document (textual) types, which is supported by our prototype tool.

**Keywords:** document classification, cross language information retrieval, training set creation

## INTRODUCTION

Oard [2] classifies free text Cross Language Information Retrieval (CLIR) approaches in two broad categories:

- Knowledge-based systems are usually based on dictionaries, ontology and thesauri.
- Corpus based approaches are usually based on parallel, comparable and monolingual corpora.

Systems using Dictionaries are very popular and translate query terms one by one using all the possible senses of the term. The main drawbacks of such systems are:

- a) The lack of fully updated Machine Readable Dictionaries (MRDs),
- b) The ambiguity of the translations of terms [3].

Since the machine translation of the query is less accurate than that of a full text, experiments have been conducted with collections having machine translations of all the collection texts to all languages of interest. Such systems are multi-monolingual systems. Parallel and comparable corpora systems are different: the parallel (or comparable) corpora are used to “train” the system and after that no translations are used for retrieval [7]. The main problem with this approach is that it is not easy to find training parallel corpora related to any collection.

This research effort has been focused on methods, models and tools for the extraction of bilingual training / test sets useful for the (semi) automatic classification of textual documents. Such documents could be tutorials, technical specifications, articles, personal notes, etc. Special emphasis will be put on covering legal, insurance, banking, medical, financial and governmental publications.

Another interesting motivation for our research is the need for managing corpus of classified texts and especially parallel (original and translated) corpora (texts). We discuss the usage of pre-selected key-phrases as attributes for

classification, and methods for classifying new documents. These methods could be applied to training data and produce (infer) the corresponding models.

In the following section 2 we describe our methods for text classification and discuss the classification of various document (textual) types, which is supported by our tool. Section 3 describes the (medical) sources used for the evaluation of our system and section 4 illustrates the extraction of bilingual training sets and test sets. We also describe the features, the capabilities and a basic interface that could be used in such systems. Eventually, in the last section, conclusions and future directions are given.

## TEXT CLASSIFICATION

Text Classification could be defined as the application of (semi) automatic methods in order to choose, from a set of predefined classification codes, the appropriate one (category / class) for a given new document. As an example, patient discharge letters could be semi-automatically classified using some technique based on the selection of the appropriate ICD (International Classification of Diseases and Diagnoses) code [4].

Various studies have been focused on the construction of Rule or Tree based model and are related to the existence of key-phrases (terms) in order to assign the class of the unclassified document. Such an approach usually uses training sets of documents, already classified, and a predefined list of key-phrases (terms). Consequently, it creates a vector, for each document of the training set, that represents the existence or not of the predefined key-phrases in the document and their frequency. The last item of each vector is the class code (the label) of document. The following figure depicts the vector and its relationship with the list of  $m$  key-phrases (terms) and the list of available classes for a short collection of four documents.

Terms	Ctf	text1	text2	text3	text4
Term 1	ctf <sub>1</sub> =3	tf <sub>1 1</sub> =0	tf <sub>2 1</sub> =1	tf <sub>3 1</sub> =0	tf <sub>4 1</sub> =2
Term 2	ctf <sub>2</sub> =3	tf <sub>1 2</sub> =0	tf <sub>2 2</sub> =2	tf <sub>3 2</sub> =1	tf <sub>4 2</sub> =0
.....	.....	.....	.....	.....	.....
Term m	Ctf <sub>m</sub> =5	tf <sub>1 m</sub> =1	tf <sub>2 m</sub> =1	tf <sub>3 m</sub> =0	tf <sub>4 m</sub> =3
Classification Code		CC <sub>1</sub>	CC <sub>2</sub>	CC <sub>3</sub>	CC <sub>4</sub>

ctf=collection term frequency, tf=term frequency, cc=classification code

Having a set of labeled vectors, it is possible to apply some the Data Mining algorithm (eg [1], [6]) and construct the classification Rules or Trees. An example of such rule could be the following:

$$(A_{\lambda_1=v_{\lambda_1}}) \wedge (A_{\lambda_2=v_{\lambda_2}}) \wedge \dots \wedge (A_{\lambda_j=v_{\lambda_j}}) \supset (A_{m+1}=B)$$

where  $1 \leq \lambda_1 < \lambda_2 < \dots < \lambda_j \leq m$  for each attribute (key-phrase) A, m is the number of key-phrases in the authority list,  $v_i \in \{\text{true}, \text{false}\}$  and  $B \in \{b \mid b \text{ is any valid classification code}\}$ .

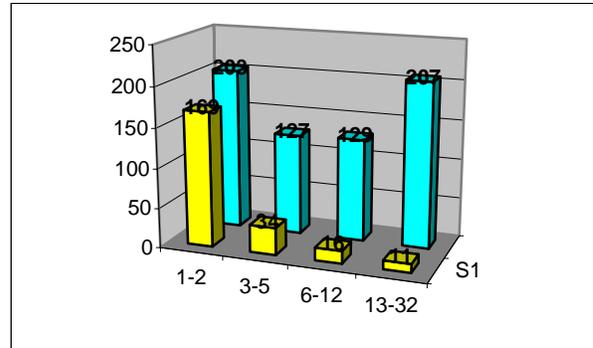
Another approach for the classification of documents incorporated into our tool is based on the similarity between existing documents (the “training” set) and the new (unclassified) documents (e.g. [5]). Such an Instance based learning method assumes that similar documents must be classified in the same category (class) or in other words must share the same classification code. In such a method the list of key-phrases is a promising set of attributes that can be used to describe and discriminate between existing and new documents.

#### SOURCES OF MEDICAL TEXT

Our initial attempt for the creation of a bilingual (Greek and English) collection of Medical texts was based on two sources:

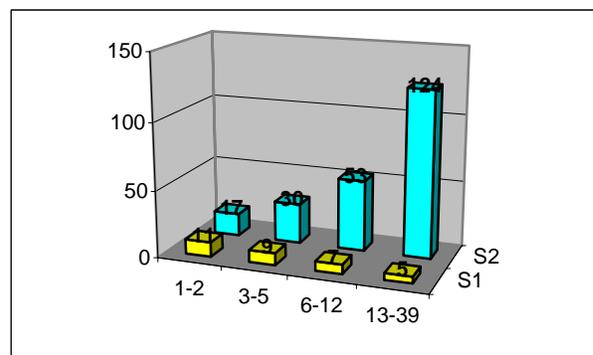
- 1) The Hospital Information System of Areteion University Hospital, in Athens, Greece, which maintains a complete Electronic Medical File (Electronic Patient Record) for the patients.
- 2) The bilingual bibliographic medical database of the Greek National Documentation Centre (NDC).

We selected Medical patient discharge letters, which have the appropriate size (are bigger than forty lines). These documents are classified into classes (diagnoses of the International system of diagnoses and diseases, ICD 9). Our documents’ collection contains 666 patient discharge letters that are classified in 230 ICD-9 classes. Fig 1 illustrates our collection. For example you can see that 34 classes occur and have 3 up to 5 documents. The total number of these documents is equal to 127. The “hidden” information is that there exist 14 classes with 3 documents, 15 classes with 4 documents and 5 classes with 5 documents ( $127 = 14 \times 3 + 15 \times 4 + 5 \times 5$ ).



The 169 classes that contain only one or two documents are not appropriate for the creation of pairs of training and test sets. Thus we are interested for 61 (34 + 16 + 11) classes extracted from the database of the Areteion University Hospital and the related 463 (127 + 129 + 207) documents.

The second source of data is the public bibliographic medical database of the Greek National Documentation Centre. We tried to choose documents that belong in the same classes of documents with documents extracted from the database of the Areteion University Hospital. Therefore, we used the Greek translations of the 61 selected ICD-9 classes and submitted them as questions (queries) to the bibliographic database. In some cases, we used synonyms. For example, we used both the phrases "malignant neoplasm of breast" and "cancer of breast". Doctors verified which of the retrieved documents are correctly classified in the class (ICD 9) which has been used as question (query). Eventually, we extracted 224 documents classified in 32 (of the 61) classes of interest in the case of the database of the Areteion University Hospital. Figure 2 presents this collection.

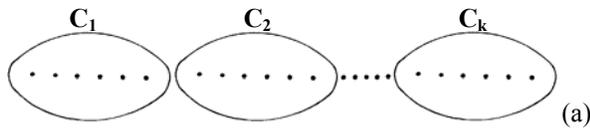


All the classes of texts from the medical database are useful since they increase the number of documents extracted from the database of Areteion. All the extracted and classified documents were stored in the database of our tool. The collection of bibliographic documents presents a higher complexity (summary and title in Greek and English, use of various bibliographic descriptors, etc)

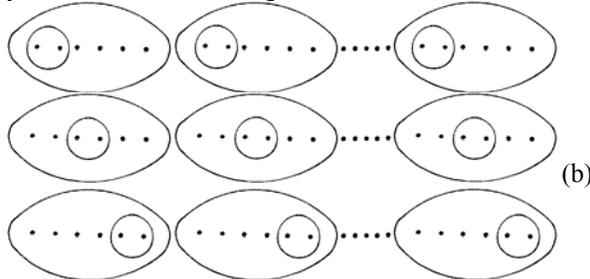
### TRAINING SET AND TEST SETS' CREATION

A simple combination of a training / test set could be extracted from a collection of documents that belong in known categories. We could construct those sets by choosing, randomly, certain texts to form the training subset and use all the other ones for the verification of the model (as a test set). But the random choice of texts for the training set has a potential disadvantage: It is possible to select texts for the training set that belong in a subset of classes (categories). Then, you may select texts, for the test set, that belong in the remaining classes. In this case the model is not representative.

Our tool supports the extraction of “representative” training/test set in the following way:  
We organize our collection of texts including  $k$  classes:

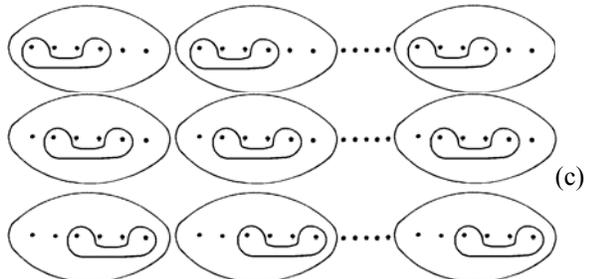


Then three pairs of Training / Test sets are constructed as you can see in the following schema:



Texts inside the circles represent the test set and the remaining ones constitute the training set.

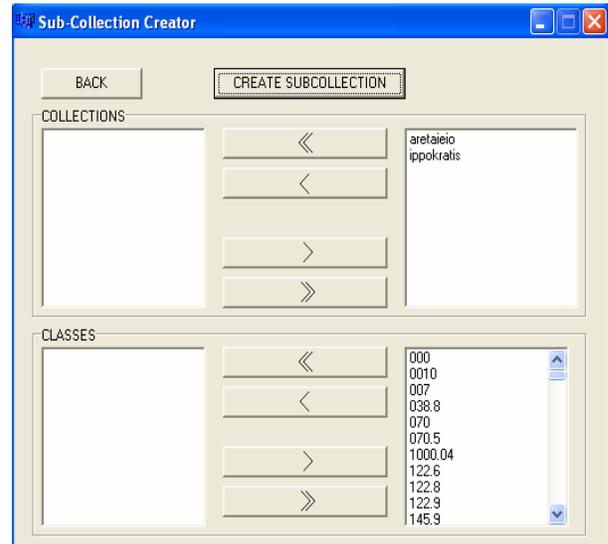
Thus the texts that constitute the test set of one of the alternative pairs become training texts in the second alternative pair of training / test set, etc. For still better dispersion of texts we can also create the three pairs of training / test sets as follows:



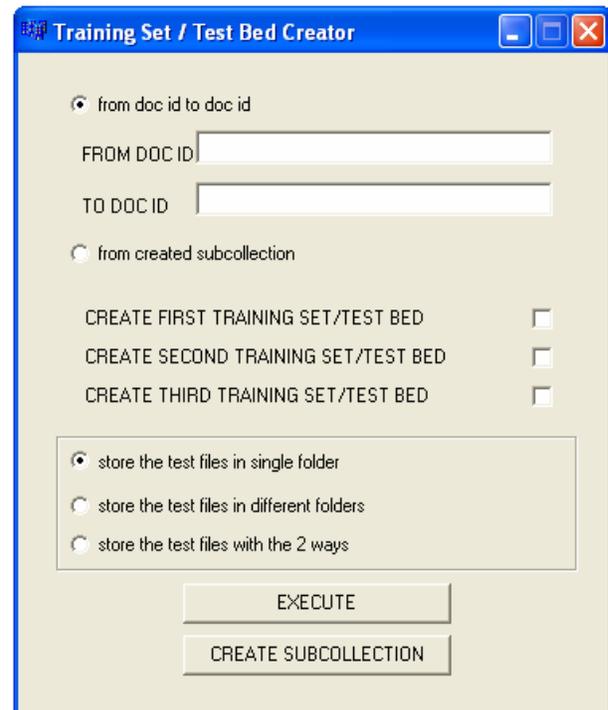
Where the texts inside the schema  constitute the test set whereas the rests constitute the training set.

### CONCLUSIONS AND FUTURE ACTIVITIES

The advantages of our system are distinguished in two main categories. It offers capabilities for the management of collections of classified documents, and also includes capabilities for the creation of training / test sets. The system can manage different collections of documents and is able to import/export all documents of a collection or import/export all the documents of a specific class. A simple interface that allows the capabilities mentioned above is illustrated in the following figure:



Another capability, as we have already mentioned, is the support of the creation of training / test sets. Our pilot system supports only three alternative pairs of training / test sets as you can see in the following figure.



We already work in the development of a new system that will cover the needs of a *trilingual – greek / english / french – test collection*. It is been built, mainly, for *CLIR experimentation*.

#### ACKNOWLEDGMENTS

This Project is co-funded by the European Social Fund and National Resources – (EPEAEK-II)-ARXIMHDHS (Archimedes). We would like to thank our student Chrissovalantis Tsavdaroglou for the first implementation of the tool.

#### REFERENCES

- [1] Ahonen H. et. al, *Mining in the phrasal frontier*, in Proc. Principles of Knowledge Discovery in Databases Conference, Trondheim, Norway, June 1997. Lecture Notes in Computer Science, Springer Verlag, 1997.
- [2] Oard, D. W., *Alternative Approaches for Cross-Language Text Retrieval*, in Cross-Language Text and Speech Retrieval, AAAI Technical Report SS-97-05. Available at <http://www.clis.umd.edu/dlrg/filter/sss/papers/>
- [3] Davis, M., *New experiments in cross-language text retrieval at NMSU's Computing Research Lab*, in D. K. Harman, ed., *The Fifth Text Retrieval Conference (TREC-5)*, NIST, 1996.
- [4] Karanikolas N. N. and Skourlas C., *Automatic Diagnosis Classification of patient discharge letters*, in MIE'2002: XVIIth International Congress of the European Federation for Medical Informatics, Budapest, Hungary, August, 2002. IOS Press, ISBN: 1-58603-279-8, ISSN: 0926-9630.
- [5] Karanikolas N. N., Skourlas C., Christopoulou A. and Alevizos T., *Medical Text Classification based on Text Retrieval techniques*, in MEDINF'2003: 1st International Conference on Medical Informatics & Engineering, Craiova, Romania, 2003. *Craiova Medicala*, volume 5, supplement 3, 375-378.
- [6] Karanikolas N. N. and Skourlas C., *Naive Rule Induction for Text Classification based on Key-Phrases*, in 6th International Conference on Data Mining, Text Mining and their Business Applications, Skiathos, Greece, 2005. *WIT Transactions on Information and Communication Technologies*, volume 35, *Data Mining VI: Data Mining, Text Mining and their Business Applications*, WIT Press, ISBN 1-84564-017-9, ISSN 1764-4463.
- [7] Marinagi K., Alevizos T., Kaburlazos V. and Skourlas C., *Fuzzy Interval Number (FIN) techniques for Cross Language Information Retrieval*, in ICEIS'2006: 8th International Conference on Enterprise Information Systems, Paphos, , Cyprus, 2006.