Automatic Translation of Court Judgments

Fabrizio Gotti
Guy Lapalme
Elliott Macklovitch
RALI-DIRO
Université de Montréal
Montréal, Québec, Canada, H3C 3J7
{gottif,lapalme,macklovi}@iro.umontreal.ca

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

Abstract

This document presents an experiment in the automatic translation of Canadian Court judgments from English to French and from French to English. We show that although the language used in this type of legal text is complex and specialized, an SMT system can produce intelligible and useful translations, provided that the system can be trained on a vast amount of legal text. We also describe the results of a human evaluation of the output of the system.

1 Context of the work

NLP Technologies[1] is an innovative enterprise devoted to the use of advanced information technologies in the judicial domain. Its main focus is the DecisionExpress™ automatic summarization technology of legal information (Farzindar *et al.,* 2004, Chieze *et al.* 2008). During the last year, a feasibility study was performed in collaboration with researchers from the RALI[2] at Université de Montréal to determine to what extent judgments from the Canadian Federal Courts could be automatically translated. As it happens, about 50 new judgments are produced weekly; 80% of which are originally written in English, and 20%, in French. By law, the Federal Courts have to provide a translation in the other official language of Canada. Currently, there is a delay of many months 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.

This paper describes the process we have followed in the development of this translation system, whose performance has been assessed with the usual automatic evaluation metrics. We also present the preliminary results of a manual evaluation of the translations. To our knowledge, this is one of the first attempts to build a large-scale translation system of complete judgments for eventual publication.

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

2 Overview of the system

We have built a *classical* phrased-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 legislations referenced, whether the decision was in favor of the plaintiff, etc. This step also attempts to identify the parts of a decision: introduction, reasoning and decision. 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++[3] for creating the translation models and SRILM[4] 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[5]. *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. Because appropriate bookkeeping 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.

[3] code.google.com/p/giza-pp/
[4] www.speech.sri.com/projects/srilm
[5] iit-iti.nrc-cnrc.gc.ca/projects-projets/portage_f.html

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.

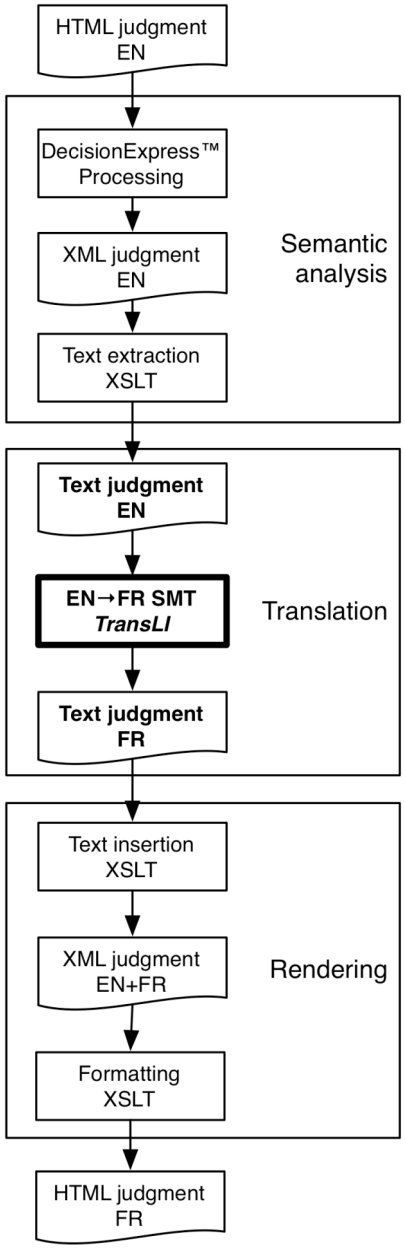

Figure 1: The translation pipeline translates an HTML court decision written in English into a French decision (also in HTML). A similar pipeline performs translations from French to English.

3 Building the corpora

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.

Corpus name	# sent pairs	# en words (K)	# fr words (K)
PRINCIPAL	245,000	6510	7510
TRAIN	244,000	6500	7500
TUNE-1	300	8	9
TEST	1300	28	33
TUNE-RECENT	400	8	10
TRAIN-LEXUM	1,000,000	22,340	25,720

Table 1: Corpora used for developing TransLI.

PRINCIPAL: we downloaded 14,400 decisions in HTML from the Federal Court of Canada web site[6] from which we extracted the text. Because many judgments did not have a translation or could not be parsed automatically with our tools, 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.

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[7] at the Université de Montréal in the context of the TransSearch[8] 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 Courts, the Tax Court and the Court of Appeal of Canada.

4 Experimentation with the system

During the development of the SMT engine, we used the classical Word Error Rate (WER) and Sentence Error Rate (SER) metrics. We also computed the BLEU score (Papineni *et al.*, 2001) that measures roughly the number of common subsequences between a reference translation and an SMT one, while penalizing important differences in sentence length. The goal is to obtain low WERs and SERs, but high BLEU scores.

All these scores are computed on the tokenized, lowercase version of the reference and SMT outputs. We have developed a series of scripts and language models for restoring proper upper and lower case and spacing between words so that the output can be easily read and evaluated by humans. This is the output that will be shown in this paper. Our experiments were limited to the *narrative* parts of the judgments. Decisions start and end with standard, stereotyped, administrative information such as name of lawyers, the name of the judge, docket numbers, the name of the parties, etc. We took for granted that this information would be appropriately translated via dictionary look-up, since the latter resource is readily available from NLP Technologies for its other product offerings.

Unless mentioned otherwise, TEST is used in all evaluations. When we translated from English to French, the English part of TEST is used as source and the French part as the reference to which the SMT output is compared. When translating in the other direction, the English part is used as source and the French part as reference. We did not take into account in our evaluations the fact that a text was an original or a translation.

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

[7] www.lexum.ca

[8] www.tsrali.com

Initial configuration

Our first tests were conducted using the default configuration of Moses, as suggested in the instructions for the WMT07 shared task (Schwenk, 2007).

The creation and tuning of the translation model takes about 24 hours on a 3 GHz computer with 8 Gb of memory. The size of a translation model file is around 3 Gb. Translating a judgment takes between 5 and 20 minutes depending on the number of sentences and their length. A typical judgment is around 60 sentences long; a sentence is 30 words long on average. Sentences in judgments are thus longer than the ones often used in other types of MT evaluations.

Table 2 shows the results we obtained for the default configuration (prefixed with an asterisk in our subsequent tables). Scores when translating into English are systematically a bit higher. This can be attributed in part to the fact that there is less morphologic variation in English than in French. We have seen a few cases in which gender agreement was incorrect in French whereas this is less problematic in English. As most of the decisions were originally written in English, we also think that judges often used similar formulations for similar cases, whereas French texts were translated by different translators who create slightly different formulations. These observations would have to be analyzed more systematically but they have been confirmed in informal discussions with administrators at the Federal Courts.

	WER	SER	BLEU
*English to French	41.5%	88.5%	43.1
*French to English	38.2%	83.2%	43.7

Table 2: Scores obtained with the initial default configuration of Moses. The * denotes this default configuration in subsequent tables.

Relatively low WERs and SERs and high BLEU scores show that we achieved quite respectable results. This only confirms the well-known observation that SMTs working on very limited domains outperform those trained on more generic, or out-of-domain texts. See for instance (Koehn *et al.*, 2007) for an illustration of this principle when translating journalistic corpora with in- and out-of-domain training corpora.

These excellent scores are no doubt due to the highly repetitive nature of both source and target texts. This correlation has already been observed in another study (Langlais *et al.,* 2005), where the authors obtain scores of about 85 when translating weather bulletins, which are much shorter, highly repetitive, texts exhibiting a simpler structure than judgments. Table 3 show some example translations produced by our system.

We now describe the series of tests we performed in order to improve on these results.

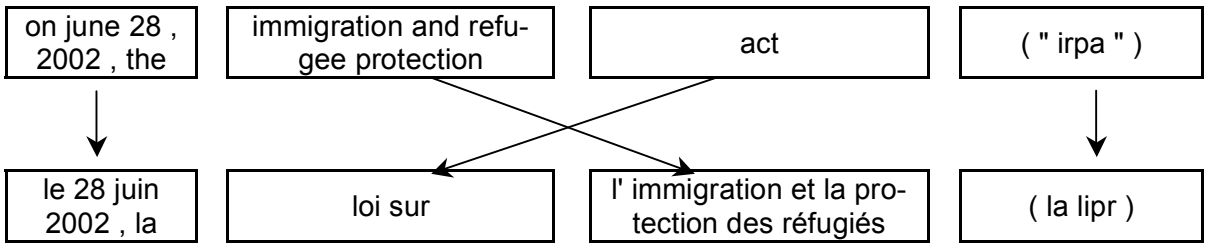

Figure 2: Reordering an English source sentence into French. The crossing arrows show the inversion (or swap) of two segments when translating. The full original sentence is "*On June 28, 2002, the Immigration and Refugee Protection Act ("IRPA") came into force and her PDRCC application was converted to an application for a pre-removal risk assessment ("PRAA").*" and the translation produced by TransLI is "*Le 28 juin 2002, la Loi sur l'immigration et la protection des réfugiés (la LIPR), est entrée en vigueur et sa demande de DNRSRC a été convertie en une demande d'examen des risques avant renvoi (ERAR).*"

Source	TransLI	Reference	BLEU
[14] In March 2000, the applicant applied for and received a passport.	[14] En mars 2000, le demandeur a demandé et obtenu un passeport.	[14] En mars 2000, le demandeur a demandé et obtenu un passeport.	100.0
Given that the test in Toth, supra, is conjunctive, I need not consider seriousness of the issue or balance of convenience.	Étant donné que le critère énoncé dans l'arrêt Toth, précité, est conjonctif, je n'ai pas tenu compte de la gravité de la question ou de la prépondérance des inconvénients	Puisque les éléments du critère de l'arrêt Toth, précité, sont cumulatifs, je n'ai pas à examiner le sérieux de la question de la prépondérance des inconvénients.	22.4
It is well known that the present procedures were put in place because a practice had grown up in which many many cases, totally devoid of merit, were initiated in the court, indeed were clogging the court, for the sole purpose of buying the appellants further time in Canada.	Il est bien connu que les présentes procédures étaient mises en place parce qu'une pratique avait grandi dans laquelle un grand nombre de nombreux cas, totalement d'honneur, a été introduite en cour, de fait n'étaient clogging la cour, dans le seul but d'acheter les appelants plus de temps au Canada.	Chacun sait que la procédure actuelle a été mise en place parce qu'une pratique s'était développée par laquelle de très nombreuses demandes, tout a fait dénuées de fondement, étaient introduites devant la cour et encombraient les rôles, uniquement pour permettre aux appelants de demeurer plus longtemps au Canada.	10.7

Table 3: The first column is the source English text and the second is the translation produced by TransLI with its BLEU score in column 4 when compared against the reference in the third column. The first sentence was translated exactly as in the reference, while the second used a different but quite acceptable formulation. An unknown word *clogging* at the end of the third example produces an unacceptable translation, which is barely understandable.

Strategy	Sophistication	Summary description	Additional models?
monotone	simple	No reordering of target segments	no
distance	medium	Counts the number of skipped segments when producing the target segments	no
toggle	complex	A reordering model allowing the translation of source segments into target ones in either monotonic or non-monotonic segment positions	yes
* msd	complex	Like "toggle", but more refined in the ways it can position target segments compared to their source counterparts' positions	yes

Table 4: Different types of reordering strategies we tested with the Moses SMT engine. These strategies are presented in increasing order of complexity.

Strategy	Translation model size	Training time	Translation time	wer (%)	ser (%)	BLEU
monotone	261 Mo	10 h	0.8 s/sent	41.7	88.4	42.6
distance	261 Mo	10 h	2.7 s/sent	41.8	88.1	42.7
toggle	no data (Moses crashes during training)					
* msd	444 Mo	10.5 h	4.1 s/sent	41.5	88.5	43.1

Table 5: Model sizes and times for the training and translation steps for the models of Table 4, along with the scores obtained during testing.

	Tuning corpus	wer (%)	ser (%)	BLEU
English → French	* TUNE-1	41.5	88.5	43.1
	TUNE-RECENT	41.5	87.5	42.9
French → English	* TUNE-1	38.2	83.2	43.7
	TUNE-RECENT	37.5	82.9	44.1

Table 6: Performance comparison of TransLI, in the two translation directions with tuning corpora TUNE-1 and TUNE-RECENT. The asterisk (*) is the default configuration. The gray row indicates the better performance.

	Training corpus	wer (%)	ser (%)	BLEU
English → French	TRAIN (244,000 sent)	41.5	87.5	42.9
	TRAIN-LEXUM (1,000,000 sent)	37.2	80.1	43.9
French → English	TRAIN (244,000 sent)	37.5	82.9	44.1
	TRAIN-LEXUM (1,000,000 sent)	34.9	79.1	45.7

Table 7: Performance comparison of TransLI, in the two translation directions with training corpora TRAIN and TRAIN-LEXUM. The gray row indicates the better performance.

	Lexicon use	wer (%)	ser (%)	BLEU
English → French	without lexicon	37.2	80.1	43.9
	with lexicon	37.3	80.3	43.8
French → English	without lexicon	34.9	79.1	45.7
	with lexicon	35.0	79.5	46.2

Table 8: Performance comparison of TransLI, in the two translation directions with and without lexicon integration. The gray row indicates the better performance.

	Translation engine	wer (%)	ser (%)	BLEU
English → French	TransLI	37.3	80.3	43.8
	Google	48.4	88.8	30.0
French → English	TransLI	35.0	79.5	46.2
	Google	45.9	88.5	31.2

Table 9: Performance comparison of TransLI and Google, in the two translation directions with and without lexicon integration. The gray row indicates the better performance.

Reordering strategies

Given the fact that the word order in French and English can differ in many cases (see Figure 2), we conducted some experiments with the reordering techniques provided by Moses, which are briefly described in Table 4. Table 5 shows the results that we obtained.

The default MSD reordering strategy produced slightly better performance. However, given the cost of an additional model it incurs, we did not consider this small improvement sufficient to warrant its use in our production context. Rather, we decided to opt for a distance-based reordering strategy, which seems a good compromise between, on the one hand, the translation quality of our pipeline and, on the other hand, the additional computing time needed by a more involved reordering model as well as the maintenance effort it would require.

Tuning corpus

After training our SMT system with TRAIN, we decided to change the tuning corpus from TUNE-1 to TUNE-RECENT in order to better reflect the context of use. Table 6 shows that results slightly improved when translating to English with a more recent tuning corpus, but

they were slightly worse when translating into French. For methodological reasons, we prefer to use a more recent tuning corpus nonetheless.

Size of training corpus

We almost quadrupled the training corpus by adding sentences taken from judgments from other courts (corpus TRAIN-LEXUM). Table 7 shows that this considerably improved the performance of the system, as expected.

Adding specialized lexicons

The judicial domain is an area in which a relatively well-defined terminology has been adopted and in which there are fixed expressions for many concepts. We wanted to see if terminological lexicons would improve TransLI's performance on these types of expressions. We compiled a list of terms taken from two sources: a series of French-English equivalents already compiled by NLP Technologies for its other activities and a list of bilingual terms in many areas of law (Canadian Passport, common law, immigration, Parliament, etc.) found on the web site of the Translation Bureau of Canada[9]. These two sources of information provided more than 33,000 bilingual entries with which we augmented our translation models.

Integrating this type of resource into a statistical translation pipeline is a delicate matter, since the associations it provides do not come with a statistical weight. Many strategies exist to overcome this. See for instance (Sadat et al., 2006) and (Och et al., 2003).

We chose to integrate these lexical entries by considering each of them as a sentence pair, with which we augmented our current training corpus, TRAIN-LEXUM. We artificially boosted the presence of these entries in the newly created corpus by repeating them 5 times each. This produced a new training corpus of 1,1670,000 sentences. Table 8 shows that this addition yielded some improvement in the French to English direction, but not in the English to French direction, probably because we did not take French morphology into account when we added the pairs to the translation model.

One might think that most of the terms found in these lexicons were already in the original translation tables built from TRAIN-LEXUM. But when we checked, we only found 10% of these entries in the translation model built from TRAIN-LEXUM. This proportion increases to 90 % when TRAIN-LEXUM is artificially augmented with the lexical entries. Given the importance of proper terminology in judicial texts, we decided to adopt the translation model built with the addition of this lexicon.

Comparison with Google translate

Google provides a *free* statistical translation service on the web[10]. We evaluated Google's engine by manually copying and pasting the text of TEST into the dialog box on the Google site. The translations were done on April 23rd 2008. Table 9 clearly shows that our system obtained much better results than that public version of Google Translate. Of course, this small scale experiment is subject to caution; it was done just to show to a potential client that the *free solution* can be improved upon by proper customization. Although quite readable, one of the main reasons for the relatively low scores of Google is the fact that the terminology used in the translation is not tailored to Canadian Court judgments.

Final configuration chosen

After all these experiments, we opted for the following configuration for TransLI as a compromise between quality, ease of deployment and maintenance and speed of translation:

- A distance based reordering strategy;
- A recent tuning corpus;
- A training corpus as large as possible;
- Integration of specialized lexicons.

5 Human evaluation

Given the fact that the output of our SMT system is intended to be provided to revisers to correct before publication, we did not want to rely solely on automatic evaluation measures.

[9] www.bureaudelatraduction.gc.ca

[10] www.google.com/translate_t

We wanted to evaluate both the quality of the language produced by TransLI independently of the source text and also the quality of the translation, taking into account the original text. These two evaluations were performed for both French to English translations and English to French, giving rise to 4 evaluation configurations.

Our TEST corpus contained thirteen texts but one was discarded in this evaluation because it was too short (150 words). The average length of the English texts was 2010 words and 2254 for the French texts.

For each configuration, we randomly selected two or three contiguous sentences (between 50 and 75 words) in each of the twelve texts. This was done in order to have a sample of each text while keeping the manual evaluation time relatively short. Of course, this had the drawback that the evaluators did not see the full context of the translations they were asked to evaluate.

Evaluation team and set-up

We organized an evaluation session with three members of the NLP Technologies Editorial Board including 2 lawyers and one certified legal translator. All three are bilingual persons (French and English) but as they are native English speakers, we concentrated our evaluation on the French to English translation direction.

In order to better calibrate the evaluations, 4 sentence groups taken from reference translations were included with 8 produced by TransLI. The selection between reference and SMT translations was random. The evaluators did not know which ones were the reference translations and which ones TransLI had produced. All evaluators were asked to evaluate the same set of sentences and translations during a single evaluation session.

Quality of the language of the translation

For the quality of language, we asked the evaluators to assign each passage 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.

Figures 3 and 4 show the mean score given by the evaluators for each text. The 8 histogram bars on the left are texts from TransLI and the 4 on the right represent reference texts. In order to facilitate the comparisons, we have sorted the scores from left to right for both TransLI and reference. Of course, this order does not reflect the order in which the evaluators saw these texts.

For the English version (Figure 3), the average scores are 2.54 for TransLI and 3.58 for the reference. From TransLI, 6 out of 8 obtained scores higher than 2. Only one reference and one TransLI output were judged perfect by all evaluators. Some reference translations were deemed imperfect by our evaluators, because their personal preferences regarding certain phrases or wording did not match those of the reference translator, something to be expected.

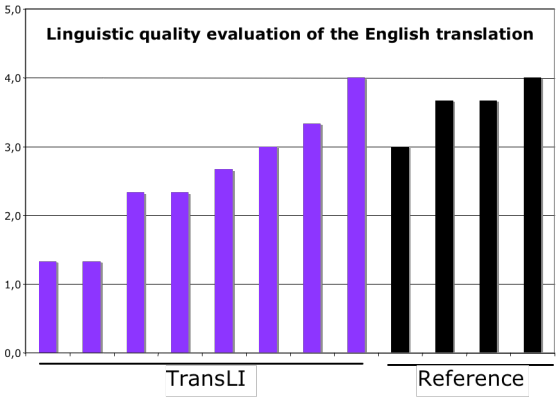

Figure 3: Mean scores from 1 (unacceptable) to 4 (perfect) for the English translations. The scores are sorted by value for both TransLI and reference translations.

For the French version (Figure 4), the average scores are 3.0 for TransLI and 3.75 for the reference. These much higher scores for French were a bit surprising to us since we were under the general impression that the English texts were of better quality. This impression had been confirmed by the slightly higher BLEU scores. The high French scores may be attributable to the fact that our evaluators, being native English speakers, were less inclined to score the French version harshly.

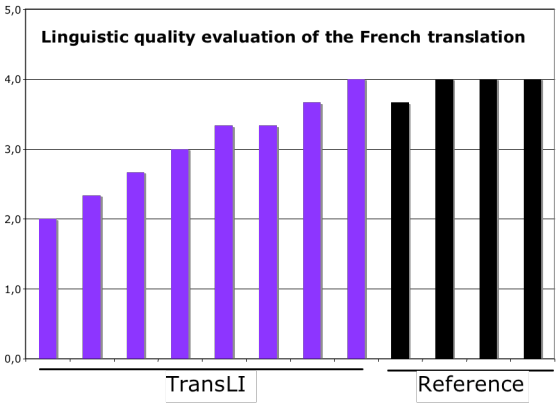

Figure 4: Mean scores from 1 (unacceptable) to 4 (perfect) for the French translations. The scores are sorted in increasing order of value for both TransLI and reference translations.

Fidelity of the translation

We also wanted to evaluate to what extent the SMT conveyed all the semantic content of the original, which we call fidelity henceforth. The same three evaluators were given couples 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 not asked to grade the translations as in the previous section, but to modify them in order to make them good enough for publication. The evaluators made their modifications by editing the text in Microsoft Word.

By comparison with the original translation, we counted the number of words modified by the evaluators. We hypothesized that the fidelity score of a translation would be inversely correlated with the number of corrections made by the evaluators. We manually separated all modified words into two categories: those that we judged as *linguistic* and *stylistic* modifications and those that we considered *semantic* modifications. A stylistic modification could be the substitution of "Finally" with "In the end", while a semantic modification generally attempts to correct an omission or error made by the SMT.

In Figure 5, we give only one measure: the number of modified words for semantic modification between the original and the final version. Overall, the evaluators modified 8.6% of the words produced by TransLI but also 3.0% of the words in the reference. We consider this an encouraging result, given the fact that these types of texts are very specialized.

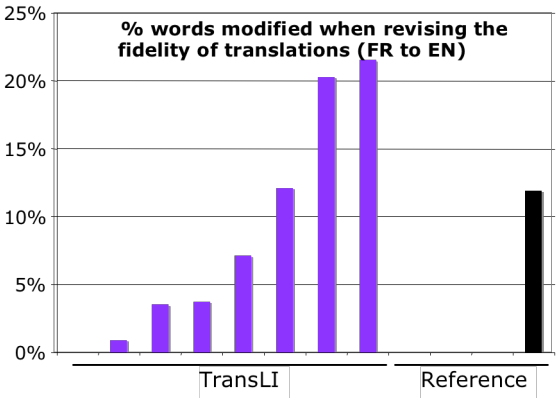

Figure 5: Percentage of modified words. The scores are sorted by value for both TransLI and reference translations.

A more detailed analysis is underway to better understand the types of modifications that were made but this would require more translation samples. This should help us identify new directions for improving the translation engine.

Time needed for revision

Although it is a bit risky to extrapolate with such a small sample, we thought it would be interesting to know if it can be faster to revise TransLI output than to revise human translation. Here, the reference translations are the human translation. Our evaluators did not know which translations had been produced by a human or which were produced by a machine.

We also kept track of the time spent by the evaluators on each text. 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 mean of about 6000 often used in the industry for revision (4 times the productivity of 1500 words translated per day per translator).

These numbers should also be compared with the time spent on reference texts. Our evaluators took an average of 6.8 minutes to revise 4 sentences (196 words), which amounts to 1717 words/hour, a 71% difference compared with the revision time for SMT output.

6 Future work

We plan to further the research presented. We will first evaluate the French output and perform 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.

It would also be interesting to perform a task-oriented evaluation to measure to what extent the SMT output can be used in a production environment without revision. We could also increase the scale of the experiment (additional evaluators and evaluation material) to obtain more statistically significant results. We would also like to know to what extent other configurations of Moses, e.g. factored translation models or training at the lemma level, could improve the translations.

7 Conclusion

To our knowledge this is one of the first times that an SMT engine has been developed specifically for judicial texts. 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. 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. A more detailed analysis is in progress to evaluate the cost-effectiveness of this approach in a production setting.

Acknowledgments

We thank the Precarn-CRIM Alliance Program for partially funding this work and the Federal Courts for their collaboration and feedback. We sincerely thank our *human* evaluators: Pia, Nancy and Mark.

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, 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. Och and H. Ney, 2003, A systematic comparison of various statistical alignment models, *Comput. Linguist.*, vol. 29, number 1, p .19–51.

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.

P. Langlais, S. Gandrabur, T. Leplus and G. Lapalme, 2005, The Long-Term Forecast for Weather Bulletin Translation. In *Journal of Machine Translation*, vol. 19, Springer Netherlands, p. 83–112.

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.

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.

K. Papineni, S. Roukos, T. Ward and W.J. Zhu, 2001, Bleu: a Method for Automatic Evaluation of Machine, Translation, IBM Research Report rc22176 (w0109022), IBM Research Division, Thomas J. Watson Research Center.

F. Sadat, G. Foster and R. Kuhn, 2006, Système de traduction automatique statistique combinant différentes ressources, *TALN 2006*, Belgique.

H. Schwenk, 2007, Building a Statistical Machine Translation System for French Using the Europarl Corpus. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.