Statistical Machine Translation prototype using UN parallel documents

Bruno Pouliquen, Christophe Mazenc
World Intellectual Property Organization
Global Databases Service

34, chemin des Colombettes
CH-1211 Geneva 20
Bruno.Pouliquen@wipo.int,
Christophe.Mazenc@wipo.int

Cecilia Elizalde, Jose Garcia-Verdugo
United Nations
Department for General Assembly
and Conference Management
Documentation Division, Spanish Translation Service
405 42nd Street
New York NY 10017
elizalde@un.org,
garcia-verdugo@un.org

Abstract

This paper presents a machine translation prototype developed with the United Nations (UN) corpus for automatic translation of UN documents from English to Spanish. The tool is based on open source Moses technology and has been developed by the World Intellectual Property Organization (WIPO). The two organizations pooled resources to create a model trained on an extensive corpus of manually translated UN documents. The performance of the SMT system as a translation assistant was shown to be very satisfactory (using both automatic and human evaluation). The use of the system in production within UN is now under discussion

1 Introduction

This paper describes a prototype for the automatic translation of United Nations documents[1].

The tool has been the subject of experiments within the United Nations, including a structured human evaluation carried out by three professional translators.

The number of documents translated by this UN Division per year is 33,670 (90 million words) in the six official languages (Arabic, Chinese, English, French, Russian and Spanish).

The UN has an extensive parallel corpus of high-quality human translations collected from 2000 to 2011 for all language combinations, since the norm is that parliamentary documents are to be translated to all the six official languages and issued simultaneously.

Quality is a paramount consideration for the translation of parliamentary documents at the UN: translators are highly skilled professionals, 50% of the translations are revised by a senior reviser, 50% are subject to self-revision, afterwards 80% of translations are subject to additional proofreading or scoping[2].

Due to the growing demand for translations and budgetary considerations, the percentage of contractual translation (currently 20%) is bound to increase. However, contractual translators do not have access to the same document and terminology databases and IT tools as internal staff, and therefore the quality of their translations suffers.

The UN documents submitted for translation in New York deal with a great diversity of subjects, including 10%-15% of documents relating to budgetary and administrative issues that are good candidates for computer-assisted translation because they contain around 30% of repetitive language.

UN translators have been exposed to machine translation through Google Translate (either directly or through CAT tools) and have found that the output quality, for the purposes of the translation of UN documents, has been decreasing over the years as documents from other organizations were added[3]. Their expectation is to explore the possibilities of a SMT tool trained only with UN documents. There is also the expectation to im-

[1] Documents provided by the Documentation Division (New York) of the Department for General Assembly and Conference Management, the main entity of the United Nations Secretariat charged with the production of parliamentary documentation. The Documentation Division in New York deals with the translation of parliamentary documentation.

[2] A lighter proofreading where only numbers, titles and number of paragraphs are checked.

[3] Google used UN documents to train its MT tool, http://www.reuters.com/article/2007/03/28/us-google-translate-idUSN1921881520070328

prove the quality and consistency of the contractual translation by providing contractors with the same toolkit as internal staff and/or applying MT and post-editing.

In parallel, WIPO has already developed such a SMT with a similar-sized corpus (called WIPO-COPPA[4], see Pouliquen & Mazenc, 2011)

A preliminary test was launched using the WIPO tool (described in Pouliquen et al, 2011) in which: a statistical machine translation (SMT) system was trained using the UN corpus in order to evaluate the quality of such a tool (especially in comparison with other tools).

UN Spanish Translation Service (STS) has been exposed to MT (mainly from English to Spanish), rule-based systems required too much work to adapt their own terminology while SMT-based systems like Google/Bing/Language Weaver yield good results. It was decided to launch an experiment with this language pair. UN STS gave WIPO access to their 64,619 English-Spanish documents.

2 Context/State of the art

At the UN, due to the large volume of translated pages and recent budgetary restrictions, there is a growing demand to decrease costs and increase throughput by leveraging IT tools as applied to translation and to increase quality and consistency, in particular for the jobs translated by contractual translators. Most in-house translators type or dictate their translations and look for information in monolingual and bilingual document databases, and terminology databases. At the Spanish Translation Service, around 25% of the documents are prepared using CAT tools[5]. Translators have been exposed to SMT and have expressed interest in including this technology in their regular toolkit.

Various techniques can be used in Machine Translation (Koehn 2010): rule-based systems, example-based translation, statistical machine translation and hybrid systems.

An international organization like the UN has 6 working languages (plus German), which means that, if such an organization wanted a translation tool in all language pair combinations, it would require 42 translation engines. A rule-based translation system would be extremely costly to build and maintain. A data-driven approach is usually more suitable when a big parallel corpus exists.

Some UN parallel corpora are already available on the Web: UN Corpora[6] (Rafalovitch et al. 2009) provides a 3.5 million word corpus which contains only a part of the General Assembly Resolutions for eight sessions only and has not been updated. Multi UN[7] (Eisele et al. 2010) has built a more extensive corpus of 370 Million words however this corpus is now outdated (up to September 2010) and not sentence-aligned.

In December 2011, the validity of a 1994 agreement with LDC was reconfirmed. The Linguistic Data Consortium (see Graff 1994) will make an updated UN corpus available for research purposes.

For the purpose of the current experiment, the Spanish Translation Service (STS) provided its full collection of English-Spanish bitexts from 2000 to 2011, composed of 64,619 documents (equivalent to about 220 million words).

With this high-quality parallel corpus, SMT was chosen, with a flexible and free engine: *Moses* (Koehn et al. 2007).

2.1 The English-Spanish parallel corpus

The SMT is trained with a parallel corpus extracted from previously translated UN documents from 2000 up to December 2011 (62,757 English-Spanish documents after filtering[8]). The provided corpus is extracted from HTML bitext files[9]. We chose to re-align every text as WIPO's

[4] English-French Patent corpus of 170 Million words Freely available for research purpose at http://www.wipo.int/patentscope/en/data/products.html#coppa

[5] UN translators use mainly SDL products with file-based translation memories. The UN is currently developing its own web-based computer-assisted translation, referencing and terminology tool in the context of a global project called *gText*, using internal developers

[6] http://www.uncorpora.org/

[7] http://www.euromatrixplus.net/multi-un/

[8] The documents all originated from UN headquarters (New York), more documents can be included in the future from UN-Geneva and UN-Vienna. We filtered out documents not in the right language or having an unrecognized format.

[9] UN document division has a simple script that matches pairs of documents with the selected language pair and a commercial alignment robot that generates the corresponding HTML table. The robot alignment algorithm relies heavily on document formatting (Microsoft Word 97/2000 format) and automatically discards document pairs that exceed a specific misalignment threshold. The resulting bitexts contain a significant amount of misaligned segment,

aligner is tailored for machine translation and produces cleaner alignments.

Starting with this material, we tried to build a reasonably clean bilingual corpus by applying the following steps (some of the cleaning techniques were successful in previous WIPO experiments):

- carrying out sentence splitting of documents (using a home-made splitter, based on sentence boundaries and a list of abbreviations)
- tokenizing each sentence (using a home-made tokenizer based on *Lucene* framework[10])
- using *Champollion* (Ma 2006) to align sentences, we developed a Java version which allows to split further long sentences (having more than 80 words). The tool uses a bilingual dictionary which we extract from previously extracted model.
- computing an "aligned-segment-matching-score" for each aligned segment (taking into account a previously learned bilingual dictionary)
- filtering out whole documents having an average-segment-matching-score below a given threshold (empirically set to 0.15)
- applying a smooth filter on the segment-matching-score ([0.1,0.2,0.4,0.2,0.1]) which will "propagate" the score of a segment to the adjacent ones, filtering out the segments having a "smooth" score lower than a second threshold (empirically set to 0.3)
- filtering out sentences having more than 80 words[11] (or only one word)
- filtering out pairs of sentences where the ratio (number of English words/number of Spanish words is more than 9)
- applying some regular expression replacement rules (deleting xml tags, uniform accents, etc.)

As a result 10,251,816 aligned pairs of segments were obtained (210 Million words in English, 240 Million in Spanish). The quality of alignment is reasonable, however attempts should always be made to improve the quality in the future.

2.2 Training the model

Moses can be trained using our parallel corpus.

2,000 segments were retained as our development set in order to carry out the optimization, (see section 3.3). As the documents are big, only two documents were part of the development set, it was decided to keep these two documents apart for the training. A first test set was selected as a random selection of 1000 segments out of remaining segments of these two documents.

Mgiza++ (Gao & Vogel, 2008) was used to align words in sentences. On a *Linux* server (48 cores of 2.5Ghz) it ran for 2.5 days on the corpus. We then stored this information in a *Lucene* index so we can use it for our concordancer (see section 3.2) or as a translation memory (TM) index.

The language model is built using SRILM toolkit (Stolke, 2002) with 5-grams. The model is generated out of the Spanish texts of the corpus (239,424,105 words).

2.3 Optimization

Attempts to optimize the performance of the system with various settings were carried out:

The generated phrase table contains 272 million entries, in such a huge table, some phrases are very unlikely to be seen in other documents. A decision was taken to try to "prune" this phrase table (in such a big corpus a phrase that occurs only once has a high probability of being useless and even erroneous). The "pruning" method as described in Johnson et al. (2007) was used with the suggested parameters i.e.: delete all phrases which occur only once in our training corpus and, for each phrase, only the first 30 translation candidates are kept. The "pruned" phrase table now contained 50 million entries (19% of the original). The speed of the translation improved and, as expected, the quality improved as well (see line "pruned" in Table 1).

The reordering model (originally containing 272 million entries) was also filtered using two criteria: the source and target ngrams were kept only if they appear in the "pruned" phrase table (resulting in 50 million entries only) and only if source and target contained less than 9 words in total (resulting in 27 million entries, this last criterion is arguable, however the differences in BLEU/METEOR scores with and without this filter are negligible while the size of the reordering model is considerably reduced, see the differ-

as well as up to 30% of segments containing no text at all (mainly figures, formatting elements and symbols).

[10] We used the standard tokenizer (McCandless, 2010) and updated it so that it recognizes email addresses, internet hostnames, URLs, XML tags, references, Greek letters, apostrophes, etc.

[11] This filter is perhaps too aggressive but the word alignment speed (and quality) will usually be poor on big sentences.

ences between line “pruned” and “prunedmax4” in table 1).

An optimization of the settings by maximizing the BLEU score on the development set (2000 segments) was carried out using the minimum error rate training (MERT).

2.4 Preliminary results

A very first BLEU score (see section 5.1) of 65.45 was quite encouraging but not reliable as it was computed on a non-representative test set (remaining sentences of the development set).

3 Translating / graphical user interface

3.1 Server configuration

We chose to set up an architecture that allows:

- various users to work at the same time
- various alternative translations
- word alignment

The *Moses* decoder is slightly modified in order to output the first 24 proposals for each submitted translation. Each decoder is encapsulated in a Java *RMI* interface server which allows the running of several concurrent decoders (on the same or on different language pairs). Each sentence submitted is queued and sent to the next free decoder.

Our phrase tables are so big (even after the “pruning”) that it is impossible to store them in memory, we store them on disk, even so the server gives good performance. The phrase tables keep the word alignment information so that users can highlight translated words in a sentence.

The server includes a *post-processor* that deletes unnecessary spaces and recases the output taking into account the input (functionality to be improved in the future).

3.2 Graphical user interface

We set up a Java Server Faces[12] Web interface to connect to the translation server. Users can interactively get translations of new documents. Alternatively they can also verify the segment previously translated through a concordancer. The interfaces were designed to be intuitive. We used Moses’ “keep alignment” functionality so that word translations are highlighted (as well as parallel segments), so that the user can immediately spot the good/bad translations.

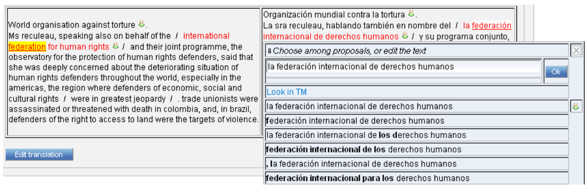
3.2.1 First Web interface: gist translation

A first Web interface allows user to submit short texts and access the corresponding automatic translation (with highlighting of parallel segments/words).

Users can access alternative translation proposals by clicking on a given segment.

However, Moses’ mechanism limits the proposal alternatives when the sentence is too long, in such cases, the user can select a chunk of source text and gets alternatives for this new segment.

A small icon indicates to the user that a segment has already been translated (is part of the TM) and links him to the concordancer (see 3.2.3)



World organisation against torture.
Ms reculeau, speaking also on behalf of the / international federation for human rights / and their joint programme, the observatory for the protection of human rights defenders, said that she was deeply concerned about the deteriorating situation of human rights defenders throughout the world, especially in the americas, the region where defenders of economic, social and cultural rights / were in greatest jeopardy / . trade unionists were assassinated or threatened with death in colombia, and, in brazil, defenders of the right to access to land were the targets of violence.

Organización mundial contra la tortura.
La sra reculeau, hablando también en nombre del / la federación internacional de derechos humanos / y su programa conjunto,

⇓Choose among proposals, or edit the text

la federación internacional de derechos humanos

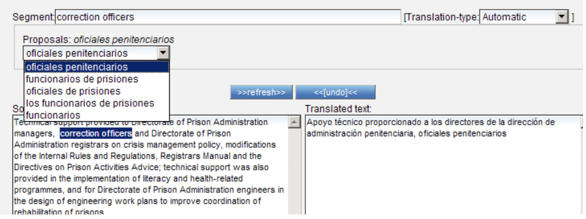
la federación internacional de derechos humanos
federación internacional de derechos humanos
la federación internacional de **los d**erechos humanos
federación internacional de los derechos humanos
, la federación internacional de derechos humanos
federación internacional para los derechos humanos

Figure 1: Gist translation, highlighting parallel segments and words (here *federación/federation*), user can access alternative translations for a given segment, the green icons are an indicator that the segment is part of the TM.

This web interface can be tested on the WIPO website, but is only suitable for Patent texts: http://www.wipo.int/patentscope/translate

3.2.2 Second Web interface: interactive translation

An alternative graphical user interface lets the user segment the text to be translated. With this interface the translator drives the translation by providing the segments he wants to translate. He can then immediately select alternative proposals.



Segment: correction officers [Translation-type: Automatic]

Proposals: *oficiales penitenciarios*

oficiales penitenciarios
oficiales penitenciarios
funcionarios de prisiones
oficiales de prisiones
los funcionarios de prisiones
funcionarios

Technical support provided to Directorate of Prison Administration managers, correction officers and Directorate of Prison Administration registrars on crisis management policy, modifications of the Internal Rules and Regulations, Registrars Manual and the Directives on Prison Activities Advice; technical support was also provided in the implementation of literacy and health-related programmes, and for Directorate of Prison Administration engineers in the design of engineering work plans to improve coordination of rehabilitation of prisons

Translated text:

Apoyo técnico proporcionado a los directores de la dirección de administración penitenciaria, oficiales penitenciarios

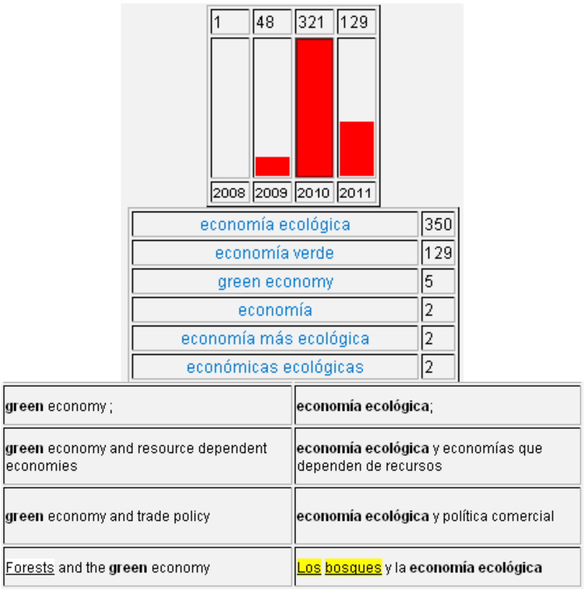
Figure 2: Interactive translation interface, user highlights the next source segment to translate and can select alternative translations

3.2.3 Concordancer

Users can access the concordancer using a Web interface. The concordancer is based on a *Lucene* index containing the information result of the word alignment (using *grow-diag-final-and file – (*Koehn 2010 p. 118)). This concordancer displays the segments containing the search term

[12] http://javaserverfaces.java.net/

and the corresponding aligned words. A first window displays the usage of the term by year, a second window displays the aligned words by order of frequency, so the user can immediately see which translation is the most common (see Figure 3 for an example).



1	48	321	129
2008	2009	2010	2011

economía ecológica	350
economía verde	129
green economy	5
economía	2
economía más ecológica	2
económicas ecológicas	2

green economy ;	**economía ecológica**;
green economy and resource dependent economies	**economía ecológica** y economías que dependen de recursos
green economy and trade policy	**economía ecológica** y política comercial
Forests and the **green** economy	Los bosques y la **economía ecológica**

Figure 3: Concordancer for term "green economy", the top graphic shows the term usage over years, then the most used translations, then the parallel segments with a link to the corresponding document.

4 Results/evaluation

4.1 Automatic evaluation

The BLEU and METEOR scores (Papineni et al. 2002, Denkowski & Lavie 2011) were used to compare human translation with automatic translation. It was decided to launch an automatic evaluation on a second test set: a random selection of 1,000 segments of new documents (i.e. documents published in January 2012, only one reference translation per document).
Table 1 gives some of our BLEU/METEOR scores according to specific experiments.

Table 1: Speed and scores computed using various configurations and tools (second test set, 1000 segments from "non-repetitive" documents published in January 2012)

Experiment	Speed seconds /segment	Model size[13]	BLEU	METEOR
Baseline	7.60	69G	47.06	60.79
Pruned	6.39	12G	47.34	61.17
PrunedMax4	6.20	7.8G	47.31	61.23
Google translate[14]	n/a	n/a	39.99	54.96
Bing translator[15]	n/a	n/a	38.20	53.81
Mert optimized	6.30	7.8G	47.87	61.34

Reading Table 1 gave us a good idea on how well the system was performing. It performed better than the two publically available commercial tools. The speed was acceptable even if the models are stored on disk (pruning our models gave better or equal scores, while improving the size and speed of the models).

However human evaluators said that this second test set was belonging to a category of "non-repetitive" documents. The scores are much lower than our first BLEU score (65.45 on the first test set, see section 2.4). The reason was that the first BLEU was calculated with a set containing many segments from a Security Council resolution, which is a repetitive document, while the second test set contained narrative reports, non-cyclic reports, and documents non-related to the parliamentary processes of the Secretariat.

Then we decided to ask UN translators for examples of such "repetitive" documents, they gave us 13 new documents, containing 786 parallel segments, and the scores were much higher as shown on Table 2. These documents contained administrative and internal reports generated at the Secretariat, usually related to an administrative cycle, including budgetary and audit cycles, as well as Security Council and General Assembly resolutions. In general, these repetitive documents, as well as resolutions, are translated closer to the English version to keep parallelism, which in turn helps parliamentary negotiations,

[13] The model size is the size of the "binarized" phrase table and reordering model with option *alignment-info* (http://www.statmt.org/moses/?n=Moses.AdvancedFeatures#ntoc2). The language model, not included, represents 1.7G.

[14] http://translate.google.com/ (February 2012)

[15] http://www.microsofttranslator.com (February 2012)

while what we called “non-repetitive” documents are translated in a much freer writing style.

According to some estimations done by the Documents Control Unit, around 30% of the documents translated in New York have some degree of reprise, which might make them suitable for MT.

Table 2: BLEU/METEOR scores on "repetitive documents" (third test set: 786 segments)

System	BLEU	METEOR
Baseline	77.25	84.35
Google translate	59.70	73.77
Bing translator	58.49	73.69
Mert optimized	79.40	85.47

This BLEU score of 79.40 is quite impressive. A first general conclusion is that two systems have to be compared on exactly the same test set as shown by high differences of the three computed scores (47.87, 65.45 and 79.40).

4.2 Human evaluation of automatic translation

4.2.1 Preliminary evaluation

The first subjective evaluation was done by UN Spanish Translation Service in January 2012, translating a real job, a performance report of a peacekeeping mission (A/66/602) with the MT system. In order to verify the accuracy with the terminology produced by the MT system as compared to the mandatory terminology databases for this category of document, the job was translated using automatic terminology recognition (Mutiterm) The overall evaluation was that the automatic translation output was very good, in particular because the terms were accurate and consistent with the official terminology, and typing was significantly reduced. Even if most sentences needed reworking during post-editing, some were totally satisfactory.

One of the translators who were assigned to translate this document was a new recruit and her output was subject to revision. The reviser found that the quality of the translation was above average for a new recruit translating this challenging category of documents, as the terminology was consistently used, the meaning was accurate and the style was adequate. In the subjective opinion of the reviser, this might be an indication that for some categories of documents, the use of MT could help new recruits to produce translations better aligned with internal stylistic preferences and terminology; this would also apply for contractors, who do not have access to the same document and terminology resources as the internal staff. This hypothesis must be further explored and validated with relevant tests.

Some other evaluations were done with different categories of documents, as notes for the President of body/organism sessions (very good quality), intranet news (poor quality) and administrative reports (good quality). As expected, a statistical machine translation tool trained with UN documents is not useful for translating all categories of documents, but a significant amount of them, in particular those that are included in the training and have some specific styles and terminology.

4.2.2 Set-up of the test

A second structured evaluation was done with three human evaluators, using the second test set. We knew that it was a “difficult” test set, however the output of human evaluation on such difficult test set is maybe more objective than on an easy one (as the third test set with close to 80 BLEU score). The three evaluators were chosen by the Chief of the Spanish Translation Service for their professionalism and were translators with more than 20 years of professional experience each. The evaluation was conducted over three full days. We have chosen to evaluate the translations using the known metrics: fluency and adequacy (see for example Denkowski & Lavie 2010).

Fluency rates how good the output Spanish is (using the following scale 5: Flawless 4: Good 3: Non-native 2: Disfluent 1: Incomprehensible) and adequacy rates the amount of information that has been transferred between original English and the Spanish translation (using the scale 5: All 4: Most 3: Much 2: Little 1: None).

At the time this evaluation was done the recasing was not working properly (partially fixed in later version), therefore we asked the evaluators to ignore case (‘*naciones unidas*’ – in lowercase – is considered as a good translation for ‘*United Nations*’).

4.2.3 Results

The three experts blindly (i.e. ignoring others’ judgments) evaluated the translation of the 1,000 segments (same segments as on Table 1). We decided not to display the reference translation, in order not to influence the judgment of the experts. A specific Web interface was built.

The experts had a minimal training on the evaluation tool and discussed about how to interpret and apply the metrics beforehand. In fact the three experts often agreed on the scores (when we compute the maximum disagreement score between the average on one evaluation, the overall average –on the 1000 evaluations– is 0.65 only).

On average the fluency is 3.94 the adequacy is 4.28.

Evaluators agreed on the final score, most of the content is maintained in the translation (adequacy more than 4), the fluency of the translation is almost "good" (fluency 3.94).

4.2.4 Feedback

The Spanish translators who participated in both evaluations as well as in other individual and informal tests found that the overall quality of the MT prototype output was good in general and very good for some specific categories of documents (for instance, peacekeeping budgets, as the ones used in Table 2), where a large volume of similar documents were included in the training. However, these particularly good MT documents were not included in the structured test. According to the feedback provided by some evaluators, the sentences included in the human evaluation were not the most repetitive and formulaic. For this reason, the use of domains might be advisable in the future. Although it is practically impossible to automatically sort the New York documents by categories using the UN symbol (an alphanumeric ID contained in all documents issued to the Official Document System[16]).

The Concordancer interface was used by the Senior Terminologist of the Spanish Translation Service, who also served as human evaluator, and she found that it was very useful to validate terminology records, despite some bugs in the current version.

Translators in other duty stations, including Vienna, Geneva and Santiago, were aware of the interfaces and were encouraged to try them. An additional training using Vienna and Geneva documents is expected at a later stage (these duty stations deal with a more limited and consistent set of subjects, so the duty station could be used as a proxy for domains).

According to the feedback received from the Chief of STS and other staff members, some translators and revisers are already using the tool for real jobs, in particular for some categories of documents, including Security Council and peacekeeping. In their opinion, the quality of the output of the system is very high and lends itself for post-editing. These translators and revisers appreciate that the terminology is consistent with UN terminology and style norms. In this respect, the feedback is particularly positive from senior revisers. In effect, as they are used to revision and are familiar with UN standards, they find useful to work with MT and post-editing. Some other translators are using the system in combination with Trados and also report very high satisfaction.

As per the feedback of some UN users, in some categories of documents, the output of MT allows translators to speed-up the translation process. However, they report that this requires a different intellectual effort that is similar to revision but still more intensive, as in some cases, the system might produce sentences with high fluency but low accuracy (for instance, grammar is acceptable but the meaning is transferred partially or not at all). Translators and managers agree that further evaluations need to be done in order to validate the benefits of MT in productivity and quality, as well as to determine the threshold of usability of MT for post-editing. There is a strong interest from translators in STS in developing a bridge between the system and CAT tools (Trados and Mercury), as well as to develop a service to translate full documents. Finally, it is important to note that as a result of this experiment, the scope of *gText*, a current global project to develop terminology, reference and CAT tools for all UN duty stations, was expanded to include also the development of machine translation systems for all the UN official languages.

5 Conclusion and future work

We had to face a scalability problem with such a big corpus. However WIPO had already successfully trained a similar scale model. This experience shows that open source solutions can sometimes provide better results than generic commercial products. The data-driven approach requires limited human resource and still provides good results. It is planned to launch similar experiments with other language pairs: English-Russian, English-Chinese and English-Arabic. We expect worse results as it is more challenging than translating from English to Spanish or French due to the highly different morphological structure of the languages.

[16] ods.un.org

In such an experiment the final word should always be left to the final users from UN.

They judged the Web interface as intuitive and requiring very little training. An integration with existing CAT tools is already on the way.

Future work includes: (a) testing of effort/productivity gains of MT and post-editing in some categories of documents and its use in conjunction with CAT tools (as in the experiment done by Plitt & Masselot, 2010), (b) testing the system with other language pairs (c) improving the user interface and (d) integrating with third party products.

Acknowledgements

The authors would like to thank the following managers and staff members of the Department for General Assembly and Conference Management of the United Nations Headquarters in New York for supporting the project: Mr. Shaaban M. Shaaban, Under-Secretary-General, Mr. Franz Baumann, Assistant Secretary-General, the members of the Departmental Management Group, the Information and Communications Technology Committee and the Technology Advisory Group of the Documentation Division, Ms. Maria Nobrega, Chief of the Spanish Translation Service, Ms. Ana Larrea, Training Officer of the Spanish Translation Service, and very specially, Igor Shpinov, who supported this project enthusiastically since its very beginning and facilitated its authorization. The project would not be possible without the dedicated collaboration and open-mindedness of Ms. Maria Barros, Ms. Rosario Fernandez and Ms. Carla Raffo, who served as human evaluators of the system and the staff of the Spanish Translation Service who provided constant feedback. .

Thank you to Laurent Gottardo for his idea in presenting bars per year in the concordancer. Special thanks to Paul Halfpenny for his valuable proof-reading.

References

Denkowski, Michael & Alon Lavie. 2011. "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems", Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation

Denkowski, Michael & Alon Lavie. 2010. "Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks", *Proceedings of AMTA*

Eisele, Andreas & Yu Chen. 2010. MultiUN: a multilingual corpus from United Nation documents. LREC 2010: proceedings of the seventh international conference on Language Resources and Evaluation, 17-23 May 2010, Valletta, Malta; pp.2868-2872.

Gao, Qin & Stephan Vogel. 2008. Parallel implementations of word alignment tool. In Proceedings of the ACL'08 Software Engineering, Testing, and Quality Assurance Workshop

Graff, David. 1994. UN Parallel Text (Complete). Linguistic Data Consortium, Philadelphia.

Johnson, Howard, Joel Martin, George Foster, Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. EMNLP-CoNLL 2007: 967-975

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris C. Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In Proceedings of ACL 07. Morristown, NJ, USA, 177-180.

Koehn, Phillip. 2010. Statistical Machine Translation. textbook, Cambridge University Press, January 2010.

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. Proceedings of LREC-2006

McCandless, Michael, Erik Hatcher, Otis Gospodnetić. 2010. Lucene in Action. 2nd Edition. Manning Press.

Papineni, K., S. Roukos, T. Ward, and WJ Zhu. 2002. BLEU: a method for automatic evaluation of machine translation, proc. of ACL 2002, pp. 311-318

Pouliquen, Bruno & Christophe Mazenc, 2011, COPPA, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO. *MT Summit XIII: the Thirteenth Machine Translation Summit*, 19-23 September 2011, Xiamen, China; pp.24-30

Pouliquen, Bruno, Christophe Mazenc & Aldo Iorio: Tapta: a user-driven translation system for patent documents based on domain-aware statistical machine translation. P*roceedings of the 15th conference of the European Association for Machine Translation*, 30-31 May 2011, Leuven, Belgium; pp.5-12

Plitt, M. & F. Masselot. 2010. A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics, 93(01/2010): p 7-16

Rafalovitch Alexandre & Robert Dale. 2009. United Nations general assembly resolutions: a six-language parallel corpus. MT Summit XII: proceedings of the twelfth Machine Translation Summit, 26-30/08/2009, Ottawa, Ontario, Canada; pp.292-299.

Stolke, Andreas. 2002. SRILM an extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing.