Patent classification experiments with the Linguistic Classification System LCS in CLEF-IP 2011

Suzan Verberne and Eva D'hondt

Information Foraging Lab
Radboud University Nijmegen
s.verberne@cs.ru.nl

Abstract. We report the results of a series of classification experiments with the Linguistic Classification System LCS in the context of CLEF-IP 2011. We participated in the main classification task: classifying documents on the subclass level. We investigated (1) the use of different sections (abstract, description, metadata) from the patent documents; (2) adding dependency triples to the bag-of-words representation; (3) adding the WIPO corpus to the EPO training data; (4) the use of patent citations in the test data for reranking the classes; and (5) the threshold on the class scores for class selection.
We found that adding full descriptions to abstracts gives a clear improvement; the first 400 words of the description also improves classification but to a lesser degree. Adding metadata (applicants, inventors en address) did not improve classification. Adding dependency triples to words gives a much higher recall at the cost of a lower precision but this effect is largely due to the class selection threshold. We did not find an effect from adding the WIPO corpus, nor from reranking with patent citations. In future work, we plan to investigate whether there are other methods for reranking with patent citations that does give an improvement, because we feel that the citations may still give valuable information.
Our most important finding however is the importance of the threshold on the class selection. For the current work, we only compared two values for the threshold and the results are much better for 1.0 than for 0.5. The 0.5 threshold gives higher recall in all runs, which was the original motivation for submitting runs with a lower threshold. However, because the much lower precision, the F-scores are lower. We think that there is still some improvement to be gained from proper tuning of the class selection threshold, and the use of a flexible threshold (also taking into account the different text representations). This is part of our future work.

1 Introduction

In this paper, we describe the classification experiments that we conducted in the context of the Intellectual Property (IP) track at CLEF 2011 (CLEF-IP[1]). In 2009, the track was organized for the first time with a prior art retrieval task. In 2010, a classification task was added to the track. In 2011, this task was continued and extented with a new optional sub-task, which is to classify a given patent document up to the subgroup level, when the subclass is given. We only participated in the main classification task: classifying documents on the subclass level.

The goal of the classification task at CLEF-IP is to classify a given patent document, according to the International Patent Classification system (IPC). For the purpose of the track, the organization released a collection of 2.6 million patent documents from the European Patent Office (EPO), extended with 400,000 documents from the World Intellectual Property Organization (WIPO). These 3 Million documents with content in English, German and French pertain to over 1 Million patents.[2] From the collection, 1,000 documents (the 'topics') per language were held out

[1] http://www.ir-facility.org/clef-ip

[2] A patent is the name for a group of patent documents that relate to the same invention; they have the same patent ID number.

as test set. The remainder of the corpus constitutes the target data, on which participants could develop their methods.

In this notebook paper, we describe our classification experiments with the Linguistic Classification System LCS. We only performed mono-lingual classification, training and evaluating our models on English texts only. We evaluate a number of classification variables: (1) the use of different patent sections, (2) adding dependency triples to the bag-of-words representation, (3) expanding the EPO training corpus with WIPO documents, (4) using patent citations to rerank the selected classes, and (5) tuning the threshold on class selection.

In Section 2, we describe the data selection, data preparation and the classification settings used. The results from the classification experiments are presented in Section 3, followed by our conclusions in Section 4.

2 Classification experiments with LCS

For our classification experiments, we used the Linguistic Classification System (LCS)[3] [2, 3]. The LCS can perform both mono-classification (each document is assigned exactly one class label) and multi-classification. In the training phase, the LCS takes as input a file which list the paths to the classification files followed by their classes. After this training phase the LCS can be used for testing the trained classifier on a test collection of documents with known classes (usually held-out training data), or for producing a classification of new documents without known classes.

Three classifiers have been implemented in the LCS: Naive Bayes, Winnow and SVM^{light}. Last year [6], we experimented with both Winnow and SVM^{light} and we found that their classification accuracy scores are comparable but that SVM^{light} is much slower. Therefore, we decided to use Winnow for this year's CLEF-IP experiments. Winnow has a number of parameters that can be tuned: α, β and *maxiters* (the number of training iterations). Based on the tuning we did last year, we decided to use $\alpha = 1.02$, $\beta = 0.98$ and $maxiters = 10$.

In our classification experiments, we compared the following experimental settings:

1. The use of different sections (abstract, description, metadata) from the patent documents;
2. The use of different document representations for classification, adding dependency triples to the bag-of-words representation.
3. The training corpus selection: EPO only, or EPO and WIPO together;
4. The use of patent citations in the test patents for reranking the assigned classes;
5. The threshold on the class scores for class selection.

We will explain how we prepared the experiments for each of these comparisons in the following subsections.

2.1 Corpus preparation: extracting IPC classes and sections

From all patents in the target data, we extracted the information needed for classification: the IPC-R classes, the textual content from the English abstract and description; and applicants, inventors en address as additional metadata. For each patent, we selected the most recent version which contains all the information needed.[4] Table 1 shows the size of the training corpus when particular patent sections are included. We allowed the abstract to be empty if either the description or the metadata sections contains content. As a result, the subcorpus 'abstract and metadata' is the largest: 1,3M documents, some of which only contain metadata.
We separately extracted the first 400 words of the description because the experiences from other participants in last year's workshop [5] taught us that the head of the description is a good alternative to the complete description, which may be too heavy to classify due to its length. We conducted experiments to validate this assumption.

[3] A demo of the application can be found at http://ir-facility.net/news/linguistic-classification-system-prototype/ for registered IRF members.

[4] E.g. in the corpus directory EP/000000/00/59/01/, EP-0005901-A3.xml is newer than EP-0005901-A2.xml and both are newer than EP-0005901-B1.xml.

Table 1. Number of EPO documents in training corpus when particular patent sections are included.

Sections	Number of docs
abstracts	855,261
abstracts and metadata	1,325,364
abstracts and description	648,441
abstracts, description and metadata	649,557

2.2 Different document representations: adding triples to words

In CLEF-IP 2010, we experimented with the addition of dependency triples to the bag-of-words representation, which is generally used in text classification. The results on the 2010 test set were mixed [6] but in follow-up experiments [3], we consistently found a significant improvement in F-score when we added dependency triples to the word-based representation of patent abstracts.

This year, we again investigated the improvement that can be gained from adding dependency triples to the bag of words, but we did not limit ourselves to classification of abstracts. We parsed the abstracts and the first 400 words of the descriptions with the AEGIR dependency parser [4, 7] version 1.8.2. AEGIR's output representation is comparable to the Stanford typed dependencies representation [1], in the sense that it generates a set of binary relations between words for an input sentence, thereby converting some function words (such as prepositions) to relations. In addition to that, AEGIR performs a number of normalizing transformations, such as passive-to-active transformation. For example, the clause "an inflammatory reaction, caused by the bowel tissue" leads to the same analysis as "the bowel tissue causes an inflammatory reaction". An example of the triple representation can be found in Figure 1 below [6].

Original text	words	triples
Heat is stored	heat is stored	[IT,SUBJ,store] [store,OBJ,heat]
at a steady	at a steady	[store,PREPat,temperature]
temperature using	temperature using	[temperature,ATTR,steady]
calcium chloride	calcium chloride	[temperature,DET,a] [chloride,ATTR,calcium]
hexahydrate and	hexahydrate and	[hexahydrate,ATTR,chloride]
up to 20 percent	up to percent	[hexahydrate,ATTR,using] [up,PREPto,20 percent]
strontium chloride	strontium chloride	[assist,OBJ,crystallization]
hexahydrate	hexahydrate	[chloride,ATTR,strontium]
to assist	to assist	[hexahydrate,ATTR,chloride]
crystallisation.	crystallisation	[hexahydrate,SUBJ,assist]

Fig. 1. Part of the original text from the abstract of document EP-0011358-A1.txt (left) and the two document representations that we created: words and triples. The classification experiments are performed on either the words representation, or words together with triples.

2.3 Training corpus selection: adding WIPO data to EPO data

In text classification, system performance usually goes up when the size of the training set increases. While the CLEF-IP test set only consisted of documents from the EPO corpus, we investigated if adding documents from another corpus, namely the WIPO, to the EPO training set led to improvements in classification accuracy. We added the WIPO corpus to two of our section subcorpora: *abstracts and description*, and *abstracts, description and metadata*. Table 2 shows the resulting document counts for the training corpora. From the table it is clear that in the WIPO corpus, there are fewer documents with the metadata fields applicants, inventors en address than in the EPO corpus.

Table 2. Number of EPO and WIPO documents in training corpus when particular patent sections are included.

Sections	No. of EPO docs	No. of WIPO docs	Total
abstracts and description	648,441	257,017	905,458
abstracts, description and metadata	649,557	16,270	665,827

2.4 The use of patent citations for reranking the classes

Some of the patent files (topics) in the test set contain citations to other EPO patents. We used these citations to rerank the LCS output using the following procedure:

1. For each topic, we extracted the patents that are cited by the topic (labelled as *patcit* in the XML file);
2. We looked up each of the citations in the training corpus and extracted their IPC-R classes. We found that 562 of the 1,000 topics contains at least one cited patent with one or more IPC-R classes.
3. These 'citation classes' get a vote each time they occur in a cited patent. A vote is worth 1.0 in addition to the LCS score.

For example, in one of the experiments, LCS selected five classes for the topic EP-1223323-A2, and assigned them the following scores:

EP-1223323-A2	F01N	4.67
EP-1223323-A2	F02D	2.22
EP-1223323-A2	B60W	1.90
EP-1223323-A2	B60K	1.55
EP-1223323-A2	F02N	1.00

Of these, F01N (1x), B60K (2x) and B60W (2x) occur in the citations of EP-1223323-A2. Their classification score is increased by the number of times they occur in the citations, and the list of classes is re-ranked:

EP-1223323-A2	F01N	5.67
EP-1223323-A2	B60W	3.90
EP-1223323-A2	B60K	3.55
EP-1223323-A2	F02D	2.22
EP-1223323-A2	F02N	1.00

2.5 The threshold on the class scores for class selection

In the case of multi-classification, LCS is flexible with respect to the number of classes that are returned per document. Internally, it produces a full ranking of classes for each document in the test set. The user can regulate the selection of classes with three parameters: (1) a threshold that puts a lower bound on the classification score for a class to be selected, (2) the maximum number of classes selected per document ('maxranks') and (3) the minimum number of classes selected per document ('minranks'). In the experiments on the target data, we kept the selection threshold to 1.0 (which is the default). Based on the average number of classes per document in the target data (2.7 according to [6]), we decided to set $maxranks = 4$. Setting $minranks = 1$ assures that each document is assigned at least one class, even if all classes have a score below the threshold.

In the submitted runs on the test data, we decided to lower the class selection threshold to 0.5 because the value of 1.0 gives an average of 1.8 classes per test document; setting it at 0.5 gives an average of 3.2 classes. The latter seemed wiser for a recall-oriented task. Also, we increased maxranks to 5. In additional experiments, we evaluated the results for a threshold of 1.0 against the results for the threshold of 0.5.

Table 3. Classification results (Precision, Recall and F1) according to *trec_eval* 9.0 for all runs on test set (1,000 topics), sorted by F1. For each measure, the best-scoring setting is printed in boldface.

Run	Sections	Text representation	Corpus	Rerank with citations?	Class selection threshold	P	R	F1
ad400WT	abs, desc400	words+triples	EPO	no	0.5	0.2995	**0.8657**	0.4206
ad400WTcit	abs, desc400	words+triples	EPO	yes	0.5	0.2995	**0.8657**	0.4206
ad400WT1	abs, desc400	words+triples	EPO	no	1.0	0.3957	0.8457	0.4932
aWT	abs	words+triples	EPO	no	0.5	0.4522	0.7778	0.5275
amWTcit	abs, meta	words+triples	EPO	yes	0.5	0.4441	0.8039	0.5311
amWT	abs, meta	words+triples	EPO	no	0.5	0.4485	0.7984	0.5313
aW	abs	words	EPO	no	0.5	0.4764	0.7644	0.5412
aWcit	abs	words	EPO	yes	0.5	0.474	0.7728	0.543
amW	abs, meta	words	EPO	no	0.5	0.4744	0.7828	0.5441
ad400W	abs, desc400	words	EPO	no	0.5	0.5197	0.8431	0.5981
admWOW	abs, desc, meta	words	EPO+WIPO	no	0.5	0.5359	0.8531	0.6118
admWOWcit	abs, desc, meta	words	EPO+WIPO	yes	0.5	0.5321	0.8625	0.6131
admWcit	abs, desc, meta	words	EPO	yes	0.5	0.5379	0.8563	0.6168
admW	abs, desc, meta	words	EPO	no	0.5	0.5436	0.8506	0.6186
adW	abs, desc	words	EPO	no	0.5	0.5518	0.8459	0.6231
aW1	abs	words	EPO	no	1.0	0.6893	0.6435	0.6235
adWcit	abs, desc	words	EPO	yes	0.5	0.5485	0.8555	0.6249
adWOW	abs, desc	words	EPO+WIPO	no	0.5	0.553	0.8505	0.6252
adWOWcit	abs, desc	words	EPO+WIPO	yes	0.5	0.5489	0.8583	0.6254
amW1	abs, meta	words	EPO	no	1.0	0.6993	0.6472	0.6302
aWT1	abs	words+triples	EPO	no	1.0	0.7068	0.6627	0.6425
amWT1	abs, meta	words+triples	EPO	no	1.0	0.7173	0.6726	0.6513
admWOW1	abs, desc, meta	words	EPO+WIPO	no	1.0	0.7198	0.7492	0.6890
admWOWcit1	abs, desc, meta	words	EPO+WIPO	yes	1.0	0.7056	0.7744	0.6925
ad400W1	abs, desc400	words	EPO	no	1.0	0.7416	0.7328	0.6926
admW1	abs, desc, meta	words	EPO	no	1.0	0.7352	0.7419	0.6932
adW1	abs, desc	words	EPO	no	1.0	0.7374	0.737	0.6939
adWcit1	abs, desc	words	EPO	yes	1.0	0.7229	0.7649	0.6991
adWOW1	abs, desc	words	EPO+WIPO	no	1.0	**0.7443**	0.7501	0.7025
adWOWcit1	abs, desc	words	EPO+WIPO	yes	1.0	0.7265	0.7754	**0.7059**

3 Results

For training the classification models, we used the target data with the exception of the 2000 most recent documents in the training corpus, which we used as test set in the development stage. A complete overview of the results on the real test data (the 1,000 topics provided by the track organization) is shown in Table 3. As opposed to last year, when we measured standard deviations over multiple runs of the same experiment, we only performed each experiment once this year. Our results on the 2010 data showed that standard deviations are small and even small differences in the results tend to be significant because of the large data set [3].

Figures 2–6 at the end of the paper show the effects of different sections, text representation, corpus selection, patent citations and class selection threshold respectively (the five experimental variables that we compare).

Figure 2 shows that adding the description to the abstract gives a clear improvement in classification accuracy: from 0.54 to 0.62 in F-score. The effect of adding the first 400 words of the description instead of the complete description, is smaller, giving an F-score of 0.60. Surprisingly, adding metadata (applicants, inventors en address) to the abstracts and descriptions does not give any improvement. This is in contrast with last year's results, when some participants reported significant improvement from adding applicants, inventors en address as metadata [5].

Figure 3 shows that adding dependency triples to the bag-of-words representation has an effect but whether this is a positive effect highly depends on the evaluation measure used. Recall is higher for the words+triples representation but this comes at the cost of a much lower precision. The experimental setting with the *lowest* F-score of all, ad400WT, has the *highest* recall of all runs (0.87). We had a look at the full ranking of the classes and found that for the runs with triples, the class scores are generally higher. This means that more classes get a score above the fixed threshold of 0.5 (in fact, the average number of classes selected per patent for ad400WT is 5.0, which is the maximum number of selected classes). As a result, recall is higher and precision is lower.

Figure 4 shows that there is no effect of adding the WIPO documents to the EPO training corpus. More data generally gives better classification results, but in this task and using this data, increasing the number of documents from 650K to 905K did not generate any effect.

Figure 5 shows that the use of patent citations in the test data for reranking the classes has no visible effect either. We plan to investigate whether there are other methods for reranking with patent citations that does give an improvement, because we feel that the citations may still give valuable information.

Figure 6 shows that the threshold on the class scores for class selection is highly important for the evaluation scores. For the current work, we only compared two values for the threshold, 0.5 and 1.0, and it is clearly visible that the results are much better for 1.0 than for 0.5. The 0.5 threshold gives higher recall in all runs, which was the original motivation for submitting runs with a lower threshold. However, because the much lower precision, the F-scores are lower. The default LCS threshold of 1.0 clearly is the better choice here. We think that there is still some improvement to be gained from proper tuning of the class selection threshold, and the use of a flexible threshold (also taking into account the different text representations). This is part of our future work.

4 Conclusion

We reported the results of a series of classification experiments in the context of CLEF-IP 2011. We investigated (1) the use of different sections (abstract, description, metadata) from the patent documents; (2) adding dependency triples to the bag-of-words representation; (3) adding the WIPO corpus to the EPO training data; (4) the use of patent citations in the test data for reranking the classes; and (5) the threshold on the class scores for class selection.

We found that adding full descriptions to abstracts gives a clear improvement; the first 400 words of the description also improves classification but to a lesser degree. Adding metadata (applicants, inventors en address) did not improve classification. Adding dependency triples to words gives a much higher recall at the cost of a lower precision but this effect is largely due to the class selection threshold. We did not find an effect from adding the WIPO corpus, nor from reranking with patent citations. Our most important finding is the importance of the threshold on the class selection. Our future work will be directed at tuning this threshold.

References

1. M.C. De Marneffe and C.D. Manning. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.
2. C.H.A. Koster, M. Seutter, and J. Beney. Multi-classification of patent applications with Winnow. *Lecture Notes in Computer Science*, pages 545–554, 2003.
3. Cornelis H. A. Koster, Jean G. Beney, Suzan Verberne, and Merijn Vogel. Phrase-Based Document Categorization. In W. Bruce Croft, Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, editors, *Current Challenges in Patent Information Retrieval*, volume 29 of *The Kluwer International Series on Information Retrieval*, pages 263–286. Springer Berlin Heidelberg, 2011.

4. Nelleke Oostdijk, Suzan Verberne, and Cornelis H.A. Koster. Constructing a broad coverage lexicon for text mining in the patent domain. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), 2010.
5. Florina Piroi and John Tait. CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain. In *CLEF 2010 LABs and Workshops Notebook Papers*, 2010.
6. S. Verberne, M. Vogel, and E. Dhondt. Patent classification experiments with the Linguistic Classification System LCS. In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2010), CLEF-IP workshop*, 2010.
7. Suzan Verberne, Eva D'hondt, Nelleke Oostdijk, and Cornelis H.A. Koster. Quantifying the Challenges in Parsing Patent Claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval (AsPIRe 2010)*, pages 14–21, 2010.

Fig. 2. The effect of different patent sections. Words only, EPO corpus, no reranking, threshold=0.5.

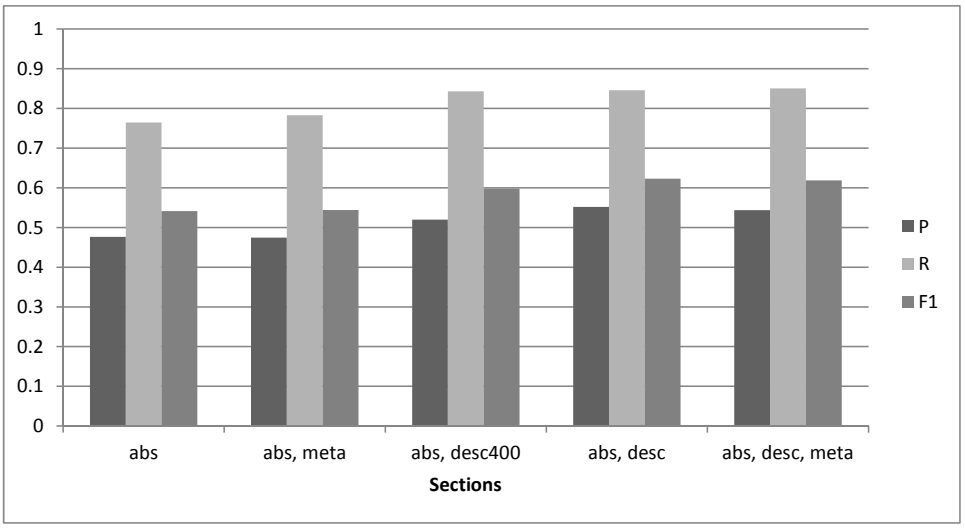

Fig. 3. The effect of adding dependency triples to the bag of words. EPO corpus, no reranking, threshold=0.5.

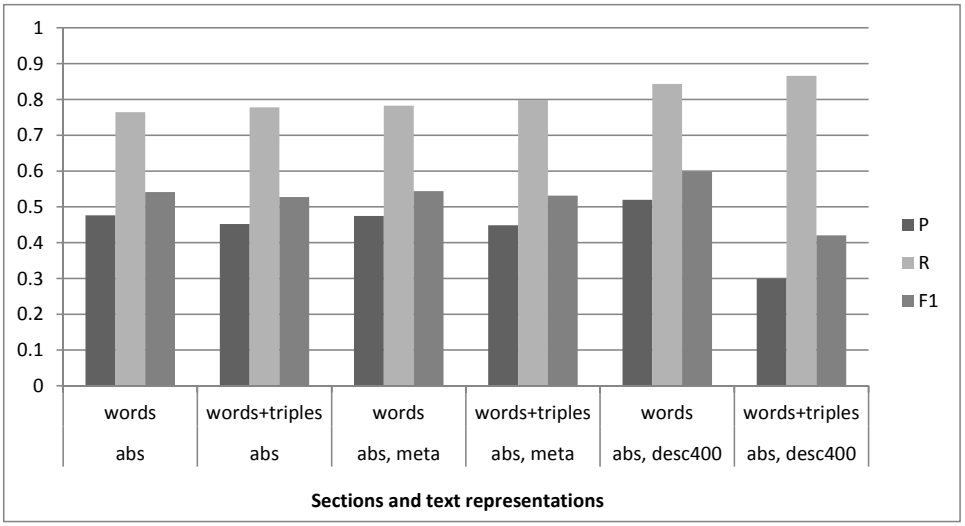

Fig. 4. The effect of adding the WIPO corpus to the EPO training set. Words only, no reranking, threshold=0.5.

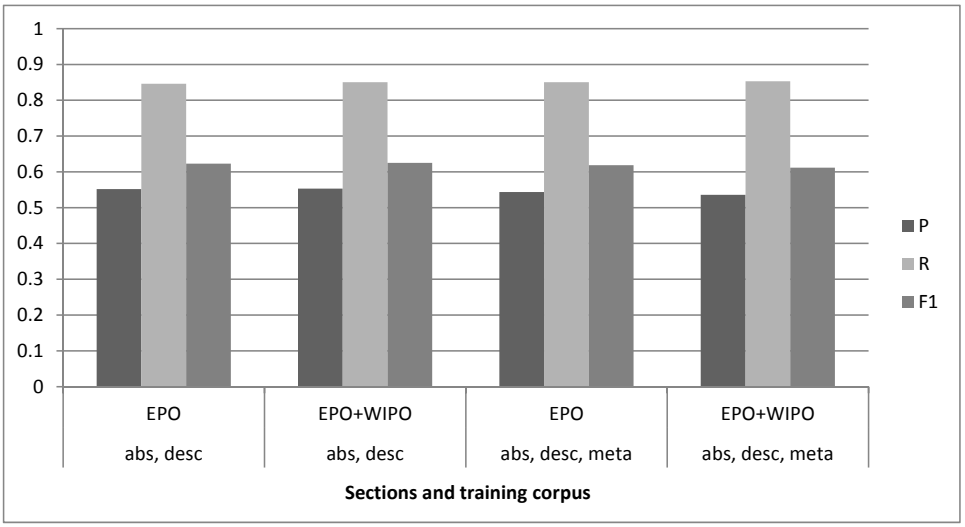

Fig. 5. The effect of reranking with patent citations. Words only, EPO corpus, threshold=0.5

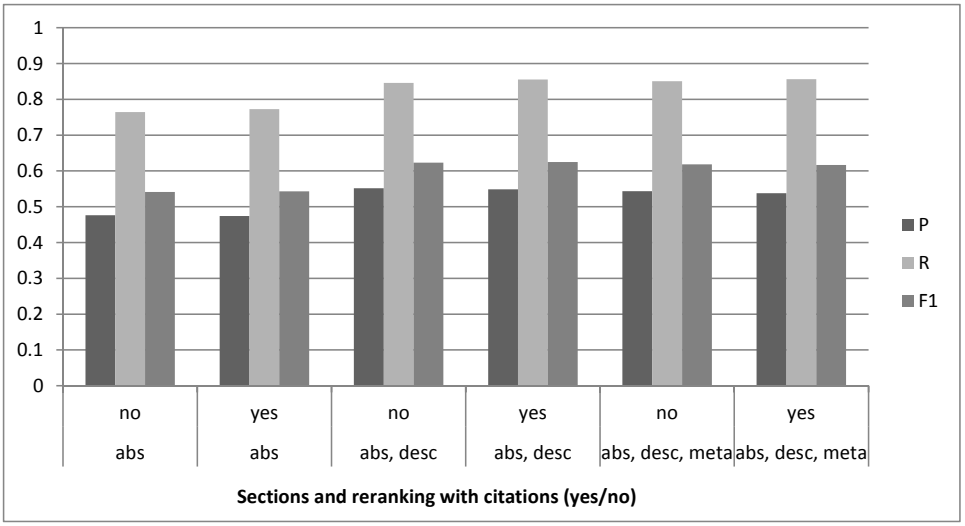

Fig. 6. The effect of changing the threshold on the class scores for class selection (1.0 or 0.5). Words only, EPO corpus, no reranking