Passage retrieval in legal texts

Paolo Rosso, Santiago Correa, Davide Buscaldi

Passage Retrieval in Legal Texts

Paolo Rosso[a], Santiago Correa[a], Davide Buscaldi[b]

[a] *Natural Language Engineering Lab, ELiRF, Universidad Politécnica de Valencia, Spain*
[b] *Laboratoire d'Informatique Fondamentale d'Orléans, LIFO, Université d'Orléans, France*

Abstract

Legal texts usually comprise many kinds of texts, such as contracts, patents and treaties. These texts usually include a huge quantity of unstructured information written in natural language. Thanks to automatic analysis and Information Retrieval (IR) techniques, it is possible to filter out information that is not relevant and, therefore, to reduce the amount of documents that users need to browse to find the information they are looking for. In this paper we adapted the *JIRS* passage retrieval system to work with three kinds of legal texts: treaties, patents and contracts, studying the issues related with the processing of this kind of information. In particular, we studied how a passage retrieval system might be linked up to automated analysis based on logic and algebraic programming for the detection of conflicts in contracts. In our set-up, a contract is translated into formal clauses, which are analysed by means of a model checking tool; then, the passage retrieval system is used to extract conflicting sentences from the original contract text.

Key words:
Passage retrieval, legal texts, treaties, patents, contracts.

1. Introduction

Nowadays, there is an increasing interest in the automatic processing of legal texts, such as contracts, treaties or patents, from both the academic and business sectors. Legal or law informatics [1], as well as forensic linguistics[1], are some of the interdisciplinary fields interested in the automatic processing of these kinds of legal texts [2] with the aim of solving legal cases with the help of a computer-assisted tool, mining relevant information from contracts, or retrieving the relevant prior art of a patent[2] as, for instance, in the research

Email addresses: prosso@dsic.upv.es (Paolo Rosso), santcg@gmail.com (Santiago Correa), buscaldi@univ-orleans.fr (Davide Buscaldi)
URL: http://www.dsic.upv.es/grupos/nle (NLE Lab)
URL: http://www.univ-orleans.fr/lifo

[1] http://www.iafl.org/

[2] Prior art (previously registered patents) and not state-of-the-art is employed in intellectual property and patent retrieval [3]

works carried out by Chieze et al. [4] and Francesconi et al. [5], or in the research work of Sarkar et al. on term burstiness [6]. Special attention needs also to be paid to legal contracts (e-contracts, social contracts, deontic contracts as well as security policies or protocols) due to the increasing interest in the study of the contract language, contract analysis, reasoning [7] and conflict discovery [8].

Legal and patent retrieval tracks have been recently introduced in the context of international competitions such as TREC[3], CLEF[4], and NTCIR[5]. Patents themselves can be considered legal texts written in a very specific jargon. There is a big interest worldwide in patent analysis and more specifically in patent retrieval and patent mining whose aim is to look for hidden information in order to create technical trend maps from a set of patents and also to search for possible cases of plagiarism of ideas [9] (conflict discovery in patents).

The above competitions are commonly considered as the reference for the last developments in *Information Retrieval* (*IR*). IR is the task of finding relevant documents given a question[6] (e.g. *When was the Lisbon Treaty signed?*) or a query[7] (e.g. *Anti-lock braking system*) which expresses a user's needs. *Passage Retrieval* (*PR*) is a specific kind of IR where the task consists in finding those portions (passages) of documents that are relevant. PR helps to filter out the information that is not relevant because it reduces the original document collection to a set of passages in which the user information needs are satisfied.

In this paper, we report on three experiments in which we have employed the Passage Retrieval system JIRS to retrieve relevant passages from three types of legal texts: treaties, patents and contracts. Contracts express potential obligations, permissions and prohibitions of different actors and can be used to protect the interest of the organisations engaged in services exchanged [10]. Due to the potentially dynamic composition of services with different contacts, as well as the combination of service contracts with local contracts, unexpected conflicts could arise. Therefore, there is a need for automatic techniques dedicated to contract analysis in order to make sure that such contracts (i.e., mechanisms that protect enterprises and organisations giving restrictions on service behaviour) are conflict-free, meaning that the contracts will never lead to conflicting or contradictory directives.

In Section 2 we briefly describe the characteristics of passage retrieval and specifically the simple language-independent *JIRS* system. Section 3 and Section 4 are devoted to the description of how *JIRS* has been adapted, respectively, for legal text and for patent retrieval tracks. Section 5 illustrates a case study of an airline check-in desk presented in [11], and how we used our passage retrieval system to find conflicting sentences in a given contract. Finally, in Section 6 we draw some conclusions and we discuss further work.

[3] http://trec-legal.umiacs.umd.edu/

[4] Cross Language Evaluation Forum (http://www.clef-campaign.org).

[5] http://www.ls.info.hiroshima-cu.ac.jp/~nanba/ntcir-8/cfp.html

[6] question: linguistic expression used to make a request for information.

[7] query: a form of questioning, a precise request for information retrieval within information systems.

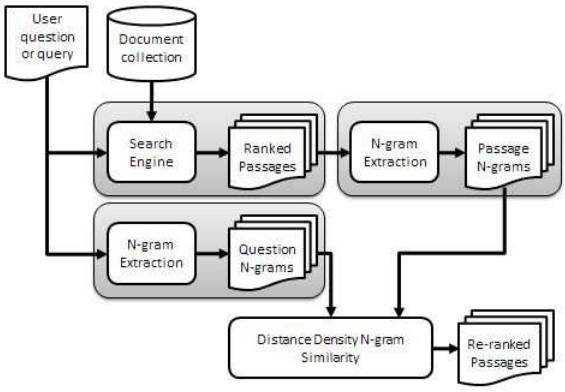

Figure 1: Architecture of *JIRS* Passage Retrieval system

2. Methodology of the Employed System

Various methods to determine the similarity between passages and a question or a query have been proposed. Among them, Tellex et al. [12] consider that the two most representative approaches are those where the similarity depends on: (i) the overlap between the question or the query and the passage terms (i.e., a large overlap means a great likeness); (ii) the density of the question or the query terms in the passage (i.e., a passage is more similar to the question if its terms are narrowly distributed). The comparison of the *PR* systems used in the *TREC*[8] competition [12] shows that the best *PR* systems are those that are based on density, with the methods of IBM [13] and SiteQ [14] being the most effective ones. The *PR* systems that participated in the *CLEF* competition can also be split into these two categories, even if most of them were based on term overlap [15, 16, 17, 18] rather than density [19, 20].

It must be noted that some systems do not fall completely into one of these two categories. Among them the most interesting ones are: (i) the system described by Adriani and Rinawati in [21] which takes into account not only the density but also the order of appearance of the question or query terms; (ii) the system illustrated by Sutcliffe et al. in [22] which combines the term overlap with question reformulation; (iii) the system of Tanev et al. [23] which calculates the similarity at the syntactical level by using edit distance between trees; (iv) the system of Hartrump [24] which transforms both the question and the passage into a semantic representation in order to find their similarity; (v) the system of Costa [25] which uses logical-syntactical patterns to determine the relevant passages.

JIRS[9] (Java Information Retrieval System) is a passage retrieval system which is based on a density n-gram distance model, where an n-gram is a

[8]Text Retrieval Conference (http://trec.nist.gov/).
[9]http://sourceforge.net/projects/jirs/

sequence of n words. For example, if we have the following sequence of words: *"anti-lock braking system"*, we can say that this sequence is a trigram but we can also say that is composed of the *"anti-lock"* 1-gram and the *"braking system"* 2-gram. The general architecture of *JIRS* is shown in Figure 1. The user's query is given to a search engine (e.g. *Yahoo*) [26] that will search a document collection to return snippets in which relevant terms from the question (or the query) occur. The n-gram extraction module will return all the k-grams of size $1 \leq k \leq n$, where n is the number of terms in the question (or query) q. This process is done for both q and each of the snippets retrieved by the search engine. Once the n-grams are obtained from the question (or query) and the snippets, a comparison is made to calculate a similarity value between them. This similarity value is used to sort the list of passages that will eventually be returned to the user. The similarity between the question (or query) and the retrieved passages is defined in Equation 1. The weighing scheme detailed here corresponds to the base JIRS scheme which has been used throughout our work.

$$Sim(p,q) = \frac{\sum\limits_{\forall x \in Q} h(x,P) \frac{1}{d(x,x_{max})}}{\sum_{i=1}^{n} w_i} \tag{1}$$

Where $Sim(p,q)$ is the function that measures the similarity of n-grams sets of the question (or the query) q with respect to the n-grams sets of the passage p. P is the n-gram set of the heaviest passage p (i.e., the one with most weight) whose terms are in the question (or the query); Q is the set of j-grams that are generated from the question q and n is the total number of terms in the question (or the query). There are three special and particular term functions:

- w_i, is the weight of the i-th term of the question (or the query) which is determined by:

$$w_i = 1 - \frac{log(n_i)}{1 + log(N)} \tag{2}$$

  Where n_i is the number of sentences in which the term t_i occurs and N is the number of sentences in the collection;

- the function $h(x,P)$ measures the weight of each n-gram and is defined as:

$$h(x,P_j) = \begin{cases} \sum_{k=1}^{j} w_k & \text{if } x \in P_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

  Where w_k is the weight of the *k-th* term (see Equation 2) and j is the number of terms that compose the analyzed n-gram;

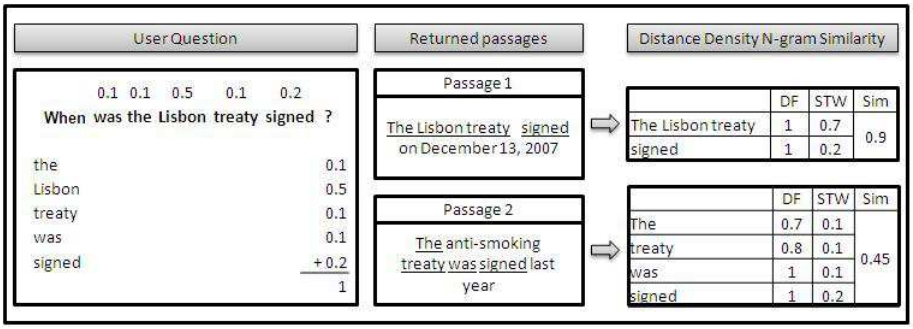

Figure 2: Density N-gram distance example (DF: Distance Factor, STW: Sum of Terms Weight, Sim: Similarity)

- and the factor $\frac{1}{d(x,x_{max})}$ that is a distance factor which reduces the weight of the n-grams that are far from the heaviest n-gram. The function $d(x, x_{max})$ determines numerically the value of the separation according to the number of words between a n-gram and the heaviest one. That function is defined as show in Equation 4 :

$$d(x, x_{max}) = 1 + k \cdot\ ln(1 + L) \tag{4}$$

  Where k is a factor that determines the importance of the distance in the similarity calculation and L is the number of words between a n-gram and the heaviest one (see Equation 3).

To better understand the passage weighing scheme of *JIRS*, let us suppose we have a collection of documents that refer to legal treaties of the European Union. Given the following question: *When was the Lisbon Treaty signed?* *JIRS PR* system returns the two following passages:

1. *The Lisbon Treaty was signed on December 13, 2007*
2. *The anti-smoking treaty was signed last year*

Figure 2 shows the details of the ranking calculated by *JIRS*. In order to calculate the Distance Factor of each n-gram we set a k factor equal to 0.3.

Buscaldi et al. [27] showed that JIRS is usually able to obtain a better answer coverage in the Question Answering task than other traditional PR models based on Vector Space Model, such as *Lucene*[10]. A comparison between these two systems is showed in Figure 3. Answer coverage is calculated as follows:

$$ac_P(Q) = \frac{\sum_{q \in Q} c(q, P)}{|Q|} \tag{5}$$

[10]http://lucene.apache.org/

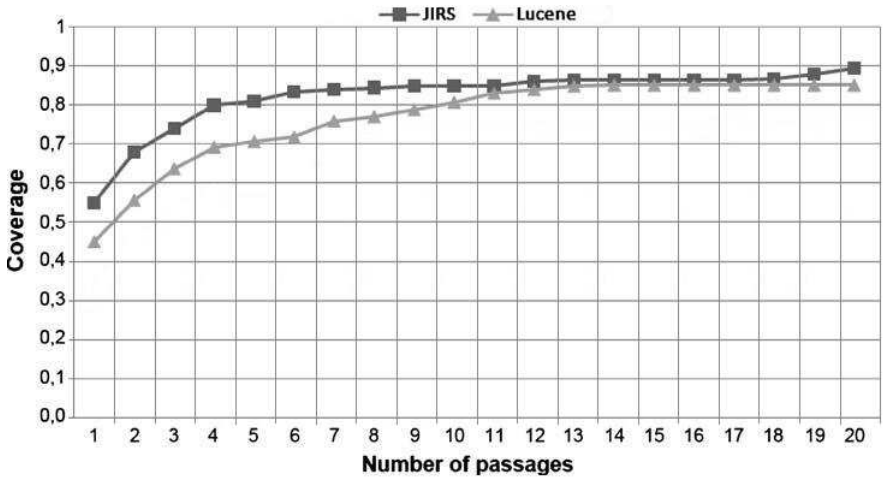

Figure 3: Comparision of answer coverage obtained with up to 20 passages ($1 \leq |P| \leq 20$), using the CLEF-2006 test set (target document set extracted from news collections): *JIRS* vs. *Lucene*

Where Q is the query (question) test set, P the set of passages p retrieved by the system (which size $|P|$ is fixed before the search), and $c(q, P)$ is the query-by-query coverage calculated as:

$$c(q, P) = \begin{cases} 1 & \text{if } \exists p \in P\text{: } p \text{ is relevant for } q \\ 0 & \text{elsewhere} \end{cases} \quad (6)$$

In the experiments related to Figure 3, the relevancy criterion consisted in p containing the right answer for question q.

An important difference between the Vector Space Model used in Lucene and the density n-gram distance model used in JIRS is that the density model privileges structure matching over frequency weights: for instance, let us suppose that given the query *"anti-lock braking system"*, the the two retrieved passages are: *"... braking system consist of disk brakes ..."* and *"... anti-lock braking system developed by ..."*. A standard system such *Lucene* would give most weight to the first passage due to the repetition of words with the same stem (*"brake"*). *JIRS* would instead give most weight to the more relevant second passage due to the presense of the trigram *"anti-lock braking system"*.

As we illustrated in Figure 1, *JIRS* makes use of a search engine in order to retrieve the relevant passages which are re-ranked at the end on the basis of its density-distance model. In order to appreciate the improvement that is obtained using the density-distance model on top of a search engine, we illustrate in Figure 4 the answer coverage of *JIRS* which exceeds by 19% the one obtained with *Yahoo* [26].

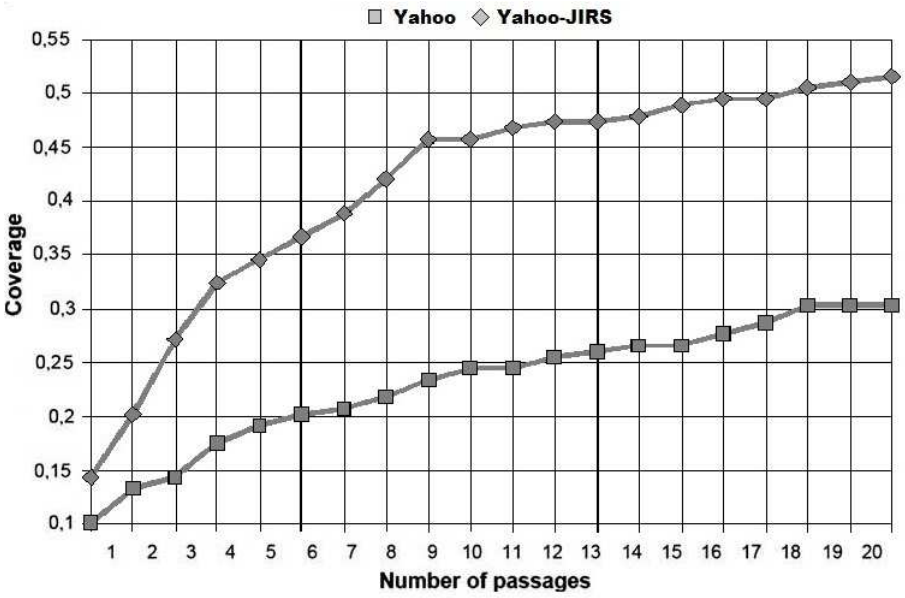

Figure 4: Comparision of answer coverage obtained with up to 20 passages ($1 \leq |P| \leq 20$), using CLEF-2006 questions with the web as collection: *JIRS* vs. *Yahoo*

3. Experiment 1: Passage Retrieval in Treaties

Many data sets consisting of legal texts are currently available. In many cases they are composed of parallel texts which are written in several languages. Examples of such kind of data sets are: the Canadian Hansards data set[11], the Europarl data set[12] and the *JRC-ACQUIS* data set[13]. In this section we describe the last data set, the way it was employed in the RespubliQA@CLEF-2009 track and our passage retrieval based approach we used for the participation in the competition.

3.1. The JRC-ACQUIS data set

The *JRC-ACQUIS* data set is composed of the total body of European Union (EU) law applicable in the the EU Member States. The collection includes articles written since 1950 and is updated constantly, due to the different languages present in the European Union. The data set is provided in 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian,

[11]For more information about the Canadian Hansards data set, refer to page: http://www.isi.edu/natural-language/download/hansard/

[12]For more information about Europarl data set, refer to page: http://www.statmt.org/europarl/

[13]For more information about the JRC-ACQUIS data set, refer to page: http://langtech.jrc.it/JRC-Acquis.html

Total of aligned documents (all languages)	4,350,447
Total of links (all languages)	243,187,303
Average of aligned documents per language	18,833
Average of links per language pair (average of all language pair)	1,052,759

Table 1: *JRC-ACQUIS* data set statistics, extracted from: http://wt.jrc.it/lt/Acquis/

Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene and Swedish. Besides, the data set also provides about $23,000$ documents per language, constituting one of the biggest parallel data sets in existence. The diversity of languages, the amount of information and the possibility of free download, make the *JRC-ACQUIS* data set a major source of data for the study of topics such as machine translation and information retrieval. Some statistics related to the data set can be seen in Table 1.

3.2. The CLEF ResPubliQA competition

The *ResPubliQA@CLEF-2009*[14] competition addresses the problem of question answering in legal texts. Given a pool of 500 independent natural language questions, each system must return the passage which answers each question from the *JRC-ACQUIS* collection of *EU* documentation where both questions and documents are translated and aligned for a subset of languages. An example of a question and a possible answer given by the organizers of the competition is as follows:

Question: *In how many languages is the Official Journal of the Community published?*

Passage: *The Official Journal of the Community shall be published in the four official languages.*

Two tasks were proposed by the organizers: monolingual and bilingual. In the monolingual task, questions, target document collection and answers are formulated in the same language, whereas in the bilingual task, questions are formulated in a different language than the target document collection and the corresponding answers. 11 groups participated in the competition with 12 runs submitted in English, 6 in Spanish, 4 in Romanian, 3 in French, 2 in German and 1 in Italian. Our group participated, submitting 5 runs, all of them for monolingual tasks: 1 in English, 2 in Spanish, 1 in French and 1 in Italian [28].

[14]For more information about the ResPubliQA@CLEF-2009 competition, refer to page: http://celct.isti.cnr.it/ResPubliQA/

The systems that participated in the *ResPubliQA* competition have used different approaches. The best system, which was developed by Rodrigo et al. [29] relied on an *IR* phase focused on improving *QA* results, a validation step for removing unpromising paragraphs and a module based on n-grams overlapping for selecting the final answer. The second best system [30] was developed on the basis of three main modules: question analysis, mainly concerned with the identification of the semantic type of the entity sought by the question (in addition it also identifies the question focus, the question type and the set of keywords relevant for the question); passage retrieval, employing a conventional *IR* search engine (in this case *Lucene*) to select a set of relevant candidate passages from the text collection; and the third module, answer extraction and ranking, where the representation of the question and the representation of the candidate answer-bearing passages are compared and a set of candidate answers is returned, ranked according to the likelihood that hey constitute the correct answer. This estimation is based on three specific answer types: answering questions asking about *Named Entities*, answering questions looking for *generic* answers and answering *definition* questions. The approach of the third best system [31] was implemented as workflow which was built on several web services: question preprocessing, query generation, search engine querying and paragraph ranking. This system is a trainable system that uses a linear combination of paragraph relevance scores to obtain a global relevance (to the question) measure which is used as the sort key. The system was trained on a specific parallel data set, but its functionality is independent of the linguistic register of the training data.

The two measures employed in the *ResPubliQA* competition [28] to assess the performance of the systems were:

- *Accuracy*: the number of right answers respect to the total number of questions:

$$accuracy = \frac{AR}{N} \tag{7}$$

- and *c@1* which rewards those systems capable of identifying wrong answers by not answering some question.

$$c@1 = \frac{1}{N}(AR + Unans\frac{AR}{N}) \tag{8}$$

  Where
  AR: is the number of correctly answered questions
  $Unans$: is the number of unanswered questions
  N: is the total number of questions

```
<TEI.2 id="" n="" lang="">
  <teiHeader lang="" date.created="">
    <fileDesc>
      <titleStmt>
        <title>Title of document</title>
      </titleStmt>
      <extent>ZZZ paragraph segments</extent>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <head n="1">Paragraph 1</head>
      <div type="body">
        <p n="2">Paragraph 2</p>
        <p n="3">Paragraph 3</p>
        ...
          <p n="m">Paragraph m</p>
      </div>
    </body>
  </text>
</TEI.2>
```

Figure 5: XML format used in JRC-ACQUIS corpus

3.3. Our Passage Retrieval based approach

Our approach consists in the use of the *JIRS* passage retrieval system, (which is based on redundancy), with the assumption that it is possible to find the answer to a question in a large enough document collection. The retrieved passages are ranked depending on the number, length and position of the question n-grams structures found in the passages. The first part of the process involves the collection of documents indexed by the *JIRS* system; for this purpose documents in XML format, must be analysed and processed. *JIRS* uses passages as basic indexing unit (i.e., passages are extracted from the documents). Due to the characteristics in which the body is provided, in each document we take the paragraph included between <p> tags as a passage to index. Therefore, each paragraph is labelled with the name of the document which contains it and its paragraph number. In Figure 5 we illustrate an example of the XML format used in the *JRC-ACQUIS* data set.

Once the collection is indexed by *JIRS*, for each question, the system returns a list of passages, according to the *JIRS* weighting scheme, which probably contain the answer to the question. In Table 2 we show some results obtained in the *ResPubliQA* competition by our *JIRS*-based NLEL-MAAT system [32].

Table 3 shows the results obtained by our *JIRS*-based system. We participated in 4 monolingual tasks (Task): English-English (en-en), French-French (fr-fr), Italian-Italian (it-it), Spanish-Spanish (es-es). Moreover, in another experiment (es-es2) we used the parallel data set to look for the answer in one of the 4 languages of the monolingual tasks (English, Spanish, Italian and French). As best answer we chose the one retrieved by *JIRS* with the highest score. On the basis of the identifier of the retrieved passage, the aligned passage in the target language (Spanish) was returned. The architecture of the multilingual *JIRS*-based system with which we participated in *ResPubliQA* is illustrated in

SIM	Q: In which countries can yards and miles be used?
0.83	P: within the Community , distances may be expressed in yards or miles in the United Kingdom or Ireland if those countries so desire
0.54	P: Member States in which vehicle speed is , at the time of adoption of this Directive , measured in miles per hour , shall be permitted to require speedometer equipment fitted to vehicles sold in their countries to be marked both in kilometres per hour and in miles per hour , until such time as their national legislation is amended to require only the use of metric (SI) units of measurement in accordance with the provisions of Council Directive No 71/354/EEC (1) of 18 October 1971 on the approximation of the laws of the Member States relating to units of measurement , as amended by the Treaty of Accession (2)

Table 2: Example of relevant passages returned by *JIRS*; SIM: Similarity measure according to *JIRS* weighting scheme; Q: Question formulated by the competition

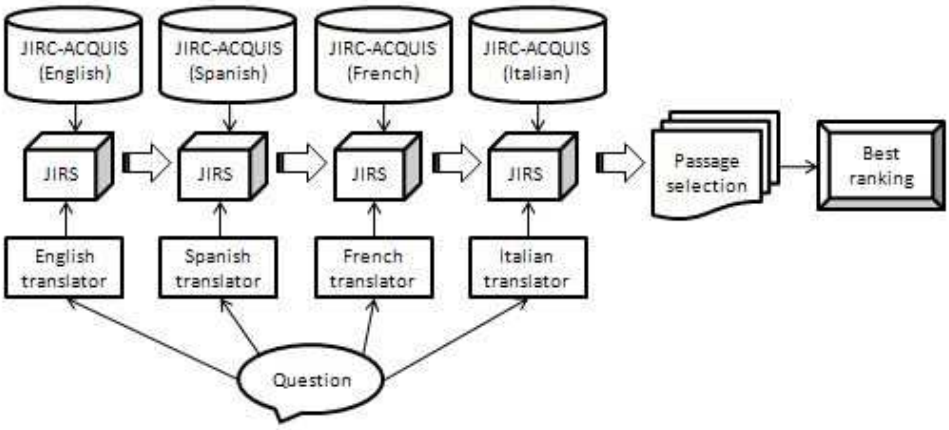

Figure 6: Architecture of NLEL-MAAT multilingual system (*ResPubliQA* competition)

Task	Ans.	Unans.	A.R.	A.W.	Accuracy	c@1
en-en	498	2	286	212	0.57	0.57
fr-fr	488	11	171	317	0.35	0.35
es-es	495	5	171	324	0.35	0.35
it-it	493	7	253	240	0.51	0.51
es-es2	466	34	218	248	0.44	0.47

Table 3: Results for submitted runs at ResPubliQA@CLEF-2009 competition; Ans.: Answered, Unans.: Unanswered, A.R.: Answered Rigth, A.W.: Answered Wrong, Accuracy: Accuracy measure, c@1: c@1 measure.

Language	JIRS-based	Best Other	Official Baseline
en	0.57	0.61	0.53
it	0.51	N/A	0.42
es	0.47	0.41	0.40
fr	0.35	0.28	0.45

Table 4: c@1 measure obtained by our system, compared to the c@1 obtained by the best system from among the other RespubliQA2009 participants and the official baseline provided by organisers.

Figure 6.

With a low complexity passage retrieval system such as the *JIRS*-based one we succeeded in obtaining very good results, better than the majority of some resource-depending systems participating in the competition. Proof of this is the best rating in three of the four tasks we participated in (see Table 4 and the official track overview [28]). The only task in which the NLEL-MAAT system was not ranked as first among participants was the monolingual *en-en* task. The system was also able to perform better than the baseline (obtained using a standard IR system with $tf \cdot idf$ weighing scheme) in all languages with the exception of French. Thanks to the characteristics of (partial) language independence of *JIRS* we obtained very good results in all languages. The use of the parallel data set (es-es2 task) allowed us to improve the results of the Spanish monolingual task (es-es). We plan to further investigate the possibility of employing this approach also for the other languages.

4. Experiment 2: Passage Retrieval in Patents

Before applying for a patent, inventors must ensure, through a search, that the invention meets the novelty requirement and does not conflict with its prior art. A search is performed by patent examiners in order to determine its legitimacy. This is the main application of passage retrieval in patents. In this section we describe the *European Patent Office* (*EPO*) data set, the *Intellectual Property* (*IP*) competition and, finally, our *PR*-based approach.

4.1. The European Patent Office data set

The European Patent Office[15] (*EPO*) data set[16] is composed of patents written since 1978. The whole collection consists of approximately 1.6 million individual patents. *EPO* patents are written in one of the three official languages (English, German and French) the documents are provided in *XML* format. Patent documents are structured documents consisting of four major sections:

[15] http://www.epo.org/

[16] For more information about the EPO data set, refer to page: http://www.epo.org/patents/law/legal-texts.html

Patent number: *EP1127544*

Title: *Measurement relating to human energy metabolism*

Description: *A heart rate measurement arrangement chargement comprises a calculating unit comprising a mathematical model arranged to form a person's energy metabolism level as an output parameter of the model unsing as input parameters of the model one or more heart rate parameters and one or more physicological parameters each describing a physiological characteristic of the person...*

Figure 7: Text passage of a patent

Documents published between	1985 and 2000
Published documents	1,958,955
Percentage of documents in English as main language	69
Percentage of documents in German as main language	23
Percentage of documents in French as main language	8

Table 5: Sub-collection EPO data set statistics (Patent documents: 1,022,388)

(i) Bibliographic data
(ii) Abstract
(iii) Description
(iv) Claims

A portion of a text file regarding a patent is shown in Figure 7:

4.2. The CLEF Intellectual Property competition

The Intellectual Property IP@CLEF-2009 task is coordinated by *Information Retrieval Facility*[17] (IRF) and *Matrixware*[18] and aims to find the prior art of a given set of patents. The basic task is to find the prior art of a set of 500 patents of a data set provided by the organizers. The statistics of this data set are shown in Table 5.

In the IP@CLEF-2009 competition, the best systems employed retrieval models (*Kullback Leibler* [33], *Okapi* [34]) as well as regression models [35, 36, 37]. Most of the systems participated in the competition despite the high complexity and obtained poor results [3]. This indicates the difficulty of the *IP* competition. There were four measures employed in the *IP* competition to assess the performance of the systems:

[17] http://www.ir-facility.org/the_irf/
[18] http://www.matrixware.com/

- *Precision* (P) is the number of relevant documents retrieved n_s over the total number of documents N_s returned by the search:

$$P = \frac{n_s}{N_s} \tag{9}$$

- *Mean Average Precision* (MAP): *Precision* is a single-value metric based on the whole list of documents returned by the system. For systems that return a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. MAP emphasizes ranking relevant documents higher. It is the average of precisions computed at the point of each of the relevant documents in the ranked sequence.

- *Recall* (R) measures the number n_s of relevant documents retrieved over the total number of relevant documents N_{db} on the data set:

$$R = \frac{n_s}{N_{db}} \tag{10}$$

- and $nDCG$ measure is defined as:

$$nDCG = M \sum_i \frac{(2^{r(i)} - 1)}{log(1+i)} \tag{11}$$

  where $r(i)$ is the relevance of document in rank position i and M is the normalizing constant chosen in order to have the score always between 0 and 1 [38, 39].

4.3. Our Passage Retrieval based approach

Our hypothesis is that the similarity between a patent candidate and another patent contained in the data set gives clues about the possibility that the latter is part of the prior art of a patent candidate [40] . In order to use *JIRS* for the *IP* task, it is necessary to clean the labels belonging to the *XML* format and filter the most important information of each patent in the data set provided to properly index the collection. In addition, for each patent it is possible to find several versions of it but, because they have an identification number that makes each patent unique, it is possible to eliminate the repeated information. Once we have finished this preprocessing, we obtain a sufficiently small portion of the data set to be indexed by the *JIRS PR* system. To search for the prior art of a patent, we construct a sequence of words that describe the content of each of the 500 patents to analyse. In order to do so, we decided to consider the title and the abstract to extract the relevant terms, employing the *Random Walks* summarisation technique [41] on the description of each patent. The *Random Walks* algorithm is inspired by graph theory, where terms are vertices and their co-occurrence within a certain distance (usually 2 words) is represented by an edge between them. The in-degree is used to assign weights to terms, according

to its contribution to the context of the document. This summarisation technique is used to extract the most relevant n-grams. In order to illustrate the method, let us consider the patent in Figure 7.

Patent title: *Measurement relating to human energy metabolism*

After applying the summary technique in the description we obtain the following relevant terms:

Relevant terms: *The heart rate*

The title and the concatenated information about the relevant terms is given to *JIRS* as a sequence of words. For the previous example we obtain the following word sequence:

Words sequence: *Measurement relating to human energy metabolism, the heart rate*

Another problem which we faced was the multilingual nature of the competition. As we described in Section 4.1, patents may be submitted in three different languages: English, French and German. Therefore, we opted for using three passage retrieval systems, one for each language. Thus we had to translate each of the 500 patents, using the google translate tool[19], in order to analyse them in the three languages and then look for the prior art for each patent in each language. The three ranking lists were merged into one in order to obtain the most similar 1,000 patents [40]. The architecture of the multilingual *JIRS*-based system we participated in the *IP* competition is illustrated in Figure 8.

Given the previous patent, our *JIRS*-based *PR* system retrieved as prior art of the patent:

Patent number: *EP1103216*

Title: *Method device measuring blood pressure heart rate environment extreme levels noise vibrations*

Description: *A method device measuring systolic diastolic blood pressure heart rate environment comprising extreme levels noise vibrations disclosed Blood pressure signals heart beat detected acoustic sensor patient artery...*

Patent number: *EP0785748*

Title: *Method and device for determining threshold values for energy metabolism*

[19] http://translate.google.com/

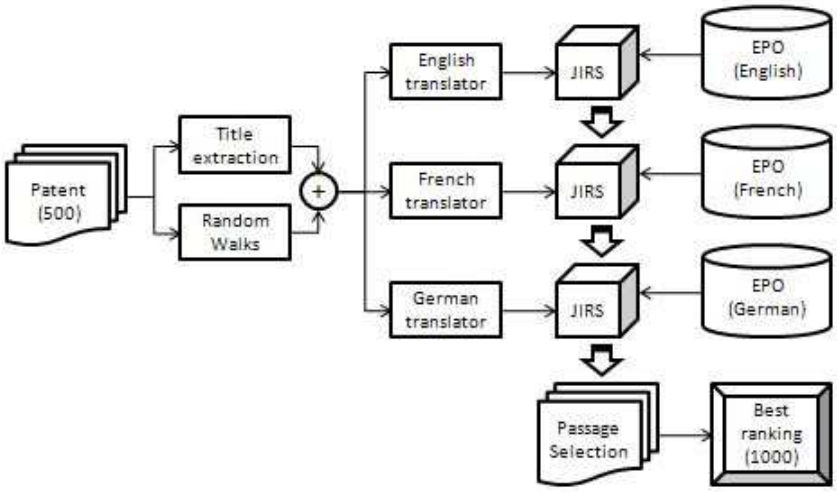

Figure 8: Architecture of NLEL-MAAT multilingual system (*IP* competition)

P	R	MAP	nDCG
0.0016	0.2547	0.0289	0.3377

Table 6: Results at IP@CLEF-2009 competition; P: Precision, R: Recall, nDCG: normalized Discounted Cumulative Gain, MAP: Mean Average Precision

Description: *The invention relates method device determining threshold values energy metabolism method testee subjected gradually increasing stress order obtain threshold values energy metabolism In DE-3439238 disclosed conventional heart rate monitor comprising chest-worn pulse transmitter wrist-worn receiver adapted wirelessly receive pulse signals transmitter...*

In Table 6 we show the results obtained by our *JIRS*-based system. As previously stated, in general most of the results obtained by the participants were also low (Figure 9), due to the complexity of the *IP* task. We have to emphasize that with an approach as simple as the one we have proposed, we have obtained results were not too far from the ones obtained by the best systems. From a practical viewpoint, our aim was to apply the simple *JIRS*-based system in order to filter out information not relevant with respect to the prior art of a patent. This allowed us to sensibly reduce the size of the data set for further investigation, eventually employing a more formal approach.

The poor performance of the proposed approach is partially due to the fact that a certain quantity of data was not used in the various stages of processing: in particular, in the document indexing phase, a lot of proprietary information was not used because of the criterion used to reduce the amount of data needed to represent the document. The criterion consisted in using the text contained

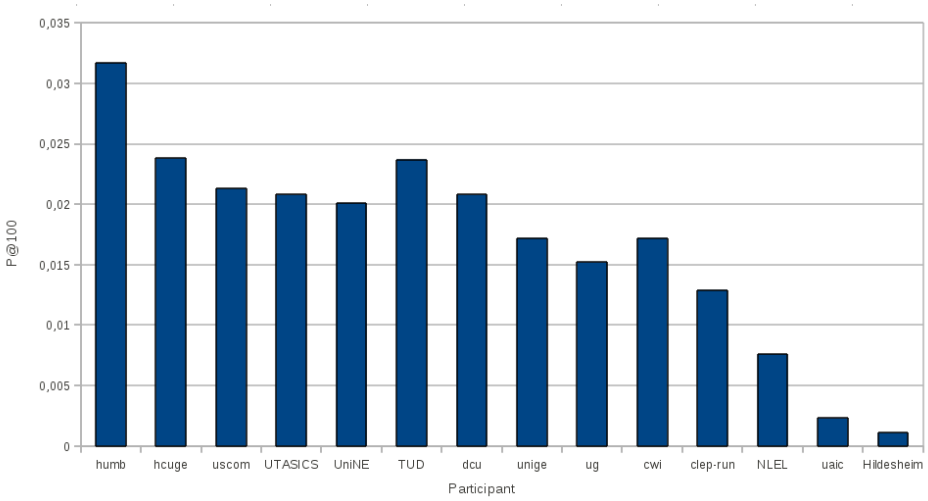

Figure 9: Comparative results (Precision @ 100) of the systems participating in the CLEF-IP 2009 competition. Data from the official track overview [3].

in the "title" and "abstract" fields for the generation of the sentence (query) for each input patent. Therefore, a substantial amount of information contained in the remaining fields of patents was lost. A text mining approach applied to these data would have allowed summarization of this information in an effective way in order to produce the query associated to the input patent. We have to remark that the need to reduce the size of the data arose from the performance limitations, as the maximum quantity of textual data that can be processed, of JIRS, and not from the method design.

Another factor responsible for the results was that *JIRS* is not able to deal with synonyms, because of the n-grams approach explained above. We noticed that different patents describe similar objects using different terms, for instance "stroller" and "buggy" may be used to describe the same baby transport device. A possible solution to this failure may be to employ term expansion or to introduce synonyms into the n-gram generation phase.

5. Experiment 3: Passage Retrieval and Conflict Detection in Contracts

There are many kinds of documents that an enterprise must deal with: contracts, complaints, manuals, incident reports, technical reports, invoices, news, blogs, etc. Table 7 summarises just some of them taking into account the areas that an enterprise could be interested in: customer service, internal management, R+D or market analysis. Due to the overload of information to be processed by an enterprise on a daily basis, there is a real need to have tools such as passage retrieval systems in order to filter out the information that is not relevant. *JIRS PR* system, thanks to its simple n-gram approach, is able

Interest area	Documentation type
Market analysis	Blogs, news, invoices, etc.
Internal management	Technical reports, contracts, etc.
Customer service	Complaints, contracts, suggestions, etc.
R+D	Articles, technical reports, manuals, etc.

Table 7: Relevant documents for internal analysis in an enterprise

to process any document that is written in natural language and to return only the relevant information to the enterprise's needs.

In this section, with the help of an example built on an invented contract between an airline and a ground operations company, we explore the possibility of employing the *JIRS* passage retrieval system to retrieve from a contract those passages where there is a conflict. To this aim, we used the formal constraint framework introduced by [11]. A contract written in English is translated to a set of Contract Language ($\mathcal{CL}$) clauses. The contract in this form is analysed by a model checker in order to detect conflicts (i.e., obligations and prohibitions to do the same thing or obligations to do contradictory things, etc.). The model checking is carried out by means of the contract analysis tool *CLAN*[20] [8]. Once a conflict is detected, the tool gives a counter-example explaining where the conflict is. This is done at the logical level; the user has to manually find the clauses in the original raw text which correspond to the detected conflict. Our goal is to link the automated analysis based on logic with the passage retrieval system in order to take the output of *CLAN* (the $\mathcal{CL}$ formula showing the conflict), write the corresponding approximate English sentence and then use the passage retrieval to find the original English text containing the conflicting sentences. It should be noted that clauses are not directly related to natural language sentences, and the size of contracts may be considerably greater than the one of the "toy" contract proposed here, thus justifying the use of an IR system to extract the relevant passages.

In [11] Fenech et al. consider a contract between an airline company and a company managing the ground operations (mainly the check-in process), where the normative specification is given as the following contract:

1. *The ground crew is obliged to open the check-in desk and request the passenger manifest two hours before the flight leaves.*
2. *The airline is obliged to reply to the passenger manifest request made by the ground crew when opening the desk with the passenger manifest.*
3. *After the check-in desk is opened the check-in crew is obliged to initiate the check-in process with any customer present by checking that the passport details match what is written on the ticket and that the luggage is within the weight limits. Then they are obliged to issue the boarding pass.*

[20] Available from www.cs.um.edu.mt/ svrg/tools/CLTool/

4. *If the luggage weighs more than the limit, the crew is obliged to collect payment for the extra weight and issue the boarding pass.*
5. *The ground crew is prohibited from issuing any boarding cards without inspecting that the details are correct beforehand.*
6. *The ground crew is prohibited from issuing any boarding cards before opening the check-in desk.*
7. *The ground crew is obliged to close the check-in desk 20 minutes before the flight is due to leave and not before.*
8. *After closing check-in, the crew must send the luggage information to the airline.*
9. *Once the check-in desk is closed, the ground crew is prohibited from issuing any boarding pass or from reopening the check-in desk.*
10. *If any of the above obligations and prohibitions are violated a fine is to be paid.*

We will illustrate the formal conflict analysis process and the way the *JIRS* passage retrieval system is linked up to a deontic logic based on $\mathcal{CL}$. First of all the airline check-in desk is translated into $\mathcal{CL}$. A generic $\mathcal{CL}$ clause has the form $[\beta]C$, meaning that if action β is performed, then contract C must be executed. Contracts may be obligations (O), permissions (P) or prohibitions (F). The notation $O_C(\alpha)$ indicates a contract in which action α must be executed, with C the contract to be taken in case α is not executed. For instance, the reparation clause that corresponds to the penalty in our contract is $O(fine)$. The & operator indicates concurrency, and $\overline{\alpha}$ means "any action except α". The clauses written in natural language that include an implicit universal quantification, that is, statements such as "After the check-in desk is open" need to be interpreted as "At any time, after the check-in is open". This is reflected by the notation ($[1^*]$). The clauses of our example contract can be translated as follows.

1. $[1^*][2hBefore]O_{O(fine)}(openCheckIn \ \& \ requestInfo)$
2. $[1^*][openCheckIn \ \& \ requestInfo]O_{O(fine)}(replyInfo)$
3. $[1^*][openCheckIn][1^*](O(correctDetails \ \& \ luggageInLimit) \wedge$ $[correctDetails \ \& \ luggageInLimit]O_{O(fine)}(boardingCard))$
4. $[1^*][openCheckIn][1^*][correctDetails \ \& \ luggageInLimit]$ $O_{O(fine)}(collectPayment \ \& \ boardingCard)$
5. $[1^*][\overline{correctDetails}]F_{O(fine)}(boardingCard)$
6. $[\overline{openCheckIn}]F_{O(fine)}(boardingCard)$
7. $([1^*][20mBefore]O_{O(fine)}(closeCheckIn)) \wedge ([\overline{20mBefore}^*]$ $F_{O(fine)}(closeCheckIn))$
8. $[1^*][closeCheckIn]O_{O(fine)}(sendLuggageInfo)$
9. $[1^*][closeCheckIn][1^*](F_{O(fine)}(openCheckIn) \wedge$ $F_{O(fine)}(boardingCard))$

The conflict discovery algorithm identifies a state in conflict labelled with the obligation to perform action *boarding Card* and the prohibition of performing

action *boarding Card* together with a trace leading to this state: once the crew opens the check-in desk (clause 3), they are always obliged to issue a boarding pass if the client has the correct details but, they are prohibited from issuing a boarding pass, once the check-in desk is closed (clause 9). Clauses 3 and 9 are in conflict once the check-in desk is closed and a client arrives to the desk with the correct details. Therefore, in order to fix this problem we need to change clause 3: after the check-in desk is opened, the ground crew is obliged to issue the boarding pass as long as the desk has not been closed. *CLAN* tools returns a trace that identifies the situation in which the check-in desk is closed at the same time as the client provides her correct details:

$< openCheckIn, closeCheckIn \ \& \ correctDetails, O(boardingCard) \ \& \ F(boardingCard) >$

In reality, a check-in desk cannot accept the passport details and close at the same time. In order to resolve this conflict, the two actions need to be mutually exclusive. In order to ensure that *2hBefore* and *20mBefore*, as well as *openCheckIn* and *closeCheckIn* occur in the correct order, we need to make use of path constraints. Therefore, clauses 3 and 4 need to be modified as follows:

- $3'$: $[1^*][openCheckIn][\overline{closeCheckIn}^*][correctDetails \ \& \ luggageInLimit]O_{O(fine)}(boardingCard)$
- $4'$: $[1^*][openCheckIn][\overline{closeCheckIn}^*][correctDetails \ \& \ luggageOverLimit]O_{O(fine)}(collectPayment \ \& \ boardingCard)$

In order to find in the contract written in natural language where the sentences containing the conflicts are, the output of *CLAN* could be translated from $\mathcal{CL}$ in natural language as follows:

- $3''$: open check-in close check-in correct details luggage in limit fine boarding card
- $4''$: open check-in close check-in correct details luggage over limit fine collect payment boarding card

JIRS is fed with queries $3''$ and $4''$, and thanks to its n-gram approach, it is able to automatically retrieve those sentences in the contract where the conflict occurs. *JIRS* returns a ranking list with the passages being most similar to each query. The obtained results for each query are shown in Table 8. It can be appreciated that the conflict clauses are ranked the top of the list. Although the airline check-in desk is a small case study, often services are frequently composed of different sub-services, each of which comes with its own contract. Therefore, not only it is important to ensure that each simple contract is conflict-free but also that the composition of all contracts has to be also conflict-free. Therefore, in the case of a real scenario, once the conflict has been detected by the *CLAN* tool, a passage retrieval system such as *JIRS* may help to automatically retrieve the passages where the conflict occurs that instead should be found manually.

query 3"		query4"	
passage	weight	passage	weight
5	0.25	4	0.35
4	0.23	5	0.20
3	0.23	3	0.18
1	0.19	1	0.16
7	0.19	7	0.16
8	0.15	8	0.13
6	0.14	6	0.12
9	0.14	9	0.12
10	0.10	10	0.09
2	0.06	2	0.05

Table 8: *JIRS* ranking of clauses

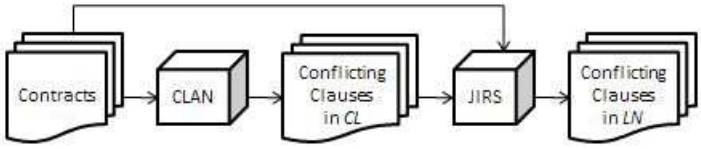

Figure 10: Architecture of the experiment *CLAN* - *JIRS*: conflicting clauses in Contract Language ($\mathcal{CL}$) and in Natural Language (*NL*)

Figure 10 shows the architecture of the experiment. We used the output of the conflict-detection tool (*CLAN*) to find the conflict clauses in natural language using JIRS.

6. Conclusions and further work

The automatic analysis of legal texts is a recent development in NLP and IR which consists in mining useful information from such closed domain documents, such as the identification of conflicts in contracts, the existence of a similar patents, or the retrieval of previous cases in order to find out how they have been resolved. In this paper we have described how a passage retrieval system may help in such scenarios. Two case studies of legal tasks have been presented and the JIRS passage retrieval system has been employed in order to retrieve relevant text passages when analysing legal treaties and patents. The described methods were tested in the CLEF framework, for cross-lingual retrieval in the legal domain, and patent retrieval.

This paper also attempts to shorten the gap between the natural language processing and the formal language and methods research communities, by exploring the possibility of employing JIRS for conflict resolution in contracts. Our experiments showed that JIRS, thanks to its n-gram approach, helped in correctly identifying and retrieving the sentences containing conflicts in a given

contract. We consider that the combination of the passage retrieval technique with the analysis of CLAN's output opens many possibility for further collaboration between the two research communities. For instance, if $\mathcal{CL}$ clauses are translated into a structured English version of a $\mathcal{CL}$ formula (and vice versa), and CLAN is used to detect conflicts, JIRS passage retrieval system could help to find the location in the original text of the sentences expressed in the restricted English version.

As future work we aim to apply formal language and methods on a reduced subset of relevant passages previously retrieved by *JIRS* when analysing legal treaties and patents as well as contracts, claims and other kinds of legal texts.

Acknowledgements

We thank the MICINN (Plan I+D+i) TEXT-ENTERPRISE 2.0: (TIN2009-13391-C04-03) research project. The work of the second author has been possible thanks to a scholarship funded by Maat Gknowledge in the framework of the project with the Universidad Politécnica de Valencia *Módulo de servicios semánticos de la plataforma G.*

References

[1] P. J. van Koppen, N. H. Roos, Rationality, Progress and Information in Psychology and Law. Liber Amicorum Hans F.M. Crombag, Metajuridica Publications, Maastricht, 2000.

[2] S. Correa, D. Buscaldi, P. Rosso, A. Rios, Passage Retrieval and Intellectual Property in Legal Texts, in: FLACOS-2009, University of Oslo, Toledo, Spain, 2009, pp. 61–71, September 24-25.

[3] G. Roda, J. Tait, F. Piroi, V. Zenz, CLEF-IP 2009: retrieval experiments in the Intellectual Property domain, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 385–409.

[4] E. Chieze, A. Farzindar, G. Lapalme, An Automatic System for Summarization and Information Extraction of Legal Information, in: E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia (Eds.), Semantic Processing of Legal Texts, Vol. 6036 of Lecture Notes in Computer Science, Springer, 2010, pp. 216–234.

[5] E. Francesconi, S. Montemagni, W. Peters, D. Tiscornia, Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language, in: Semantic Processing of Legal Texts, Vol. 6036 of Lecture Notes in Computer Science, Springer, 2010.

[6] A. Sarkar, P. H. Garthwaite, A. De Roeck, A Bayesian Mixture Model for Term Re-occurrence and Burstiness, in: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 48–55.

[7] C. Prisacariu, G. Schneider, An Action-Based Logic for Reasoning about Contracts, in: WoLLIC, 2009, pp. 335–349.

[8] S. Fenech, G. J. Pace, G. Schneider, CLAN: A Tool for Contract Analysis and Conflict Discovery, in: Z. Liu, A. P. Ravn (Eds.), ATVA, Vol. 5799 of Lecture Notes in Computer Science, Springer, 2009, pp. 90–96.

[9] B. Martin, Plagiarism: Policy Against Cheating or Policy For Learning?, in: Newsletter of the Australian Sociological Association, Vol. 16, 2004, pp. 15–16.

[10] G. J. Pace, M. Rosner, A Controlled Language for the Specification of Contracts, in: Controlled Natural Language, Workshop on Controlled Natural Language, CNL 2009, Marettimo Island, Italy, June 8-10, 2009. Revised Papers, Vol. 5972 of LNCS, Springer, 2010, pp. 226–245.

[11] S. Fenech, G. J. Pace, G. Schneider, Automatic Conflict Detection on Contracts, in: M. Leucker, C. Morgan (Eds.), ICTAC, Vol. 5684 of Lecture Notes in Computer Science, Springer, 2009, pp. 200–214.

[12] S. Tellex, B. Katz, J. Lin, A. Fernandes, G. Marton, Quantitative evaluation of passage retrieval algorithms for question answering, in: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM, New York, NY, USA, 2003, pp. 41–47. doi:10.1145/860435.860445.

[13] A. Ittycheriah, M. Franz, W.-J. Zhu, A. Ratnaparkhi, R. J. Mammone, IBM's Statistical Question Answering System, in: TREC, 2000.

[14] G. G. Lee, J. Seo, S. Lee, H. Jung, B.-H. Cho, C. Lee, B.-K. Kwak, J. Cha, D. Kim, J. An, H. Kim, K. Kim, SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP, in: TREC, 2001.

[15] G. Bouma, G. Kloosterman, J. Mur, G. van Noord, L. van der Plas, J. Tiedemann, Question Answering with Joost at CLEF 2007, in: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, D. Santos (Eds.), CLEF, Vol. 5152 of Lecture Notes in Computer Science, Springer, 2007, pp. 257–260. doi:10.1007/978-3-540-85760-0_30.

[16] S. Roger, K. Vila, A. Ferrández, M. Pardiño, J. M. Gómez, M. Puchol-Blasco, J. Peral, Using AliQAn in Monolingual QA@CLEF 2008, in: Peters et al. [42], pp. 333–336. doi:10.1007/978-3-642-04447-2_38.

[17] Á. Martínez-González, C. de Pablo-Sánchez, C. Polo-Bayo, M. T. Vicente-Díez, P. M. Fernández, J. L. Martínez-Fernández, The MIRACLE Team at the CLEF 2008 Multilingual Question Answering Track, in: Peters et al. [42], pp. 409–420.

[18] R. Strötgen, T. Mandl, R. Schneider, A Fast Forward Approach to Cross-Lingual Question Answering for English and German, in: Peters et al. [43], pp. 332–336.

[19] C. Amaral, H. Figueira, A. F. T. Martins, A. Mendes, P. Mendes, C. Pinto, Priberam's Question Answering System for Portuguese, in: Peters et al. [43], pp. 410–419.

[20] R. Besançon, M. Embarek, O. Ferret, The OEDipe System at CLEF-QA 2005, in: Peters et al. [43], pp. 337–346.

[21] M. Adriani, Rinawati, Finding Answers to Indonesian Questions from English Documents, in: Peters et al. [43], pp. 510–516.

[22] R. F. E. Sutcliffe, M. Mulcahy, I. Gabbay, A. O'Gorman, D. Slattery, Cross-Language French-English Question Answering Using the DLT System at CLEF 2005, in: Peters et al. [43], pp. 502–509.

[23] H. Tanev, M. Kouylekov, B. Magnini, M. Negri, K. I. Simov, Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005, in: Peters et al. [43], pp. 390–399.

[24] S. Hartrumpf, Extending Knowledge and Deepening Linguistic Processing for the Question Answering System InSicht, in: Peters et al. [43], pp. 361–369.

[25] L. Costa, 20th Century Esfinge (Sphinx) Solving the Riddles at CLEF 2005, in: Peters et al. [43], pp. 467–476.

[26] J. M. Gómez, P. Rosso, E. Sanchis, Re-ranking of Yahoo snippets with the JIRS Passage Retrieval system, in: In Proceedings workshop on cross lingual information access (CLIA-2007), 2007, joint conf. on artificial intelligence (IJCAI-07).

[27] D. Buscaldi, P. Rosso, J. M. G. Soriano, E. Sanchis, Answering Questions with an n-gram based Passage Retrieval Engine, J. Intell. Inf. Syst. 34 (2) (2010) 113–134.

[28] A. Peñas, P. Forner, R. F. E. Sutcliffe, Á. Rodrigo, C. Forascu, I. Alegria, D. Giampiccolo, N. Moreau, P. Osenova, Overview of respubliqa 2009: Question answering evaluation over european legislation, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 174–196.

[29] A. Rodrigo, J. Pérez, A. Peñas, G. Garrido, L. Araujo, Approaching Question Answering by means of Paragraph Validation, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 242–252.

[30] G. Puscaşu, A. Iftene, I. Pistol, D. Trandabăţ, D. Tufiş, A. Ceauşu, D. Ştefănescu, R. Ion, C. Orăsan, I. Dornescu, A. Moruz, D. Cristea, Developing a Question Answering System for the Romanian-English Track at CLEF 2006, in: C. Peters, P. Clough, F. C. Gey, J. Karlgren, B. Magnini, D. W. Oard, M. de Rijke, M. Stempfhuber (Eds.), Working Notes for the CLEF 2006 Workshop, Vol. 4730 of Lecture Notes in Computer Science, Springer, Alicante, Spain, 2006.

[31] R. Ion, D. Ştefănescu, A. Ceauşu, D. Tufiş, E. Irimia, V. Barbu-Mititelu, A Trainable Multi-factored QA System, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 257–264.

[32] S. Correa, D. Buscaldi, P. Rosso, NLEL-MAAT at CLEF-ResPubliQA, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 223–228.

[33] C. D. Manning, P. Raghavan, H. Schtze, Introduction to Information Retrieval, Cambridge University Press, New York, NY, USA, 2008.

[34] S. Jones, S. Walker, M. Gatford, T. Do, Peeling the Onion: OKAPI system architecture and software design issues, Journal of Documentation 53 (1) (1997) 58–68.

[35] P. Lopez, L. Romary, Multiple Retrieval Models and Regression Models for Prior Art Search, CoRR, 2009 abs/0908.4413.

[36] D. E. L. José Carlos Toucedo, University of Santiago de Compostela at CLEF-IP09, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 418–425.

[37] P. R. Julien Gobeill, Douglas Theodoro, Exploring a wide Range of simple Pre and Post Processing Strategies for Patent Searching in CLEF IP 2009, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF

2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 444–451.

[38] Y. Fang, L. Si, A. Mathur, FacFinder: Search for Expertise in Academic Institutions, Tech. rep., SERC-TR-294 and Department of Computer Science, Purdue University (2008).

[39] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (4) (2002) 422–446. doi:10.1145/582415.582418.

[40] S. Correa, D. Buscaldi, P. Rosso, NLEL-MAAT at CLEF-IP, in: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers, Vol. 6241 of LNCS, Springer, 2010, pp. 438–443.

[41] S. Hassan, R. Mihalcea, C. Banea, Random-Walk Term Weighting for Improved Text Classification, in: ICSC '07: Proceedings of the International Conference on Semantic Computing, IEEE Computer Society, Washington, DC, USA, 2007, pp. 242–249. doi:10.1109/ICSC.2007.71.

[42] C. Peters, T. Deselaers, N. Ferro, J. Gonzalo, G. J. F. Jones, M. Kurimo, T. Mandl, A. Peñas, V. Petras (Eds.), Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers, Vol. 5706 of Lecture Notes in Computer Science, Springer, 2009.

[43] C. Peters, F. C. Gey, J. Gonzalo, H. Müller, G. J. F. Jones, M. Kluck, B. Magnini, M. de Rijke (Eds.), Accessing Multilingual Information Repositories, 6th Workshop of the Cross-Language Evalution Forum, CLEF 2005, Vienna, Austria, 21-23 September, 2005, Revised Selected Papers, Vol. 4022 of Lecture Notes in Computer Science, Springer, 2006.