Patent translation within the MOLTO project

Cristina España-Bonet[*], Ramona Enache[†], Adam Slaski[†],
Aarne Ranta[†], Lluís Màrquez[*] & Meritxell Gonzàlez[*]

August 25, 2011

Abstract

MOLTO is an FP7 European project whose goal is to translate texts between multiple languages in real time with high quality. Patents translation is a case of study where research is focused on simultaneously obtaining a large coverage without loosing quality in the translation. This is achieved by hybridising between a grammar-based multilingual translation system, GF, and a specialised statistical machine translation system. Moreover, both individual systems by themselves already represent a step forward in the translation of patents in the biomedical domain, for which the systems have been trained.

1 Introduction

MOLTO[1] is an European project within the Seventh Framework Programme. Its main goal is to develop a set of tools for translating texts between multiple languages in real time with high quality.

MOLTO clearly bets for high quality translation, the cost to pay is to limit the coverage to restricted domains which can be covered by a grammar. As its main technique, the project uses domain-specific semantic grammars and ontology-based interlinguas. These components are implemented in GF (Grammatical Framework) [8], which is a grammar formalism where multiple languages are related by a common abstract syntax. Up to now, GF has been applied in several small-to-medium size domains such as dialogue systems[2] or the translation of mathematical exercises[3].

When dealing with real text from a given domain, a grammar fails to cover any ungrammatical construction used. However, empirical machine translation systems in general, and statistical machine translation systems (SMT) in particular, have good coverage on any sort of text. The aim of MOLTO is to get the best of both worlds by building a hybrid GF-SMT system that achieve high-precision and good coverage.

Patents have been chosen for the opening of the system to non-restricted language. This election has two main reasons. First, the language of patents, although having a large amount

[*]{cristinae,lluism,mgonzalez}@lsi.upc.edu, TALP Research Center, LSI Department, Universitat Politècnica de Catalunya

[†]{ramona.enache,aarne}@chalmers.se, asl@mimuw.edu.pl, Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg

[1]MOLTO: FP7-ICT-247914, 2010–2013, www.molto-project.eu

[2]TALK project, Tools for Ambient Linguistic Knowledge: IST-507802, 2004–2006, www.talk-project.org

[3]WebALT project, Web Advanced Learning Technologies: EDC-22253, 2005–2007, webalt.math.helsinki.fi

Proceedings of the 4th Workshop on Patent Translation, pages 70-78,
MT Summit XIII, Xiamen, China, September 2011

SET	Segments	EN tok	DE tok	FR tok
Training	279,282	7,954,491	7,346,319	8,906,379
Development	993	29,253	26,796	33,825
Test	1,008	31,239	28,225	35,263

Table 1: Numbers for the patents aligned corpus in English (EN), German (DE) and French (FR).

of vocabulary and richness of grammatical structure, still uses a formal style that can be interpreted by a grammar. And second, there is nowadays a growing interest for patents translation. The high and increasing number of registered patents has created a huge multilingual database of patents distributed all over the world. So, there is an actual need for building systems able to access, search and translate patents, in order to make these data available to a large community.

The objective of MOLTO with respect to patents translation is twofold. On one hand, research on hybrid translation systems is being carried out to study the best approach to combine GF and a SMT system. On the other hand, a prototype for machine translation and retrieval within patents will be built. The purpose of this paper is to focus on the first part and depict the current status and prospects for the translation system.

The paper is organised as follows. The next section, Section 2, describes the corpora and linguistic processors used in this work. We detail in Section 3 the two independent systems to translate patents. Afterwards, Section 4 depicts the hybridization prospects for these systems, and finally Section 5 summarises and outlines future work.

2 Patents Domain

A patent is an official document granting a right. Besides the terms of the patent itself, it also contains information about its publication, authorship and classification for example. Being an official document, the structure giving the terms of the patent is quite fixed. Every patent has a title, a description, an abstract with the most relevant information and a series of claims.

A claim is a single (possibly very long) sentence composed mainly of two parts: an introductory phrase and the body of the claim, usually linked by a conjunction. It is in the body of the claim where there is the specific legal description of the exact invention. Therefore, claims are written in a lawyerish style and use a very specific vocabulary of the domain of the patent.

2.1 Corpus

MOLTO works with European patents and the task is restricted to English, French and German. A first domain of application includes biomedical and pharmaceutical patents. We select patents with IPC (International Patent Classification) code A61P, corresponding to "Specific therapeutic activity of chemical compounds or medical preparations".

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[4]. These data are an extract of

[4]http://clef2010.org/

the MAREC corpus, containing over 2.6 million patent documents pertaining to 1.3 million patents from the European Patent Office[5] (EPO). Our parallel corpus is a subset with those patents with translated claims and abstracts into the three languages. From this first subset we selected those patents that deal with the appropriate domain.

The final corpus built this way covers 56,000 patents out of the 1.3 million. That corresponds to 279,282 aligned parallel fragments as it can be seen in Table 1. A fragment is the minimum segment aligned in the three languages, so, it is shorter than a claim and, consequently, shorter than a sentence. Two small sets for development and test purposes have also been selected with the same restrictions: 993 fragments for development and 1008 for test.

2.2 Linguistic processors

The detection and correct tokenisation of chemical compounds has been shown to be crucial in the performance of translators (see Section 3 for the analysis). A regular tokeniser would for example split the *compound* "cis-4-cyano-4-(3-(cylopentyloxy)-4-methoxyphenyl)cyclohexane-1-carboxylic" by the puctuation into 9 tokens and, consequently, each of the tokens would be translated as an independent word. To deal with this peculiarity of the domain, we developed a pipeline to detect, tokenise and translate compounds.

2.2.1 Compound recogniser and tokeniser

As a first approximation we devise a recogniser and tokeniser based on affix detection. A list with 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 tokenisation. 103,272 compounds are found with this procedure within the training corpus defined in the previous section.

However, this list of compounds contains some noise. Examples of noise are in this context proper names with the defined affixes (Hôpital), words that do not appear in the dictionary (extracorporeal) or simply typos (comparoate). The amount of noise is considerable, but extra words do not in general imply a wrong tokenisation. So, the method works better as a (non-)tokeniser than as a compound detector and it bets for high recall instead of precision.

Given the power of GF, one can also build a simple grammar for translating compounds. What makes the difference between this rule-based approach and a mere translation of each word in the compound is that in this case the possible reordering of the words is already defined by the grammar. So, functional words like acid, ester or aldehyde swap its position with the radical words whenever necessary.

2.2.2 Part-of-speech tagger, lemmatiser and named entity recogniser

Part-of-speech (PoS) tagging and lemmatisation are necessary in the lexicon building of the patents grammar. `GENIA` [10], a linguistic processor prepared specially to process texts from the biomedical domain, is used for both purposes.

Named entities are marked in the text and are not translated by GF, but translated independently and substituted afterwards. In the biomedical domain, a simple heuristic works as well tagging proper names as a state-of-the-art tagger not specifically trained. We consider

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

to be proper names the words starting with a capital letter (after lowercasing the sentences), and the words containing numbers or special characters inside. This simple methodology lead to 100% precision and recall for the first 200 fragments in the training corpus of Section 2.1, where the proper names were manually annotated and the output was compared to that of the named entity recogniser. In this case 176 proper names were properly classified and replaced with a place holder name.

3 Individual translation systems for patents

The translation of patents can be approached through different methods. In this work we focus on GF and SMT systems, and specialise the two of them into the patents domain.

3.1 Interlingua-based translation, GF

The key concept of GF is the division of a grammar in an abstract syntax part and the concrete syntaxes corresponding to each of the target languages. The largest and most general example of such a grammar is the *resource library* [7], comprising 20 languages, for which the main grammatical constructions are provided. The library can be further used by domain-specific grammars, which can use the grammatical constructions from here alleviating the burden of handling linguistic difficulties and allowing a better focus on the higher-level details.

Even with this easiness, building a rule-based general-purpose translation system is a laborious task. However, we assume that most of the claims can be covered by a limited set of grammatical constructions and extend the GF resource grammar with these constructions.

Grammars like this one with non-trivial coverage usually are ambiguous, the number of the interpretations is the product of the number of parse trees for each subconstruction. On the good side, the grammar covers all possible interpretations, but on the other side, in order to make it usable, statistical based disambiguation needs to be used.

The task of translation is resumed to parsing from the source language to an abstract syntax tree and linearising it in the target language. Still, the system is restricted to the language generated by the grammar. Lexicon building is then an important step since the vocabulary of patent claims is virtually unlimited. The GF library multilingual lexicon contains the most common entries for structural parts-of-speech and it is used as a base to be extended with nouns, adjectives, verbs and adverbs. The abstract syntax for these PoS is created from the claims in one language (English). Once it is built, it is lemmatised and manually corrected from noise and ambiguities. Then, the proper inflection is generated using the implemented GF paradigms and the English dictionary of the GF library. Base forms are translated into the necessary languages and the inflection is generated for each of them.

The following figure shows the basic steps of the full system's behaviour:

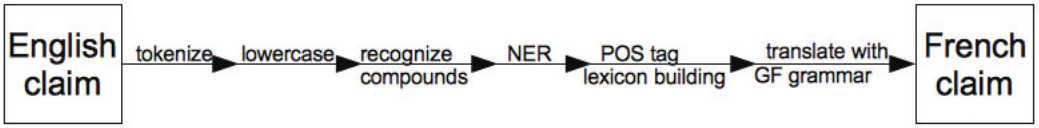

Figure 1: GF translation system for patent claims.

	DE2EN			EN2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.52	0.64	**0.72**	0.42	0.51	**0.69**
1−TER	0.59	0.67	**0.76**	0.45	0.53	**0.71**
BLEU	0.43	0.58	**0.65**	0.33	0.45	**0.58**
NIST	8.25	9.67	**10.12**	6.53	8.05	**9.40**
ROUGE-W	0.40	0.48	**0.52**	0.34	0.41	**0.48**
GTM-2	0.30	0.40	**0.47**	0.25	0.32	**0.43**
METEOR-pa	0.60	0.69	**0.74**	0.36	0.45	**0.57**
ULC	0.09	0.29	**0.41**	0.03	0.19	**0.43**

Table 2: Automatic evaluation using a set of lexical metrics of the in-domain SMT system for the English-German language pair. Results of two state-of-the-art systems, Bing and Google, are showed for comparision.

	FR2EN			EN2FR		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.54	0.66	**0.78**	0.57	0.63	**0.73**
1−TER	0.59	0.70	**0.80**	0.60	0.66	**0.74**
BLEU	0.45	0.62	**0.70**	0.43	0.53	**0.62**
NIST	8.52	10.01	**10.86**	8.39	9.21	**9.96**
ROUGE-W	0.41	0.50	**0.54**	0.39	0.45	**0.49**
GTM-2	0.32	0.43	**0.53**	0.31	0.36	**0.45**
METEOR-pa	0.61	0.72	**0.77**	0.57	0.65	**0.71**
ULC	0.07	0.28	**0.44**	0.10	0.23	**0.39**

Table 3: As in Table 2 for the English-French language pair.

Up to now, performance of the grammar aimed to parse full claims is still unsatisfactory. The high level of ambiguities remaining results in slowness, and coverage is up to now a 15% of the working corpus. Hybrid systems can deal with ambiguities, i.e., multiple translation options, and can complete with statistical translations the parts not covered by GF. However, the grammar must be expanded so that the two systems can collaborate on equal terms.

3.2 Statistical translation, SMT

The statistical system is a state-of-the-art phrase-based SMT system trained on the biomedical domain with the corpus described in Section 2.1. Its development has been done using standard freely available software. A 5-gram language model is estimated using interpolated Kneser-Ney discounting with `SRILM` [9]. Word alignment is done with `GIZA++` [5] and both phrase extraction and decoding are done with the `Moses` package [3, 2]. The optimisation of the weights of the model is trained with MERT [4] against the BLEU [6] evaluation metric.

Table 2 shows a first evaluation of this system (Domain) using a variety of lexical metrics. This set of metrics is a subset of the metrics available in the `Asiya` evaluation package [1]. We specifically select this set of metrics because all of them are available for the three languages: English, German and French. Together with our in-domain system we show the same

	DE2FR			FR2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.42	0.52	**0.76**	0.30	0.43	**0.65**
1−TER	0.47	0.56	**0.68**	0.32	0.46	**0.66**
BLEU	0.29	0.43	**0.56**	0.24	0.39	**0.53**
NIST	6.72	8.21	**9.10**	5.35	7.30	**8.88**
ROUGE-W	0.31	0.38	**0.45**	0.29	0.37	**0.44**
GTM-2	0.24	0.30	**0.41**	0.21	0.28	**0.41**
METEOR-pa	0.45	0.56	**0.64**	0.26	0.39	**0.51**
ULC	0.03	0.22	**0.41**	-0.03	0.19	**0.44**

Table 4: As in Table 2 for the French-German language pair.

evaluation for two public SMT systems for general translation: Bing[6] and Google[7]. These systems can be considered the state-of-the-art of a SMT open domain translator.

In general, our in-domain trained system performs significantly better than the two general purpose ones mainly because of two reasons. First, it has been trained on the specific domain and second, the tokenisation tools have been specifically developed to deal with chemical compounds. The concrete values can be read in Tables 2, 3 and 4 for the language pairs English-German, English-French and French-German respectively.

Even though the Domain system shows a good performance among SMT systems, some of the observed translation errors would not be produced by a rule-based system, which, on the other hand, would probably produce different ones. Table 5 displays two translations from German into English where this is made evident. In the first one, systems are not able to capture the different order in the verb position, although the translation is adequate lexically. The second sentence is an example of the importance of the chemical names. Google, for instance, tokenises the compound by the punctuation. Some of the tokens are then translated, but the full compound is not recovered. Bing and Domain do not tokenise the compound, but according to the results, the word does not appear in the training corpus and has not been translated. These kinds of errors can be easily alleviated by the GF grammar and are a motivation to combine GF and SMT for the translation of patents.

4 Hybridisation approaches

Hybrid approaches in MOLTO depart from three key assumptions when facing the combination of paradigms: 1) the quality of a completely translated sentence by a GF-based system will be always better than the translation obtained with SMT; 2) when the GF-based systems fails at producing a complete translation it can probably produce a set of partial translations (phrases) with confidence scores or probabilities; 3) the SMT system is always capable of generating an output translation. Assumption number one implies that our combination setting will be set as a fall-back strategy, i.e., in general SMT will be seen as a back-off for GF-based MT. Assumption number two makes it possible to combine partial outputs from GF with the SMT system in a real hybrid approach. Assumption number three guarantees that a translation will be always output by the combined system.

[6]http://www.microsofttranslator.com/

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

DE	Verwendung nach Anspruch 23 , worin das molare Verhältnis von Arginin zu Ibuprofen 0,60 : 1 beträgt .
EN	**The use** of claim 23 , wherein the molar ratio of arginine to ibuprofen **is** 0.60 : 1 .
Domain	The use of claim 23 , wherein the molar ratio of arginine to ibuprofen 0.60 : 1 .
Google	The **method** of claim 23 , wherein the molar ratio of arginine to ibuprofen 0.60 : 1 **is** .
Bing	Use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

DE	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
EN	(±)-N-(3-**a**minopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradeceneyloxy)-1-propanaminium **bromide**
Domain	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
Google	(±)-N-(3-aminopropyl)-N , N-dimethyl-2 , 3-bis (syn-9-tetradecenyloxy) is 1-propanaminiumbromid
Bing	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid

Table 5: Examples of wrong German-to-English translations in SMT systems. This kind of errors are not produced by the GF grammar for translating compounds.

Keeping these premises in mind we develop combination schemes to integrate grammar-based and statistical MT systems in a hybrid approach. We can divide the schemes in three big groups:

Hard integration: Force fixed GF fragment translations within a SMT system.

Soft integration led by SMT: Make available GF fragment translations to a SMT system.

Soft integration led by GF: Complement with SMT options the GF translation structure.

Each one of these options involves either the modification of the original systems or the construction of a new architecture. The most important thing in order to combine methodologies is that GF is able to parse general text robustly, it must be able to skip those structures not covered by the grammar and give some general information so that the statistical component of the engine takes care of the fragments. The first work on this task is the robust parser being developed for GF. Current experiments use shallow parsing as a first approximation and efforts are being made to increase the coverage.

Similarly, it is important that systems can share information. In order to make available GF translations to a SMT system one mainly needs to be able to feed an SMT decoder with translation pairs. GF translation pairs can be obtained by using its high quality alignments and extract the phrases in the SMT style. Since GF alignments are reliable, this will add a set of high quality phrases to be combined with those coming from the pure SMT system in the translation table. GF has been adapted for this purpose so that it is able to generate both alignments in the usual format[8] and with a text Giza-like nomenclature.

4.1 Ongoing work: robust GF with an extended lexicon

A first combination of GF and statistical methods is being developed for the English-to-French translation of patents. The kernel of the system is the Interlingua based translation

[8]Graphviz, an open source graph visualization software (http://www.graphviz.org/).

of Section 3.1. The system uses the patents grammar together with the resource grammar, builds automatically the lexicon from the English text and translates it into French. The GF translation mechanism is then applied on sentences that can be parsed, otherwise a chunker is used to fragment the sentence and only the parts of the sentence that can be handled by the grammar are translated. The other parts can be sent to the SMT system, or alternatively the SMT system can be fed with the phrases translated with GF.

5 Conclusions

One of the goals of the MOLTO project is to build a high-quality and robust translator for patents in at least three European languages: English, German and French. In order to achieve this purpose several systems are being developed. One of them is a multilingual rule-based translation system, and another one is a statistical translation system. Both of them depart from general systems and have been specialised into the patents domain. Besides, these two approaches will be merged to forge hybrid systems, and some ongoing work is devoted to build independent modules of the individual systems that can ease the integration.

The work is still in progress. The GF grammar needs a more thorough evaluation, in order to decide upon future extensions that would improve its coverage. A dedicated chunker in the three languages is also being built to divide claims and allow a separate treatment. And probably the most difficult step is to implement a disambiguation module that deals with the open language found in patents.

On the other hand, the in-domain SMT system already outperforms state-of-the-art general translation systems. The more advanced hybrids will combine the large coverage shown by SMT together with the capabilities of GF in generating grammatically correct translations.

Acknowledgements

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). The work was conducted using the Matrixware Research Collection, provided by IRF `www.ir-facility.org`. Authors are thankful to Aurélien Max and Xavier Auvray for their help with the French processing tools and the preliminary evaluation respectively.

References

[1] Giménez, J., and Màrquez, L. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, 94 (2010), 77–86.

[2] Koehn, P., Hoang, H., Mayne, A. B., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session* (Jun 2007), pp. 177–180.

[3] Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst,

E., and Moran, C. Open Source Toolkit for Statistical Machine Translation. Tech. rep., Johns Hopkins University Summer Workshop. http://www.statmt.org/jhuws/, 2006.

[4] Och, F. J. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics* (Sapporo, Japan, July 6-7 2003).

[5] Och, F. J., and Ney, H. A systematic comparison of various statistical alignment models. *Computational Linguistics 29*, 1 (2003), 19–51.

[6] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics* (2002), pp. 311–318.

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

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

[9] Stolcke, A. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing* (2002).

[10] Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics.*, P. Bozanis and e. Houstis, E.N., Eds., vol. 3746. Springer Berlin Heidelberg, 2005, p. 382–392.