The Patents Retrieval Prototype in the MOLTO Project

Milen Chechev[†] Meritxell Gonzàlez[§] Lluís Màrquez[§] Cristina España-Bonet[§]

[†]Ontotext AD
Sofia 1504, Bulgaria
milen.chechev@ontotext.com

[§]TALP Research Center
Barcelona 08034, Spain
{mgonzalez,lluism, cristinae}@lsi.upc.edu

ABSTRACT

This paper describes the patents retrieval prototype developed within the MOLTO project. The prototype aims to provide a multilingual natural language interface for querying the content of patent documents. The developed system is focused on the biomedical and pharmaceutical domain and includes the translation of the patent claims and abstracts into English, French and German. Aiming at the best retrieval results of the patent information and text content, patent documents are preprocessed and semantically annotated. Then, the annotations are stored and indexed in an OWLIM semantic repository, which contains a patent specific ontology and others from different domains. The prototype, accessible online at `http://molto-patents.ontotext.com`, presents a multilingual natural language interface to query the retrieval system. In MOLTO, the multilingualism of the queries is addressed by means of the GF Tool, which provides an easy way to build and maintain controlled language grammars for interlingual translation in limited domains. The abstract representation obtained from the GF is used to retrieve both the matched RDF instances and the list of patents semantically related to the user's search criteria. The online interface allows to browse the retrieved patents and shows on the text the semantic annotations that explain the reason why any particular patent has matched the user's criteria.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

Keywords

Patent Translation, Multilingual Information Retrieval, Automatic Semantic Annotations

1. INTRODUCTION

The MOLTO project[1] addresses high quality domain specific machine translation (MT). In particular, it is focused on the study and development of novel tools and new resources with sufficient level of speed and automation for real-time translation tasks. Examples of such advances are the tools for multilingual translation on the Web [9] and the Translator Tools API described and implemented in [2]. The MOLTO project has also several prototypes in which to integrate these technological advances. The aim of each prototype, defined as specific case studies, is to demonstrate the usability and productivity of the techniques developed in the project.

In this paper we describe the multilingual patent retrieval prototype. It addresses a case study of the MOLTO project centered on the patents domain. This case study aims to create a prototype for MT and retrieval of patents, allowing translation of patent abstracts and claims, cross-language retrieval of patent data and multilingual queries.

In the following, Section 2 presents the general guidelines of the MOLTO project. Section 3 gives an overview of the patents case study and the main challenges that motivated our work. Section 4 gives a detailed description of the patent retrieval system and Section 5 introduces the online demo and the functionalities integrated in its interface. Finally, in Section 6 we lay out the future directions of the prototype.

2. THE MOLTO PROJECT

The mission of the MOLTO project is to develop a solution for real-time multilingual translation of web documents with high quality. Nowadays there is a real need to maintain the synchrony of multilingual repositories of any kind of documents. The Wikipedia and the EU Parliament are common examples that represent this reality. However, while in the EU Parliament the number of translations is limited to the official languages and the content is officially edited, in the Wikipedia, instead, the number of languages is higher and the edition of the articles is not synchronized at all. In general, multilingual environments can be characterized by the following features: i) the number of languages, ii) the number of authors, iii) the frequency of the updates, iv) the synchrony across languages, and v) the quality of the texts.

In big environments where the number of languages and the frequency of the updates are high, the synchrony of the documents is generally incompatible with translations quality. The main reason is that human translation is not feasible in such environments, and while state-of-the-art translation systems such as Google Translate[2] or Systran[3] are good for general purpose translations, they cannot produce high quality texts in specific domains where reliability is required.

[1]http://www.molto-project.eu

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

[2]http://translate.google.com/
[3]http://www.systransoft.com

The MOLTO technology addresses simultaneously the five features described above, covering up to 15 languages. Quality is achieved by means of the use of a restricted language. Edition and synchrony is possible due the MOLTO software toolkit, making the restricted language translation much more practical and scalable. The toolkit consists of a family of open-source software products: a grammar development tool IDE and API, a translator's tool API and browser interfaces, a grammar library for linguistic resources, and a grammar library for the domains addressed in the project. The tools are also being extended to a semantic wiki platform and the user-friendliness of these solutions for non-expert grammar writers will be tested.

The technological basis of MOLTO are the Grammatical Framework[4](GF, a grammar-based system that provides multilingual robust translation of controlled language), the semantic web technology and the interoperability between both. In MOLTO, the grammar-based MT benefits from the semi-automatic creation of grammars from the ontologies that are available in standard formats such as RDF[5] and OWL[6]. Similarly, the knowledge datasets can benefit from the conceptual models captured by the multilingual grammars.

The feasibility and performance of the techniques developed in MOLTO are being shown in three case studies: Mathematics, a multilingual dialog system to help math students to solve problems; Cultural Heritage, a multilingual ontology to describe museum objects and a cross-lingual retrieval system to query the ontology; and Patents, a multilingual patents retrieval system able to translate patent abstracts and claims in English, German and French. The prototype based on the latter is the focus of this paper.

3. OVERVIEW OF THE PATENTS CASE STUDY

Nowadays, there are five main patent offices around the world: Japan, Korea, China, Europe and the United States. These offices manage a huge amount of documents describing the patented inventions. There is a clear need to exchange the information related to such inventions, either for carrying out the legal tasks characteristic of the patent offices, or for building systems able to search, access and translate patents content and make it available to the community. However, these documents are written in several languages and it is not possible to undertake this task through human translation (either due the outsize of the databases or the update frequency of the documents). This is an interesting scenario in which to apply the technologies developed within the MOLTO project. First, because the automatic translation of patents must be accurate since it involves legal texts; and second, the number of documents is huge and increasing, and the content of claims may change from the applications to the final reviews.

The Patents Structure. The files associated to every patent, normalized to an XML format, contain the terms of the patent and the bibliographic data. The standardized fields include dates, countries, languages, references, author names and companies as well as rich subject classifications. Moreover, every patent has a title, a description, an abstract with a short and general summary and a series of claims.

The MOLTO *patents case study* comprises the two basic scenarios described in the next subsections: online patent retrieval and multilingual patent translation, which can be viewed as a joint multilingual patent retrieval paradigm.

3.1 Patents Retrieval

In the patent retrieval scenario, end-users have access, in the chosen language, to any patent information, which may be originally produced in another language. In a batch process, all patent documents are stored in the databases of the retrieval system. Then, they are classified in multiple indexes according to the patent information that is obtained from the bibliographic data, the content of the claims and the language in which they are written. An end-user accessing the system to search any patent information may want to specify a matching criterion. In MOLTO, such criteria are written in natural language (NL) and they are translated into a relational representation between the terms of the query and the content of the patent indexes. The online translation of NL queries is grounded on the abstract syntax representation produced by the GF. The current interface, described in Section 5, allows to query the system in English and French under a controlled language designed for the patents domain. When the retrieval system returns a hit list of patents, they are shown to the user along with a brief NL answer, produced also in the query's language, and the set of ontological concepts that matched the query. This way, the retrieval interface introduces some interaction in which the user can have an idea of the results, select any of the documents to see the whole content and the semantic annotations, or update the query to obtain further results.

The Patents Database. In order to limit the scope of the problem and cope with the technological requirements, the case study is limited to the biomedical and pharmaceutical domain (IPC[7] code A61P). The retrieval database contains 7,705 patent documents obtained from the European Publication Server[8] published during 2011 and 2012 and having the IPC code A61P. 4,274 out of the 7,705 documents have claims, and 2,058 out of them are trilingual. 2,116 documents have claims written only in English, 66 have claims only in German and 34 only in French. No extra files have other combination of languages. None of them belong to the corpus used to train the machine translation systems described below. Hence, the text of the documents having multiple translations can be used for testing purposes related to the accuracy of the translation system and to compare the performance of the retrieval system when using original or machine translations.

3.2 Patents Translation

The patent translation scenario refers to the off-line translation of the patent documents. Patents can be translated when they are added to the repository (e.g., by the editor) or when they are retrieved from the repository (e.g., by the end-user). In MOLTO we focus on the first case. The translation is limited to the patent's claims and abstracts texts and the official languages of the European Patent Of-

[4]http://www.grammaticalframework.org/
[5]http://www.w3.org/RDF/
[6]http://www.w3.org/OWL/
[7]http://www.wipo.int/classifications/ipc/
[8]https://data.epo.org/publication-server/

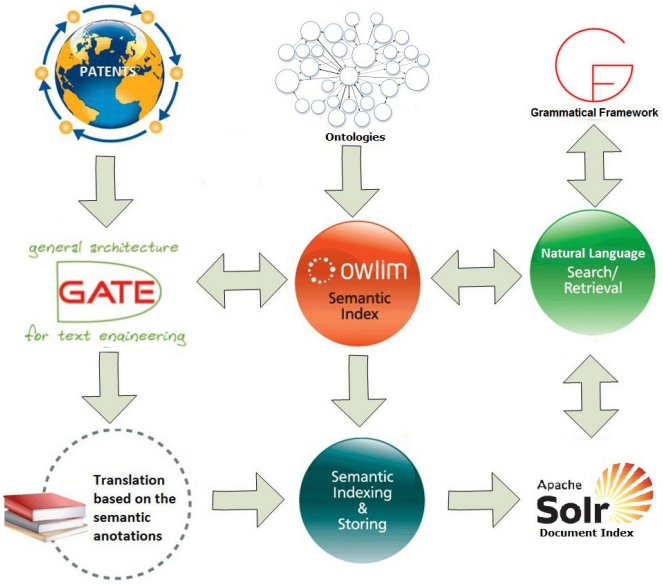

Figure 1: The retrieval architecture

fice: English, German and French. The research approach in MOLTO's machine translation is the hybridization between interlingual grammar-based GF [8] and specialized statistical machine translation (SMT) systems, which provide wide coverage and translation of unrestricted input.

Currently, patents are translated using a specialized SMT system based on Moses [7]. For training the system, a parallel corpus in the three languages has been gathered from the corpus of patents given for the CLEF-IP track in the CLEF 2010 Conference[9]. Our parallel corpus is a subset with those patents with claims and abstracts translated into the three languages and from the A61P domain. The final corpus built this way covers 56,000 patents which results into 279,282 aligned parallel fragments. Further details about the corpus, the MOLTO approach to patent translation and the system evaluation can be found in [5].

4. THE PATENTS RETRIEVAL PROTOTYPE

The MOLTO patent prototype consists of several modules (see Figure 1). First, there is a module for collecting patents from the European Publication Server. It is configured to collect patents published at a concrete period of time and having a concrete IPC classification code. The patent files, which use a specific XML and DTD format, are next processed by the semantic annotation tool. The semantic annotation tagger is based on [1] and [4]. It uses GATE [3] as the framework for processing the XML patent files to add useful annotations on the patent's abstracts and claims. The annotation is carried out by a pipeline, integrated in GATE, that includes the use of populated gazetteers with a large coverage of terms retrieved from the Linked Life Data[10] ontologies. These data include the FDA[11] drug information obtained from the FDA Orange Book and a wide range of diseases and other medical terms, also defined in the Unified Medical Language System (UMLS) standard[12]. The semantic annotations mark all the possible names of a particular drug or disease as a single instance, providing the possibility of better retrieval results. Furthermore, the goal of MOLTO is not just to receive the results but also to receive them in the user's language. So, the yet annotated patents are translated by the aforementioned MT systems.

The semantic repository OWLIM [6] and free text indexes are the places where the patents and their annotations are stored for further retrieval procedures. The semantic repository contains also an ontology that represents the patent information and the semantic items from the biomedical domain. This ontology contains 17 classes that describe concepts such as active ingredient, drug, applicant of the drug, dosage form, etc.

Finally, the search module is connected to the semantic repository. The user input is used to access the semantic repository and the free text indexes, providing the search functionality to both modules. The interfaces provided for both searching and browsing the repository content are described in the next section.

5. THE ONLINE DEMO

As previously mentioned, the retrieval system is available online at `http://molto-patents.ontotext.com` and can be queried in three different ways. The NL-based interface allows the user to query the system in English and French using written natural language. The SPARQL interface, more suitable for advanced users, allows to accurately browse the repository using SPARQL queries. The keyword-based visual browsing interface uses the RelFinder tool[13] in which the user can search for keywords using the autocomplete functionality. The results from the RelFinder search are visualised as graphs of results. For example, Figure 2 shows the relations found between `AMPICILLIN`, `TETRACYCLINE` and `ACETIC ACID`, including the patent document `EP-0092182-B1` found in the repository.

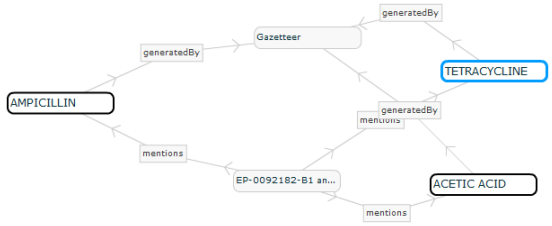

Figure 2: The Relfinder interface

The NL based Interface. The NL input is processed by a query grammar that produces an abstract representation of the user search criteria which is translated into SPARQL. The interface includes an autocomplete function to help the user writing queries under the controlled language supported by the grammar. The grammar covers a set of query topics (e.g., active ingredients of a drug, strength of a drug, dosage forms) for which we wrote a number of query examples. The initial set of query examples consisted of 131 sentence templates in English and their translation into French.

[9] http://clef2010.org/
[10] `http://linkedlifedata.com`
[11] Food and Drugs Administration, http://www.fda.gov
[12] http://www.nlm.nih.gov/research/umls
[13] http://relfinder.dbpedia.org

Nonetheless, the current version of the grammar generates (and therefore can process) a wider spectrum of sentences (591 sentence templates in English and 504 in French).

The Semantic Annotations. The visualization of the results displays the list of classes from the ontologies that match the query and the list of patent documents indexed under the matching criteria. The interface provides also a link to access the semantically annotated documents and the original patents. The interface that shows the annotated documents highlights on the text the words that are related to any semantic item. Colors are given according to the semantic annotations type. The right side of the page gives the list of semantic types and colors that are present in the text. The example in Figure 3 shows the annotations in the file `EP-0092182-B1` retrieved by the query: *`What is the information about "AMPICILLIN"?`*.

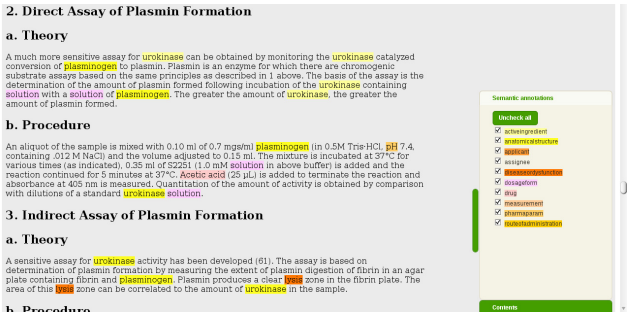

Figure 3: A semantically annotated patent text

6. FUTURE DIRECTIONS

The current prototype of the system is fully functional and serves to demonstrate the usability and productivity of the technologies developed in MOLTO for each scenario of the patents case study. After integrating all the components of the system, the future work points in multiple directions, such as new features in the interface and improvements in the annotation process and the translation methodology.

Regarding the online interface, the current prototype is able to parse natural language user queries written in English and French, although it will also be able to deal with German queries in short. Some experiments combining free text search and natural language queries under controlled language will be also conducted. Furthermore, the MOLTO project works on the generation of NL responses. In the multilingual patents retrieval prototype, the interface will incorporate a generation of natural language responses based on the abstract representation of the relations between the query, the matched semantic triples in the ontology and the labels of the actual retrieved instances. Then, the multilingual surface generation of the responses will be provided by the GF and a specific grammar that will be written for the case. The interface interaction will be also enhanced with additional functionalities. For instance, the interface will indicate whether the text showed has been obtained by machine translation and will provide links to the original text in the original language. Also, during the visualization of results the user will be able to select the desired language and change from one to another.

With respect to the patent annotation process, the semantic tagger will be enhanced in order to obtain more accurate annotations. Furthermore, the translation process will include a mechanism to keep the synchrony of the semantic annotations across several languages. An extension of this feature would project or transfer the annotations also to the original translations of the patent text.

We are currently involved in the evaluation of each component integrated in the prototype. Namely, the semantic annotator accuracy, the retrieval system efficiency, the automatic translations quality and the user interface usability. In MOLTO, the evaluations will be carried by both real users and automatic metrics.

7. ACKNOWLEDGMENTS

This work has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247914 (MOLTO project, FP7-ICT-2009-4-247914).

8. REFERENCES

[1] M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, and V. Tablan. Large-scale, Parallel Automatic Patent Annotation. In *Proc. of the 1st ACM workshop on Patent information retrieval*, PaIR '08, pages 1–8, New York, NY, USA, 2008. ACM.

[2] L. Carlson. D3.1. The Translation Tools API. MOLTO Project, 2011.

[3] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proc. of the 40th Meeting of the ACL.*, 2002.

[4] H. Cunningham, V. Tablan, I. Roberts, M. Greenwood, and N. Aswani. *Information Extraction and Semantic Annotation for Multi-Paradigm Information Management*, volume 29 of The Information Retrieval Series. Springer Berlin Heidelberg, 2011.

[5] C. España-Bonet, R. Enache, A. Slaski, A. Ranta, L. Màrquez, and M. Gonzàlez. Patent Translation within the MOLTO project. In *Proc. of the 4th Workshop on Patent Translation, MT Summit XIII*, pages 70–78, Xiamen, China, sep 2011.

[6] A. Kiryakov. OWLIM: Balancing Between Scalable Repository and Light-Weight Reasoner. In *In Proc. of WWW2006*, Edinburgh, Scotland, 2006.

[7] P. Koehn, H. Hoang, A. B. Mayne, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177–180, Jun 2007.

[8] A. Ranta. Grammatical Framework: A Type-Theoretical Grammar Formalism. *Journal of Functional Programming*, 14(2):145–189, 2004.

[9] A. Ranta, K. Angelov, and T. Hallgren. Tools for Multilingual Grammar-based Translation on the Web. In *Proc. of the ACL 2010 System Demonstrations*, pages 66–71, Stroudsburg, PA, USA, 2010.