MT Techniques in a Retrieval System of Semantically Enriched Patents

Meritxell Gonzàlez[1], **Maria Mateva**[2], **Ramona Enache**[3], **Cristina España**[1], **Lluís Màrquez**[1], **Borislav Popov**[2] and **Aarne Ranta**[3]

[1] Universitat Politècnica de Catalunya, Spain
{mgonzalez,cristinae,lluism}@lsi.upc.edu
[2] Ontotext AD, Bulgaria
{maria.mateva,borislav.popov}@ontotext.com
[3] University of Gothenburg, Sweden
{ramona.enache,aarne}@chalmers.se

Abstract

This paper focuses on how automatic translation techniques integrated in a patent retrieval system increase its capabilities and make possible extended features and functionalities. We describe 1) a novel methodology for natural language to SPARQL translation based on a grammar–ontology interoperability automation and a query grammar for the patents domain; 2) a devised strategy for statistical-based translation of patents that allows to transfer semantic annotations to the target language; 3) a built-in knowledge representation infrastructure that uses multilingual semantic annotations; and 4) an online application that offers a multilingual search interface over structural knowledge databases (domain ontologies) and multilingual documents (biomedical patents) that have been automatically translated.

1 Introduction

The five major international patent offices in the world maintain patent databases written mainly in English, French, German, Chinese, Japanese and Korean. Many other smaller offices maintain their databases as well, having documents written in their official languages. These offices and their users have the clear need to exchange the content of their databases. This need has actually promoted the organization of international conferences and competitions related to the field, such as patent classification, retrieval and translation, and the development of systems able to search, access and translate patent contents, either from monolingual or cross-lingual databases, and make them available to the community.

This paper presents a translation and multilingual retrieval system for biomedical patents, developed within the MOLTO project[1]. The principal characteristics of the MOLTO system are a multilingual repository of patents (all of them semantically annotated[2] and translated automatically) and the automatic translation from natural language (NL) queries to SPARQL. The integration of different translation methodologies into the system has been crucial to increase its capabilities and make possible extended features and functionalities, with respect to a preliminary version of the system described in (Chechev et al., 2012).

1.1 Related Work

In relation to the querying methods, most of the public search interfaces, either from the patent offices as the European Patent Office (EPO)[3] or independent ones as PatentLens[4], offer keyword-based search on the title, or multifaceted searching and browsing through the bibliographic data. Also, systems as Google Patents[5] allow free text search through the original text of the patents. In addition to these, the MOLTO system supports also controlled natural language queries, which allow to write richer and more expressive inquiries.

With respect to patent translation systems, the EPO public service, in combination with the

[1] http://www.molto-project.eu
[2] The semantic annotation is based upon http://www.ontotext.com/kim/semantic-annotation
[3] http://worldwide.espacenet.com/
[4] http://www.patentlens.ne
[5] http://www.google.com/patents

Google Patent Translate service, offers automatic translation of abstracts, descriptions and claims excerpts. A different approach is that of the Patent Language Translations Online (PLuTO) project, a dedicated project to patent translation. Its machine translation framework is a web service whereby users can request translations of excerpts. The translation engine uses the MaTrEx (Machine Translation Using Examples) system developed at DCU (Armstrong et al., 2006). It is a hybrid data-driven system built following established design patterns, with an extensible framework allowing for the interchange of novel or previously developed modules as it is defined in (Tinsley et al., 2010). The MOLTO approach to patent translation (España-Bonet et al., 2011) addresses full XML-like document translation using hybrid approaches and keeping the document structure.

Regarding the cross-lingual search, we use semantic repositories to annotate and index the patents content, which makes possible the use of more expressive queries that can deal with higher classes of the ontologies. The use of ontologies provides a shallow representation of the information space. Recent works use these light-weight ontologies to provide controlled lexicons for the classification of the content. However, few systems take a real advantage of the full potential of an ontological representation. This is the case of the ontology-based retrieval model described in (Vallet et al., 2005). Their model supports semantic search in document repositories through the exploitation of full-fledged domain ontologies and knowledge bases. As in MOLTO, full documents are returned in response to the user inquiries.

Finally, it is worth to mention the CLEF-IP track[6] launched by the Cross Language Evaluation Forum in 2009, although the purpose of the tasks is significantly different from the mission in MOLTO. The CLEF-IP track investigates the use of use of Information Retrieval techniques for patent document retrieval. One of their tasks consists in finding the set of patents that are relevant to a given topic. This is also the purpose of BioPatentMiner (Mukherjea and Bamba, 2004), a system for biomedical information retrieval especially designed to discover relationships among the concepts in the knowledge resources.

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

1.2 Patent Translation and Retrieval

The MOLTO patents prototype addresses automatic translation and retrieval of patents, allowing translation of patent abstracts, claims and descriptions, cross-language retrieval of patent documents and multilingual queries.

Two different approaches to machine translation (MT) have been used. For the massive translation of text, an statistical machine translation (SMT) system has been trained and adapted to transfer the semantic annotations to the target languages. On the other hand, an rule-based MT system (RBMT) is built in order to translate from natural language to the semantic query language.

Figure 1 shows a general architecture of the system. First, the English sections of the patents are annotated semantically with a GATE pipeline (Cunningham et al., 2011) designed especially for the biomedical domain. The RDF triples extracted from the documents and the domain ontologies are aligned and loaded in the semantic repository. Next, the annotated documents are automatically translated beforehand in order to fill the knowledge bases. Both, the original and the translated documents are indexed.

An interactive web-based user interface is accessible at http://molto-patents.ontotext.com. It allows for querying the system in English, French and German, using the controlled natural language (CNL). These queries are finally translated to SPARQL and the search engine in the retrieval system returns the documents that are relevant to the query.

In the following, Section 2 describes the re-

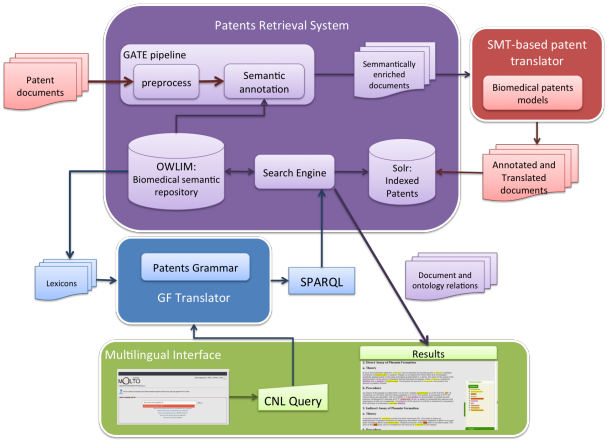

Figure 1: Patents prototype architecture

sources used to build up the whole platform. Next, Section 3 analyzes a methodology for building a query language on top of the ontology, sharing grammar components with the retrieval system. Then, Section 4 details the translation of the documents that feed the databases and how the method used endows the system with additional features. Finally, Section 5 describes the online user interface, and Section 6 summarizes the main contributions of this work.

2 Base System Resources

2.1 Grammatical Framework

The Grammatical Framework (GF, Ranta (2011)) is a type-theoretical grammar formalism, mainly used for multilingual NL applications. Grammars in GF are represented as pairs of an *abstract syntax*, an interlingua that captures the semantics of the grammar on a language-independent level, and a number of *concrete syntaxes* representing target languages. There are two main operations defined, *parsing* text to an abstract syntax tree and *linearizing* trees into raw text.

The GF Resource Library (Ranta, 2009) is the most comprehensive grammar for dealing with natural languages. It covers the morphology and basic syntax of 27 natural languages. The layered representation makes it possible to regard multilingual GF grammars as a RBMT system, where translation is possible between any pair of languages for which a concrete syntax exists.

2.2 Knowledge Representation

A generic knowledge representation infrastructure (KRI) was built for the purpose. It provides a mature basis for storage and retrieval of both knowledge and content. KRI uses OWLIM (Bishop et al., 2011) as a semantic data repository and allows browsing RDF data with the use of generic views. Moreover, it enables NL querying over the ontologies. The patent retrieval system is an overlay of KRI, which includes a module for document indexing and snipetting.

From the point of view of semantic data, the retrieval system is based on the Exopatent[7] project. It aligns several public ontologies and dictionaries to bring up an integral knowledge base and semantic annotation module in the biomedical patents domain. The conceptual model for the final ontology is given in Figure 2, where the following ontologies are included: PROTON[8], an ontology based on the FDA Orange book[9], UMLS[10], MeSH[11], as well as the auxiliary ontologies: the structural Semantic Annotation Repository(SAR), and the mapping SAFEPat.

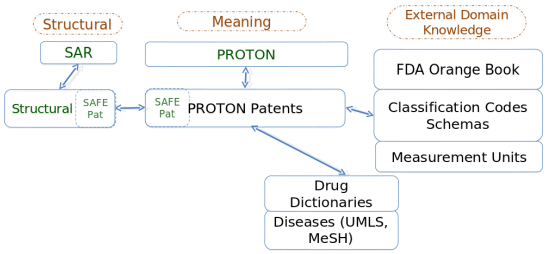

Figure 2: Conceptual model for the ontologies on the biomedical patents domain

2.3 Biomedical Patent Document Collections

The SMT system, used to translate the patent documents, requires a large parallel dataset to train the models. For this purpose we use the MAREC[12] corpus which contains European patents published between 1976 and 2008.

On the other hand, the EPO provided a website from where we downloaded 7,705 patent documents, also in the biomedical domain, all published between 2010 and 2012. These patent documents follow the XML specifications defined by the EPO, which consists namely of bibliographic data, abstract, description, claims and references. The abstract, the description and the claims are always written in one of the three official languages, i.e., English, French and German, and sometimes they contain also the translation to any of the other two languages or both of them. These documents constitute our patents retrieval knowledge base.

2.4 Moses Machine Translation System

The text of the patents is translated using a variant of the SMT system described in (España-Bonet et al., 2011). The automatic translator uses a phrase-

[7] http://exopatent.ontotext.com/

[8] http://www.ontotext.com/proton-ontology

[9] http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm

[10] http://krono.act.uji.es/people/Ernesto/UMLS_SN_OWL

[11] http://www.nlm.nih.gov/mesh/

[12] http://www.ir-facility.org/

based system adapted to and trained on the parallel patents in the biomedical domain mentioned above. A 5-gram language model is estimated using interpolated Kneser-Ney discounting with SRILM. Word alignment is done with GIZA++ and both phrase extraction and decoding are done with the Moses package. The optimization of the weights of the model is trained with MERT against the BLEU measure.

Besides a standard preprocess, a correct tokenization is important within the biomedical domain. For this reason, we devise a chemical compound recognizer and tokenizer based on affix detection. A list with approximately 150 affixes has been compiled and it is used to select the candidate tokens to be a compound from the corpus. The candidates selected this way are matched against a dictionary and those without a match are considered to be compounds and do not get an internal tokenization. $103,272$ compounds were found with this procedure within the training corpus, out of $7,954,491$ English tokens.

The translation strategy we follow allows to transfer the markup, including the semantic annotations, in the original source documents to the target text, yet enabling the multilinguality in the retrieval system. For this reason, the original SMT system was adapted in order to cope with additional encoding and tokenization requirements. To do so, all the training datasets were cleaned-up, the HTML markup was converted into UTF-8 symbols and the system was re-trained again.

3 CNL to SPARQL Translation with GF

We have explored the interoperability between query languages built with GF and ontologies. Traditionally, semantic data is queried using SPARQL, a query language for semantic datastores. However, this is not a friendly query method for unskilled users. The use of controlled languages is a possible solution, since they provide formal representation of a NL query, which is easily translated to any machine language syntax.

We devise a methodology for building a query language on top of ontologies, using the ontology relations and sharing the grammar lexicons (dictionaries) with the query language designed. This model is used in different domains within MOLTO (e.g., painting descriptions, business models and patents retrieval) and for different purposes, such as NL generation of objects' descriptions, and the query translation described in here.

GF provides multilinguality at a low cost, on the grounds of the Resource Grammar Library (RGL) and multilingual lexicons. It generates the abstract representation of the user's request. Then, the abstract syntax can be translated to any other language for which there is a concrete syntax. The transition between NL and GF concrete grammar is done automatically by the GF parser. The abstract syntax is defined according to the ontologies, whereas the concrete syntax is driven by the ontologies and the resource grammars.

3.1 The Patents Query Grammar

The patents query grammar includes the base YAQL module, whose principles are explained in (Ranta, 2012). It provides the basis for SPARQL generation from any RGL language, with just the minimum of lexical types. This generic grammar module can be reused for any domain.

The patents grammar builds on existing ontology classes (types) and relations to allow the formulation of NL queries that are in turn translated into a single SPARQL query. From the back-end perspective, it runs a GF process and the SPARQL is generated by a single translation command. For instance, the following command:

```
PatentQuery> p -lang=PatentQueryEng -cat=Query
    "show me the patents that mention AMPICILLIN" |
    l -lang=PatentQuerySPARQL
```

returns the SPARQL query shown below:

```
PREFIX pkm:<http://proton.semanticweb.org/protonkm#>
PREFIX psys:<http://proton.semanticweb.org/protonsys#>
CONSTRUCT {
    ?doc pkm:mentions ?d . }
WHERE {
    ?d psys:mainLabel "AMPICILLIN" .
    ?doc pkm:mentions ?d .
};
```

The constructors from the abstract syntax describe individual SPARQL queries. For example, the GF operation that builds the query that ask about patents mentioning a certain concept (e.g. drug) is matched to a parametrized function, which takes the concept's label as parameter and returns a corresponding SPARQL. (Dannells et al., 2013).

In the patents prototype, large dictionaries of terms (see Table 1), extracted from the ontologies, are used for the creation of grammar lexicons. In this way, the designed CNL provides all the necessary support for the queries over the semantic annotations in the documents and the ontologies in the information retrieval system.

We reach full coverage of the query language designed, and we extend it with semantically equivalent variants of each sentence. The next examples are different ways to paraphrase the query `show me the patents that mention AMPICILLIN` in English and French.

```
English:
what are the patents that mention AMPICILLIN
what are all the patents that mention AMPICILLIN
patents that mention AMPICILLIN
the patents that mention AMPICILLIN
all the patents that mention AMPICILLIN
give me the patents that mention AMPICILLIN
show me the patents that mention AMPICILLIN
give me all the patents that mention AMPICILLIN
show me all the patents that mention AMPICILLIN

French:
que sont les brevets qui mentionnent AMPICILLINE
que sont tous les brevets qui mentionnent AMPICILLINE
des brevets qui mentionnent AMPICILLINE
les brevets qui mentionnent AMPICILLINE
tous les brevets qui mentionnent AMPICILLINE
montre les brevets qui mentionnent AMPICILLINE
montre tous les brevets qui mentionnent AMPICILLINE
```

The current patent query grammar has high quality of the CNL generated and provides a direct mapping of the NL queries to SPARQL, which favors scalability and domain adaptation. Also, GF provides translation between the queries in the covered natural languages, as demonstrated in the online prototype.

3.2 Ontology Coverage by the Query Language

We defined different query topics, and we specified eight ontology types whose instances were of interest, and we lexiconized them for the query language. The semantic types with examples and number of corresponding entities in the ontology are shown in Table 1.

The extension of the query language introduced new useful queries. For example, we added the following queries that allow to search over the relations between concepts (e.g., `what are the drugs with the active ingredient ACTIVE_INGREDIENT`), and the combination of two variables (e.g., `what are the patents that mention Y and Z`). Our observation is that the more the grammar approximates the RDF predicates relations names, the easier and the more precise SPARQL queries can be introduced.

Annotation (Entity) type	Examples	Dict. size
Drug	FLUONID, HEXADROL	5,529
ActiveIngredient	BUDESONIDE, NIFEDIPINE	1,803
RouteOfAdministration	DENTAL, CARDIAC	53
Applicant	ASTRAZENECA, BAXTER	1,822
ApplicationNumber	10184202, 09814154	4,609
PatentNumber	EP2219477B1, EP2382981A3	4,609
TECode	AA, AB, BX	17
Market	DISCN, OTC, RX	3

Table 1: Dictionary types and number of entries extracted for the GF grammar lexicons

We have conducted a large number of tests on the queries to search if the documents returned reply to the NL queries. The lack of any variance in the possible answers, due the control over the SPARQL queries, makes the system reach high precision and high recall results. We practically execute a query over a semantic datastore and hence the only probability for incorrectly retrieved patents are errors in the design of the SPARQL grammars or incorrect semantic annotations.

4 Cross-lingual Patents Retrieval

As we have pointed earlier, this system is a domain specific overlay of the general purpose KRI. It integrates a specific GF-based queries grammar designed to cover the relations in the ontologies and the semantic types with their annotated instances in the documents. Furthermore, the system features full-text search enabled by Apache Solr[13], with index built from the documents' content in the different languages.

The SPARQL generation was delegated completely to GF, as explained in Section 3. In addition, Solr is also used for document snipetting, which results in significantly faster response of the system compared to its previous versions.

4.1 Document Collection and Semantic Annotations

From all the documents gathered from the EPO website, up to 4,609 out of the 7,705 documents contain at least one section with abstracts, claims or descriptions written in any of the languages. This is the final dataset that we annotated using domain-specific knowledge. Table 2 gives a numerical description of the dataset, i.e., the number of documents having the required sections in English, German and French, respectively.

The original text of the patent documents contains special elements, such as superscripts, subscripts, images and chemical formulae. For this reason, the pipeline consists of several steps:

[13] http://lucene.apache.org/solr/

	Documents	Claims	Descriptions	Abstracts
EN	4,485	62,638	3,832	2,518
DE	2,047	32,007	192	80
FR	2,011	31,487	130	44

Table 2: Number of sections in the patents dataset

1) preprocessing the text in order to get rid of non-relevant marks, 2) converting the HTML codes into UTF-8 symbols and preserve the form of the chemical compounds during tokenization (there are a total of $1,097,243$ compounds in the documents, $243,823$ different names, out of $178,213,580$ English tokens) and 3) annotating the text with the semantic types that describe biomedical concepts and the patents structure. Table 3 gives a summary of the exposed concepts and the number of the corresponding instances found in the patents.

Concept	Instances	Concept	Instances
ActiveIngredient	285,192	AnatomicalStructure	2,170,330
Applicant	26,177	DiseaseOrDysfunction	1,218,063
DosageForm	241,279	Drug	160,698
Measurement	1,704,078	PharmaParam	39,330
Receptor	32,885	RouteOfAdministration	99,468

Table 3: List of semantic concepts and number of instances found in the dataset

The annotation technology is based on the GATE framework and a customized pipeline for biomedical patents that consists of gazetteers populated from the ontology resources, and various JAPE[14] rules. The semantic annotation step is followed by a process in which the annotations are triplified/RDF-ized, i.e., transformed into RDF triples. Then, the patent identifier is related to the annotations that are mentioned via the "mentions" predicate from the PROTON ontology. Consequently, the RDF triples are loaded and stored in the OWLIM repository. This allows to obtain information regarding both the patent documents and the characteristics of the drugs, diseases and other entities of interest available in the semantic knowledge base.

4.2 Automatic Translation of Semantically Enriched Patent Documents

The semantic annotations created with the above process are projected to the translated sections of the patent documents. From semantic data's perspective, this is a huge step forward as detailed in the next section.

The designed process for patents translation allows for building a translated document having the same XML structure as the original patent, including the semantic information. Our purpose is to translate semantically enriched text in order to transfer the semantic knowledge to the translations. As a result, the retrieval system can index the translated patents, no matter in which language they are written, and the interface of the system can show the translated text and its relevant concepts.

The patent text is extracted from the sections in a structured manner. The resulting text is marked, segmented and tokenized as required by the machine translation system described in Section 2. The goal is to avoid the excessive segmentation of the sentences and improve the translation quality yet having longer segments. To do so, we make use of the "zone" and "wall" functionalities of the Moses translator. This way, we can maintain the position of the marks in the text while certain degree of word reordering is still allowed. After this step, the structural marks are removed and the remaining consists of raw text.

Then, the translated text is post-processed in order to recover the original structure of the document, including original formatting, claims enumeration and images, yet following the original XML document structure.

An online demo is available[15] to show the patent translation process. Patent documents should be written according to the EPO specifications, either with or without semantic annotations. It can also be used remotely to facilitate its integration with other tools.

4.3 Document Snippeting with Respect to Semantically Annotated Text

For each patent from the result we show the part of it (a snippet), where the annotated entity of interest appears. In previous versions of the retrieval system, its response time was too high due to the documents' snippetting. The reason for this is that we return documents that contain certain drug, active ingredient, etc. without track of the place in the text where it appears.

[14] http://gate.ac.uk/sale/tao/splitch8.html#chap:jape

[15] http://nlp.lsi.upc.edu/molto

Previously, for each returned document, all annotations had to be checked, in order to find the one whose label matches certain search criteria. We propose a speed up approach with two major components - an Apache Solr index, which is over the whole documents' content, separate for the different languages, and an auxiliary index, that offers the exact text string from the document for each {`document_id, annotation_label, query_language`} triple. When a semantic entity is searched, the system returns a Solr highlight over the mention in the text, that is defined by the above criteria and the auxiliary index (the document identifier comes from the RDF results and the annotation label is usually taken from the SPARQL query itself).

These changes optimized the system speed about 5 times.

5 Online Interface to Access the Patents Retrieval System

The user interface, which is publicly available online, allows for querying the system in English, French and German.

The queries are written using a *incremental query* input text box that accepts the CNL defined by the GF query grammar for patents. The GF engine parses the user's query and translates it to SPARQL; which in turn is executed against the semantic repository. As a result, the user receives the set of retrieved documents along with their snippets and the RDF facts matched. The graphical interface displays also an interactive list of domain concepts and documents that allows for browsing the ontology and inspecting the patent documents.

There has been a general observation that the controlled language lacks extensive coverage of the human speech. Even in a closed domain, such as biomedical patents, a user may want to experiment with different queries and end ask a question that is not covered by the the CNL, but is considered essential from a human's perspective.

Back-off mechanisms to the interpretation of controlled natural language queries can vary. We have proposed full-text search (FTS) as a feasible solution for our system. The major motivation behind this decision is the fact that our system, apart from knowledge from a semantic database, provides documents as results. Therefore, for those cases in which the user's query cannot be parsed by GF, we have enabled the FTS that returns to the user only the documents and snippets that contain the keywords in the search. This enables for searching on entities that are not in the ontology and our predefined dictionaries.

The interface has an incremental query functionality which is smart enough to suggest only the possible controlled queries and the applicable lexical types (e.g., drug, therapeutic equivalence code, and so on) for them.

5.1 Use Case Example

The CNL used in the interface was expanded and improved with respect to the ontology. Some of the new queries are essential for the usefulness of the system. For example, it is now possible to search a patent by its number, or by its application number (these were introduced as new annotation types, extracted from the patents metadata). We also introduced series of queries like `give me all the drugs with the` *`market RX`*, where we collect drugs with common characteristics - e.g. market, active ingredient, therapeutic equivalence (TE) code, dosage form, etc.

Hereby we present a query example in English and French that the system allows. The question is `what is the information about AMPICILLIN`, and `quelle est l' information á propos de AMPICILLINE`, respectively. The system returns a number of documents from which we selected $EP2397497A2$. On Figures 3 and 4 we show the English original and the French automatic translation, thus including the translation names.

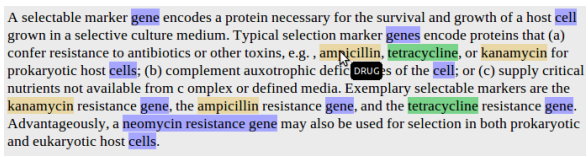

Figure 3: `AMPICILLIN` example in English

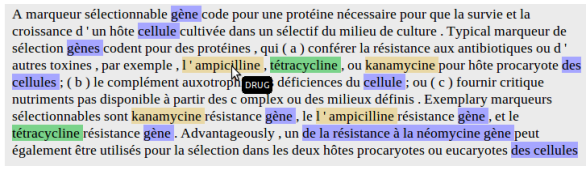

Figure 4: `AMPICILLINE` example in French

6 Conclusions and Challenges

The system presented in this work sets up the grounds where to combine machine translation, semantic data and retrieval techniques in order to come up with a useful platform for multilingual patent retrieval system. The main challenges addressed in the system were: a) to provide integrate semantic data model for the domain and annotate the corpus of documents, b) to translate biomedical patent documents, c) to design the mechanisms to enable the multilingual indexing and retrieval of the patents, d) to define and develop a query language and the query grammars in several languages for the system and e) to set up a website for online retrieval of patent document that serves as a testbed of our work.

The translation of the patent documents is mainly based on statistical machine translation techniques, although certain hybridization with rule-based systems (the GF) improves the quality of the translation. One of the challenges in this task was to come up with a mechanism to translate the semantics of the source texts to the target languages. What remains as a future challenge is the use of these annotations to increase either the accuracy of the annotations or the quality of the translations, or to favor domain adaptation in MT.

GF has been proved to be an efficient technology of generating the SPARQL queries, as if SPARQL was "Yet Another Query Language". This methodology facilitates the interoperability between the query grammar and the ontologies and speeds up the development and maintenance of the querying subsystem. In previous prototypes of the system, we used canned grammars build on a domain specific language. These grammars suffered from ambiguities and inconsistencies. In contrast, the GF-to-SPARQL approach minimizes the need for this maintenance. In gross, instead of a grammatical representation of a sentence using the Resource Grammar Library, the new approach provides a SPARQL representation. This automation saves lots of additional efforts with respect to providing a mapping from natural language to SPARQL.

Acknowledgments

This work has been funded by the European Community's Seventh Framework Programme (MOLTO project, FP7-ICT-2009-4-247914).

References

[Armstrong et al.2006] Armstrong, S., M. Flanagan, Y. Graham, D. Groves, B. Mellebeek, S. Morrissey, N. Stroppa, and A. Way. 2006. MaTrEx: Machine Translation Using Examples. In *TC-STAR OpenLab on Speech Translation*, Trento, Italy.

[Bishop et al.2011] Bishop, Barry, Atanas Kiryakov, Damyan Ognyanoff, Ivan Peikov, Zdravko Tashev, and Ruslan Velkov. 2011. OWLIM: A family of scalable semantic repositories. *Semantic Web Journal*, 2:33–42, June.

[Chechev et al.2012] Chechev, Milen, Meritxell Gonzàlez, Lluís Màrquez, and Cristina España-Bonet. 2012. The patents retrieval prototype in the MOLTO project. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 231–234, New York, NY, USA. ACM.

[Cunningham et al.2011] Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

[Dannells et al.2013] Dannells, Dana, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. 2013. Multilingual access to cultural heritage content of the semantic web.

[España-Bonet et al.2011] España-Bonet, Cristina, Ramona Enache, Adam Slaski, Aarne Ranta, Lluís Màrquez, and Meritxell Gonzàlez. 2011. Patent translation within the MOLTO project. In *Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII*, pages 70–78, Xiamen, China, sep.

[Mukherjea and Bamba2004] Mukherjea, Sougata and Bhuvan Bamba. 2004. BioPatentMiner: an information retrieval system for biomedical patents. In *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pages 1066–1077.

[Ranta2009] Ranta, Aarne. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2(1).

[Ranta2011] Ranta, Aarne. 2011. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

[Ranta2012] Ranta, A. 2012. *Implementing Programming Languages. An Introduction to Compilers and Interpreters, with an appendix coauthored by Markus Forsberg*. College Publications, London.

[Tinsley et al.2010] Tinsley, John, Andy Way, and Páraic Sheridan. 2010. PLuTO: MT for Online Patent Translation. In *Proceedings of the 9th Conferences of the Association for Machine Translation in the Americas (AMTA 2010)*.

[Vallet et al.2005] Vallet, David, Miriam Fernández, and Pablo Castells. 2005. An ontology-based information retrieval model. In *In ESWC*, pages 455–470. Springer.