Machine Translation of Legal Information and Its Evaluation

Atefeh Farzindar
NLP Technologies Inc.
3333 Queen Mary Road, suite 543
Montréal, Québec, Canada, H3V 1A2
farzindar@nlptechnologies.ca

Guy Lapalme
RALI-DIRO
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
lapalme@iro.umontreal.ca

Abstract. This paper presents the machine translation system known as TransLI (**Trans**lation of **L**egal **I**nformation) developed by the authors for automatic translation of Canadian Court judgments from English to French and from French to English. Normally, a certified translation of a legal judgment takes several months to complete. The authors attempted to shorten this time significantly using a unique statistical machine translation system which has attracted the attention of the federal courts in Canada for its accuracy and speed. This paper also describes the results of a human evaluation of the output of the system in the context of a pilot project in collaboration with the federal courts of Canada.

1. Context of the work

NLP Technologies[1] is an enterprise devoted to the use of advanced information technologies in the judicial domain. Its main focus is *DecisionExpress*™ a service utilizing automatic summarization technology with respect to legal information. *DecisionExpress* is a weekly bulletin of recent decisions of Canadian federal courts and tribunals. It is an tool that processes judicial decisions automatically and makes the daily information used by jurists more accessible by presenting the legal record of the proceedings of federal courts in Canada as a table-style summary (Farzindar *et al.,* 2004, Chieze *et al.* 2008). NLP Technologies in collaboration with researchers from the RALI[2] at Université de Montréal have developed TransLI to translate automatically the judgments from the Canadian Federal Courts. As it happens, for the new weekly published judgments, 75% of decisions are originally written in English

[1] http://www.nlptechnologies.ca
[2] http://rali.iro.umontreal.ca

and 25% in French. By law, the Federal Courts have to provide a translation in the other official language of Canada.

The legal domain has continuous publishing and translation cycles, large volumes of digital content and growing demand to distribute more multilingual information. It is necessary to handle a high volume of translations quickly.

Currently, a certified translation of a legal judgment takes several months to complete. Afterwards, there is a significant delay between the publication of a judgment in the original language and the availability of its human translation into the other official language.

Initially, the goal of this work was to allow the court, during the few months when the official translation is pending, to publish automatically translated judgments and summaries with the appropriate caveat. Once the official translation would become available, the Court would replace the machine translations by the official ones. However, the high quality of the machine translation system obtained, developed and trained specifically on the Federal Courts corpora, opens further opportunities which are currently being investigated: machine translations could be considered as first drafts for official translations that would only need to be revised before their publication. This procedure would thus reduce the delay between the publication of the decision in the original language and its official translation. It would also provide opportunities for saving on the cost of translation.

We evaluated the French and English output and performed a more detailed analysis of the modifications made to the translations by the evaluators in the context of a pilot study to be conducted in cooperation with the Federal Courts.

This paper describes our statistical machine translation system, whose performance has been assessed with the usual automatic evaluation metrics. We also present the results of a manual evaluation of the translations and the result of a completed translation pilot project in a real context of publication of the federal courts of Canada. To our knowledge, this is the first attempt to build a large-scale translation system of complete judgments for eventual publication.

2. Methodology

NLP Technologies' methodology for machine translation of legal content consists of the following steps:

- Translated judgments are gathered;
- The HTML markup is removed from the judgments, which are then aligned at the level of the sentence;
- a translation model is created using the pairs of translated sentences;
- The court tests the usability of the Statistical Machine Translation (SMT) in the context of a pilot project;
- The SMT is then deployed.

In the context of our project, NLP Technologies in collaboration with RALI used the existing translated judgments from the Federal Court of Canada as a training corpus for our SMT system. The next section provides more details on the translation system:

3. Overview of the system

We have built a phrase-based statistical translation system, called TransLI (Translation of Legal Information), that takes as input judgments published (in HTML) on the Federal Courts web site and produces an HTML file of the same judgment in the other official language of Canada. The architecture of the system is shown in Figure 1.

The first phase (semantic analysis) consists in identifying various key elements pertaining to a decision, for instance the parties involved, the topics covered, the legislation referenced, whether the decision was in favor of the applicant, etc. This step also attempts to identify the thematic segments of a decision: **Introduction**, **Context**, **Reasoning** and **Conclusion** (see section Evaluation in a pilot project). During this phase, the original HTML file is transformed into XML for internal use within NLP Technologies in order to produce DecisionExpress™ fact sheets and summaries. We extract the source text from these structured XML files in which sentence boundaries have already been identified. This is essential, since the translation engine works sentence by sentence.

The second phase translates the source sentences into the target language using SMT. The SMT module makes use of open source modules GIZA++ (Och and Ney, 2003) for creating the translation models and SRILM for the language models. We considered a few phrase-based translation engines such as Phramer (Olteanu et al, 2006), Moses (Koehn et al., 2007), Pharaoh (Koehn, 2004), Ramses (Patry et al., 2006) and Portage (Sadat et al., 2005). Moses was selected because we found it to be a state-of-the-art package with a convenient open source license for our testing purposes.

The last phase is devoted to the rendering of the translated decisions in HTML. Since the appropriate bookkeeping of information has been maintained, it is possible to merge the translation with the original XML file in order to yield a second XML file containing a bilingual version of each segment of text. This bilingual file can then be used to produce an HTML version of the translation, or for other types of processing, like summarization.

Indeed, since summaries of judgments produced by NLP Technologies are built by extracting the most salient sentences from the original text, producing summaries in both languages should be as simple as selecting the translation of every sentence retained in the source-language summary.

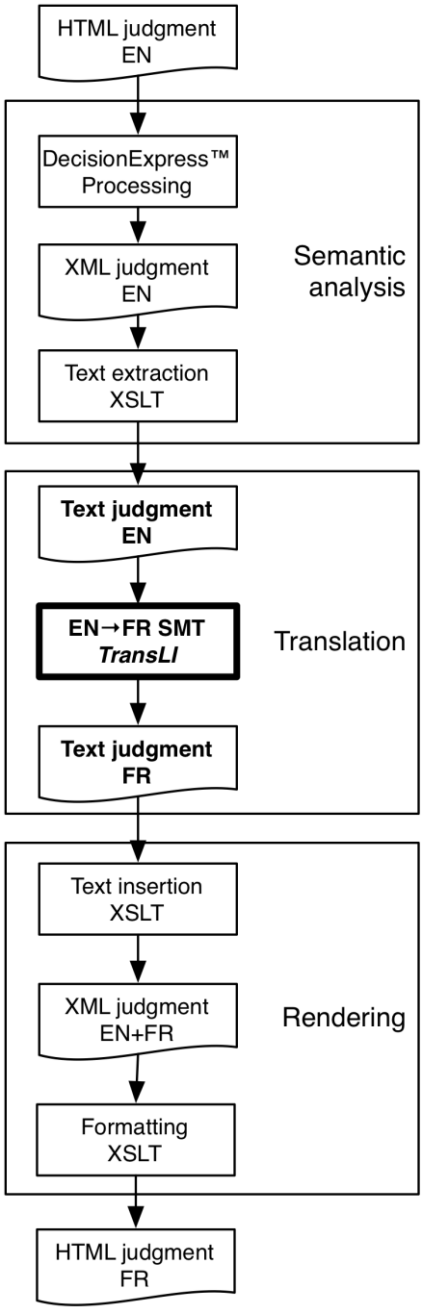

Fig 1: The translation pipeline translates an HTML court decision written in English into a French decision (also in HTML). A similar pipeline performs translations from French to English

Corpus name	# sent pairs	# en words	# fr words
principal	245,000	6,510,000	7,510,000
train	244,000	6,500,000	7,500,000
tune-1	300	8,000	9,000
test	1300	28,000	33,000
tune-recent	400	8,000	10,000
train-lexum	1,000,000	22,340,000	25,720,000

Table 1: Corpora used for developing TransLI

Gotti et al. (2008) describe the development and the testing of the TransLI statistical machine translation system. The final configuration is a compromise between quality, ease of deployment and maintenance and speed of translation with the following features: a distance based reordering strategy, a tuning corpus based on recent decisions; a large training corpus and the integration of specialized lexicons.

Although these types of texts employ a specialized terminology and a specific cast of sentences, the availability of large amounts of high quality bilingual texts made it possible to develop a state-of-the-art SMT engine. These excellent results prompted us to perform a human evaluation also described in (Gotti et al. 2008) on 24 randomly selected sentences from our test set. This evaluation centered on the quality of the produced translation and on its fidelity, i.e. to what extent the SMT conveys all the semantic content of the original.

A key element in the success of an SMT system lies in the availability of large corpora of good quality. In the Canadian judicial domain, we are fortunate enough to have access to public web sites providing translations of excellent quality for almost all judgments of the most important Canadian courts. For our work, we built a set of corpora, the characteristics of which are shown in Table 1.

`principal`: we downloaded 14,400 decisions in HTML from the Federal Court of Canada web site[3] from which we extracted the text. Because many judgments did not have a translation or could not be parsed automatically with our tools because of inconsistent original formatting, we ignored them and we were left with 4500 valid judgment pairs. From these pairs, we extracted the sentences and aligned them to produce a bi-text of around 260,000 sentence pairs. A number of them had English citations in the French text and vice-versa. Once these cases were filtered out, we were left with 245,000 sentence pairs.

`train`: 99% of the sentences from `principal`, used to train the SMT system.

`tune-1`: 1% of `principal` used to adjust the parameters of the system. There is no overlap with `train`.

`test`: 13 recent decisions that were published after the decisions occurring in `principal`. This better simulates the application context for our system, which will be used for translating recent decisions.

[3] decisions.fct-cf.gc.ca/en/index.html

`tune-recent`: 6 recent decisions that were published after the decisions in `principal`.

`train-lexum`: Since the RALI has a long experience in dealing with judicial texts in collaboration with the Lexum[4] at the Université de Montréal in the context of the TransSearch[5] system, we decided to add 750,000 bilingual sentence pairs from our existing bilingual text database. These sentences are taken from decisions by the Supreme Court, the Federal Court, the Tax Court and the Federal Court of Appeal of Canada.

For the quality of language, we asked three evaluators to assign each of the 24 passages a score: 1 (unacceptable), 2 (bad), 3 (fair), and 4 (perfect), according to whether they found it to be in a correct and readable target language, independently of the source language. This would correspond to the case where a non-French speaking person wanted to consult an English translation of a French text. Our evaluators did not know which translations had been produced by a human or which were produced by a machine.

The same three evaluators were given groups of two or three sentences containing the source French text and the English translation produced either by TransLI or by a human translator (the reference text). The evaluators were asked to modify them in order to make them good enough for publication. Overall they took an average of 27 minutes to revise 8 TransLI texts (475 words), which corresponds to 1070 words/hour. That would amount to 8000 words per day compared to the average of about 6000 often used in the industry for revision (4 times the productivity of 1500 words translated per day per translator).

4. Evaluation in a pilot project

Although still not of publishable quality, the translations of the TransLI system that we developed in this project can be readily used for human revision, with promising productivity gains. Following those encouraging results on a small sample of a few sentences, we conducted a pilot study with the Federal Courts of Canada in which we translated a certain number of complete judgments from French to English and from English to French. We herein set out the more detailed evaluation of the revision process that we performed on a randomly selected set of 10 decisions (6 from French to English and 4 from English to French).

We also describe how we evaluate the quality of our current automatic judgment translations and the effort needed to revise them so that they can be published. As the summarization system of *NLP Technologies* already divides a judgment into four main thematic segments: **Introduction**, **Context**, **Reasoning** and **Conclusion**, we describe the evaluation using those divisions. In order to give an idea of the source text, of the raw SMT translation produced and of the revised output judged acceptable for publication, Table 2 shows a few sentences from each division.

[4] www.lexum.ca
[5] www.tsrali.com

The thematic segmentation is based on specific knowledge of the legal field. According to our analysis, legal texts have a thematic structure independent of the category of the judgment (Farzindar and Lapalme, 2004) Textual units dealing with the same subject form a thematic segment set. In this context, we distinguish four themes, which divide the legal decisions into thematic segments, based on the work of judge Mailhot (1998):

- **Introduction** describes the situation before the court and answers these questions: who did what to whom?
- **Context** explains the facts in chronological order: it describes the story including the facts and events related to the parties and it presents findings of credibility related to the disputed facts.
- **Reasoning** describes the comments of the judge and the finding of facts, and the application of the law to the found facts. This section of the judgment is the most important part for legal experts because it presents the solution to the problem between the parties and leads the judgment to a conclusion.
- **Conclusion** expresses the disposition, which is the final part of a decision containing the information about what is decided by the court.

In order to evaluate the results of the automatic translation, we computed two automatic measures over the space-separated tokens of a sentence. A token is thus a word plus any accompanying punctuation or symbols. A token can also be any sequence of contiguous non-space characters:

- **Edit distance:** the number of tokens that differ in the source and revised text as computed by the classical Levenshtein distance algorithm (Levenshtein, 1966).
- **Number of operations:** the number of consecutive insertion, deletion and replacement operations to transform the source into the revised text. For example, replacing 5 consecutive words would count as 5 in the edit distance but for only one operation. This measure approximates the number of cut and paste operations needed to revise an SMT translation.

	dist	ops	French original	SMT Fr->En output	Post edited version
Introduction	1	1	[1] Il s'agit d'une requête visant à obtenir un sursis d'exécution de l'ordonnance de déportation émise contre le demandeur prévue pour le 3 novembre 2008 à 18 h 30.	[1] This is a motion for a stay of execution of the deportation order issued against the applicant scheduled for November 3, 2008 at 6:**30.**	[1] This is a motion for a stay of execution of the deportation order issued against the applicant scheduled for November 3, 2008 at 6:**30p.m.**
Context	5	5	[8] Le 13 avril 2007, le demandeur s'est prévalu d'un Examen des risques avant renvoi (« ERAR ») et, le 16 mai 2007, il présentait une deuxième demande de résidence permanente pour raisons humanitaires. Ces deux dernières demandes furent entendues par le même agent i.e. Patricia Rousseau, laquelle, par décision du 31 juillet 2008, rejetait les deux demandes.	[8] On April 13, 2007, the Applicant availed • of a pre-removal risk assessment ("PRRA") and, on May 16, 2007, he submitted a second application for permanent residence on humanitarian and compassionate grounds. These last two applications were heard by the same • **officer** Patricia Rousseau, ~~i.e.~~ **that**, by decision dated July 31, 2008, dismissed both applications.	[8] On April 13, 2007, the Applicant availed *himself* of a pre-removal risk assessment ("PRRA") and, on May 16, 2007, he submitted a second application for permanent residence on humanitarian and compassionate grounds. These last two applications were heard by the same officer, **i.e.** Patricia Rousseau, **who,** by decision dated July 31, 2008, dismissed both applications.
Reasoning	5	3	[16] Quant au préjudice allégué, il s'agit de celui subi par tous ceux qui sont l'objet d'une ordonnance de renvoi.	[16] As to the harm • **alleged** is that • by all those who are subject to a removal order.	[16] As to the harm alleged, **it** is that which is experienced by all those who are subject to a removal order.
	4	3	[17] En conséquence, les conditions exigées par l'arrêt Toth n'étant pas respectées, la demande de sursis ne peut être accueillie.	[17] Accordingly, the conditions required by • Toth **is** not • met, the application for a stay cannot be allowed	[17] Accordingly, the conditions required by the Toth **case** not having been met, the application for a stay cannot be allowed.
Conclusion			[18] Pour ces motifs la Cour ordonne que la demande de sursis de la mesure de renvoi soit rejetée.	[18] For these reasons, the Court orders that the application for • stay of removal is dismissed.	[18] For these reasons, the Court orders that the application for a stay of removal is dismissed.

Table 2: Sentences from a decision (2008fc1224 from http://decisions.fct-cf.gc.ca/fr/2008/2008cf1224/2008cf1224.html). The first column indicates the theme in which the summarization system has classified the source sentence of the fourth column; the second column indicates the Levenshtein edit distance in terms of space separated tokens between the original SMT output (fifth column) and the revised output (sixth column). Replacement of tokens is shown in bold in the original and revised. Insertion in the revision is underlined and the insertion point is shown with a bullet in the original. Deletion of original is indicated by a strike-through the deleted text in the original. Because the sentences are tokenized at blank spaces, some indications may not reflect accurately the minimum distance or the sequence of editing operations performed by the revisor: for example, in paragraph [8] «officer» and «officer,» were considered as distinct tokens so the editing sequence is reported as an insertion of «officer,» and a replacement of «officer» by «i.e.».

	English to French (4 texts)				French to English (6 texts)			
Theme	**Nb ops**		**Nb tokens**		**Nb ops**		**Nb tokens**	
Introduction	31	5%	397	8%	13	2%	350	5%
Context	297	47%	2046	41%	154	18%	1246	19%
Reasoning	281	44%	2243	45%	646	76%	4642	69%
Conclusion	28	4%	298	6%	38	4%	457	7%
Total	637	100%	4984	100%	851	100%	6695	100%

Table 3: Number and percentage of editing operations and tokens in each division over ten judgments

Edit distance	**English to French**		**French to English**	
Introduction	51	13%	19	5%
Context	626	31%	263	21%
Reasoning	518	23%	1213	26%
Conclusion	43	14%	68	15%
Overall	1238	25%	1563	23%

Table 4: Edit distance in tokens for each division, the percentages are taken over the number of tokens given in the fourth and eighth column of Table 3

Table 2 shows examples of values of these measures on a few sentences. Even though, the exact values of the number of operations might differ from what a careful reviewer might do, we think this value is a good approximation of the work needed for revision.

Table 3 shows that for both translation directions, the number of editing operations is roughly equivalent to the number of tokens in each division. Table 4 shows that the global proportion of differences is similar for both directions of translation. The results are slightly better on the French to English direction, which is expected due to the complexity of the French language (with the accents and exceptions) bringing more complications to the machine translations. When we compare the different themes, we see that the Introduction and Conclusion themes require significantly less editing than the Context or the Reasoning themes. The type of text used in these themes in part explains these differences. In the legal field, the sentences used for the Introduction and Conclusion of the judgments often use the same expressions while the Context and Reasoning contain more sentences which are seldom seen in multiple judgment. Sentences from the Context that explain the litigation events are more variable.

5. Conclusion

The volume of legal content is growing rapidly. In Canada it is even more problematic because it is created in two languages and different formats. As a result, the amount of data that must be translated in short time has grown tremendously, making it difficult to translate and manage.

Legal organizations need solutions that enable them to handle quickly a high volume of translations. Our goal was to study the ability to train translation systems on a specific domain or subject area like the legal field so as to radically increase translation accuracy. This process recycles existing translated content to train the machine on the terminology and style of the requested domain.

To our knowledge this is one of the first times that an SMT engine has been developed specifically for judicial texts and evaluated in a pilot study. We managed to establish that an SMT engine trained on an appropriate corpus can produce a cost-effective revisable text.

6. Future Work

An interesting aspect of our findings is that review and post-editing of judicial translations are an important part of an SMT-integrated work flow. Reviewers with subject knowledge need to have direct access to the translation process in order to provide a feedback loop to the SMT training process.

We will therefore continue further investigation into an optimization of the post-editing and reviewing process, specifically with a focus on quantifying the distance, measured in number of operations and edits, to arrive at a fully acceptable translation.

As part of an ongoing collaboration with Palomino System Innovations Inc., a Canadian web content management system provider, we will evaluate integration of TransLI SMT into a translation work flow system – with a view to apply SMT to generic web content in the future.

Acknowledgments

We thank the Precarn Program for partially funding this work and the Federal Courts for their collaboration and feedback. We sincerely thank our lawyers for leading the human evaluation: Pia Zambelli and Diane Doray. The authors thank also Fabrizio Gotti and Jimmy Collin for technical support of experiments. We also thank Elliott Macklovitch, the Coordinator of the RALI and Markus Latzel, CEO of Palomino System Innovations Inc. for the support and many fruitful discussions.

References

E. Chieze, A. Farzindar, and G. Lapalme, 2008, “Automatic summarization and information extraction from Canadian immigration decisions”. In *Proceedings of the Semantic Processing of Legal Texts Workshop*, LREC 2008.

A. Farzindar and G. Lapalme. LetSum, an automatic Legal Text Summarizing system. In Thomas F. Gordon (editors), *Legal Knowledge and Information Systems, Jurix 2004: the Sevententh Annual Conference*, p. 11-18, IOS Press, Berlin, Dec 2004.

A. Farzindar, G. Lapalme and J.-P. Desclés, 2004, Résumé de textes juridiques par identification de leur structure thématique. In *Traitement automatique de la langue (TAL)*, vol. 45, number 1, p. 39-64.

F. Gotti, A. Farzindar, G. Lapalme and E. Macklovitch. Automatic Translation of Court Judgments. In *AMTA'2008 The Eighth Conference of the Association for Machine Translation in the Americas*, p. pp 1-10, Waikiki, Hawai'i, Oct 2008.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, 2007, Moses: Open Source Toolkit for Statistical Machine Translation In *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

P. Koehn and J. Shroeder, 2007, Experiments in Domain Adaptation for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on SMT*, Prague, Czech Republic.

P. Koehn, 2004, Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004*, p. 115–124.

V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10 (1966):707–710.

M. Olteanu, C. Davis, I. Volosen and D. Moldovan, 2006, Phramer, An Open Source Statistical Phrase-Based Translator. In *Proceedings of the Workshop on Statistical Machine Translation*, p. 146–149.

L. Mailhot. Decisions, Decisions: a handbook for judicial writing. Editions Yvon Blais, Québec, Canada, 1998.

F. J. Och, H. Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

A. Patry, F. Gotti and P. Langlais, 2006, Mood at work: Ramses versus Pharaoh. In Workshop *on Statistical Machine Translation*, HLT-NAACL, New-York, USA.

F. Sadat, Johnson, H., Agbago, A., Foster, G., Kuhn, R., Martin, J., Tikuisis, A., Portage: A Phrase-based Machine Translation System, *ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond.* Ann Arbor, Michigan, USA. June 29-30, 2005. pp. 133-136.