NRC Publications Archive
Archives des publications du CNRC

Filtering and routing multilingual documents for translation

Carpuat, Marine; Goutte, Cyril; Isabelle, Pierre

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.
For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:
https://doi.org/10.1109/CISDA.2012.6291536
2012 IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA 2012), pp. 1-7, 2012-07-13

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=12610fd4-0110-475e-bc5c-0dfc0481f7de
https://publications-cnrc.canada.ca/fra/voir/objet/?id=12610fd4-0110-475e-bc5c-0dfc0481f7de

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at
PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada
Conseil national de recherches Canada

Canada

Filtering and Routing Multilingual Documents for Translation

Marine Carpuat, Cyril Goutte and Pierre Isabelle
Interactive Language Technology
National Research Council Canada
Gatineau, QC J8X3X7 Canada
Email: Firstname.Lastname@nrc.ca

***Abstract*—Translation is a key capability to access relevant information expressed in various languages on social media. Unfortunately, systematically translating all content far exceeds the capacity of most organizations. Computer-aided translation (CAT) tools can significantly increase the productivity of translators, but can not ultimately cope with the overwhelming amount of content to translate. In this contribution, we describe and experiment with an approach where we use the structure in a corpus to adequately route the content to the proper workflow, including translators, CAT tools or purely automatic approaches. We show that linguistically motivated structure such as document genre can help decide on the proper translation workflow. However, automatically discovered structure has an effect that is at least as important and allows us to define groups of documents that may be translated automatically with reasonable output quality. This suggests that computational intelligence models that can efficiently organize document collection will provide increased capability to access textual content from various target languages.**

I. Introduction

Organizations interested in monitoring social media for commercial or security purposes must handle growing amounts of content, but also an increasing number of languages. Accessing multilingual information is a key capability for analyzing this data efficiently. However, traditional translation performed by professional human translators suffers from two main drawbacks in the context of social media data analysis. First it is a slow process that would bottleneck automatic analysis in the large text collections generated by social media users. Second, and more importantly, systematic translation of multilingual content far exceeds the capacity of most organizations. Although many tools are available to ease the translators job, such as translation memories, terminology banks or bilingual concordancers, the gains in productivity cannot match the growth in content that needs to be translated.

Machine translation promises to bring productivity gains, but suffers from a number of problems. Although its output is often good enough for gisting, it is still far from reaching the same quality as human translation. More problematic is the fact that state-of-the-art, statistical machine translation (SMT) is highly sensitive to mismatches between the text to translate and the training material used to estimate the translation model. As a consequence, text that is unlike the training data may produce poor translations. Typical experimental settings considered in SMT research (e.g. [1], [2]) use well defined benchmark data in a well controlled setting. In contrast, social media data is highly dynamic in nature, with many different sources spanning a range of topics, genres and quality, such that appropriate representative training data is hard to find.

We address the following questions. First, how does translation quality of computer-aided translation (CAT) tools vary across text genres? Heterogeneous text collections can be organized along many dimensions. We propose to start with genre, since the genre of a document attempts to represent its medium and communicative goal. Genre is therefore particularly relevant in the context of social media, where documents vary widely depending on their source and purpose, and are quite different from the texts that have been historically studied in machine translation (e.g., Canadian or European parliament proceedings [3]). We will show that there are systematic differences in CAT performance across genres. This suggests that genre information (either given or automatically attributed [4]) is useful to route documents to be translated to the appropriate CAT tools.

The second question we address is whether we can automatically discover relevant structure from heterogeneous data in order to translate a wide variety of documents more efficiently and accurately. We show that indeed, we discover clusters of textual documents on which CAT tools have very different and consistent behaviour. This suggests a natural way to route documents to the appropriate combination of fully automatic translation, semi-automatic translation involving, e.g. post-edition, or fully human translation when CAT tools are ineffective.

As there is currently no publicly available large corpus of translated social media data, we illustrate our approach and results on a multilingual document collection from the United Nations. This data set offers several realistic and desirable features: it is large, it covers several languages, and it contains text of very different genres. Contrary to most data used in security and defence applications, it is publicly available and can be used to demonstrate the approach.

In the following two sections, we describe the methods used in our approach. Section IV describes the data that we use in our experiments, and sections VI and VII shows how corpus structure may be leveraged to improve the translation workflow, using both existing manual distinctions, as well as automatically discovered corpus structure.

II. Methods: Computer-Aided Translation

This section gives a brief overview of the two main computer-aided translation methods that are widely used today: translation memories and statistical machine translation. Since our contribution is focused on evaluation of translation quality and its impact on routing documents for translation, we are not interesed in proposing new methods here. On the contrary, we aim to use well-established tools in order to build a pipeline that is representative of current practice.

A. Translation Memory

The first type of CAT tool we consider in this study are Translation Memories (TM). While they have not been studied much in the field of machine translation, TM are widely used by professional translators. TM improve translators' productivity and consistency by providing examples of similar sentences that have already been translated.

We use an NRC-internal implementation [5] that simulates commercial systems while giving us more control over parameters. Given a corpus of existing translated sentences, and a new source sentence that we call the *query*, we compute the *distance* between the query and every source sentence in the corpus. The translation memory outputs the target side of the sentence pair with the minimum distance between its source side and the query. We report results using $1 - smoothedBLEU$ as the distance metric, where $smoothedBLEU$ is a smoothed version of BLEU score [6], the standard metric of translation quality based on n-gram precision between system output and reference translations [7]. We experimented with other distance metrics and found that $smoothedBLEU$ worked best when using BLEU as the evaluation metric.

B. Statistical Machine Translation

Our second CAT tool is Statistical Machine Translation (SMT). In contrast with TM, SMT is a very active topic of research and a wide variety of models and systems exist. We use the NRC's PORTAGE system [8], a state-of-the-art phrase-based SMT system which placed 3rd out of 19 participants in the constrained Chinese-to-English task at the 2009 NIST OpenMT evaluation, and performed very well in the latest 2012 evaluation for both Chinese-English and Arabic-Englich. In this paper, the SMT decoder scores translation candidates for each sentence using a loglinear mixture of the standard set of features (conditional translation probabilities and lexical weights for the translation model, lexicalized distortion and target language model). Features and their weights are estimated from bilingual parallel text. We refer the reader to [9] for a good overview of phrase-based SMT modeling.

III. Methods: Corpus Organization

We need tools to automatically learn how to organize our large heterogeneous collection of documents. We use a two step clustering approach for this purpose: (1) a Latent Dirichlet Allocation (LDA) model is used to discover latent topics in the document collection, and (2) documents are then clustered using their topic distribution as a low-dimensional representation.

We use the MALLET[1] implementation of LDA [10] to learn topics in a document collection. LDA is a generative model of documents, where a topic t is a distribution over words ϕ_t, and each document d is a mixture of T latent topics, given by probabilities θ_d. Dirichlet priors are placed over both ϕ and θ. In a document d, the word tokens $w^d = \{w_n^d\}_{n=1}^{N_d}$ are associated with topic assigments $z^d = \{z_n^d\}_{n=1}^{N_d}$, drawn from the document-specific topic distribution. Given a corpus of observed words w, the posterior distribution over z can be estimated using Gibbs sampling, and used to compute the word-topic distributions and the topic-document distribution. In practice, we run the sampler for 1000 iterations. The concentration parameter β for the topic-word concentration is initialized to 0.01, while the document-specific topic distributions are initialized with a concentration parameter $\alpha = 0.01/T$ and a uniform base measure [11]. Hyperparameters are reestimated every 10 sampling iterations.

Each document in the corpus can be represented as a vector $d \in [0,1]^T$, where T is the number of topics and d is the document-specific topic distribution from the LDA model. The resulting topic vectors are clustered using the k-means algorithm, as implemented in the WEKA toolkit[12]. The main parameters are therefore the number of topics T and the number of clusters k. In this paper, we use $k = 8$, so that we have the same number of genres and clusters, and we experimented with a range of topic numbers. We will report results with $T = 80$, and will give an overview of topics discovered in our data in Section VII.

The resulting two-step approach of LDA followed by k-means is representative of current approaches to document categorization [13, for a good overview]. It remains to be seen whether purely Computational Intelligence (CI) techniques such as fuzzy or neural systems can improve models for this task. In this contribution, we focus on evaluating the impact of standard approaches, and leave the investigation of more advanced CI models to future work.

IV. Data: a heterogeneous multilingual corpus

We work with a bilingual corpus of Chinese-English documents extracted from United Nations (UN) data and made available by the Linguistic Data Consortium.[2] The corpus comprises two parts:

1) A genre annotated subset of 972 Chinese-English documents, all published in 2000, with a genre label for each document. We refer to this subset as UN-2000.
2) An unannotated, parallel, sentence-aligned corpus of 33,174 Chinese-English documents published between 1993 and 1999. We refer to this subset as UN-9399.

Note that all documents come from the same general source (the United Nations), but the dates of the two subsets do

[1] http://mallet.cs.umass.edu/

[2] http://www.ldc.upenn.edu/

TABLE I
GENRE CATEGORIES IN THE DATASET

#	Title	Description
A	Biography/Résumé	Account of persons life; résumé/CV
B	Lecture/Briefing	Formal statement presented to committee; one side of debate, speech
C	Letter	Communication introduced by word "Letter" with addressee; line 1 stating "I have the honor to inform you that"; ending with signature
D	Meeting Documentation	Agenda, addendum/addenda, Minutes, Summary of Meeting; references to Draft Resolution(s), statements, reports, and comments
E	Draft Resolution	Detailed quotation of resolution or draft; series of paragraphs beginning with gerunds, then requests/decides (3rd s./present verbs:); allusion to session agenda
F	Message	Short or long communication, including the note verbale, written note by official (instruction or brief modification to agenda item/addendum)
G	Organizational Charter	Articles of organization, org. chart/lists, provisions for procedures and responsibilities; guidance and rules for group/members; job description(s)/vacancy announcement(s)
H	Study/Report	Detailed official statements, conclusions and recommendations; progress report and summary of issues on given topic

not overlap. This ensures that no document from the genre-annotated corpus is present in the large sentence-aligned corpus. This also ensures that the genre annotated documents come from the same source, but may cover new topics that appeared in 2000 but were not necessarily represented in previous years, a situation that is common in typical applications.

The genre annotation was originally collected by collaborators for an independent sentence alignment project. The list of genres is given in table I with a brief description of each. These categories are quite different from genres used in the traditional genre categorization literature (see e.g. [4] and references therein) and most are specific to the UN data. Some are defined by the communication medium, or by rules based on text patterns (e.g., Genre C, Letter). While some genres cover a wide variety of documents (G, org. charter), others make very fine grained distinctions: for instance, agenda and addenda should be in "Meeting Documentation" (D), while "Draft Resolution" (E) contains allusions to session agenda, and "Message" (F) contains written notes by officials which can be brief modifications to agenda items or addendum.

Each document was annotated independently by two human judges, and a third judge was asked to adjudicate documents with a final reference annotation. It is important to note that the disagreement between judges was fairly high, with an agreement ratio between 0.44 and 0.76 depending on the pair of judges, and a *kappa* between 0.31 and 0.70. This suggests that the genre annotation task is non trivial, and proved difficult for human judges, even with precise guidelines and definitions (cf. Table I). We therefore expect that genre cues are not simple to identify in text.

This annotation process yielded a total of 972 documents, each annotated with one of eight genres, which we refer to as UN-2000. Note that the distribution of documents per genre is

TABLE II
TEXT STATISTICS PER GENRE FOR UN-2000 CORPUS: NUMBER OF DOCUMENTS, NUMBER OF SENTENCES PER DOCUMENT; FOR EACH LANGUAGE: AVERAGE SENTENCE LENGTH ($\bar{\ell}$), NUMBER OF SINGLETONS PER DOCUMENT (SNGL), AND TOKEN TO TYPE RATIO (TTR).

	nb.	sent	Chinese			English		
#	docs	/doc	$\bar{\ell}$	sngl	ttr	$\bar{\ell}$	sngl	ttr
A	5	11.2	13.5	46.2	2.05	15.0	47.2	2.41
B	7	95.7	24.1	143.0	6.94	26.9	199.6	6.41
C	190	10.1	22.7	10.7	9.68	25.6	14.1	9.09
D	264	67.8	23.4	16.5	37.66	26.0	22.7	31.50
E	165	29.3	30.3	12.9	25.14	33.3	18.0	21.96
F	95	13.5	19.4	16.2	7.39	21.4	19.6	7.39
G	26	111.7	19.7	54.2	15.18	21.3	75.9	13.24
H	215	111.1	25.3	28.6	40.85	27.3	44.6	30.82

very unbalanced (Table II). The biography (A) and lecture (B) genres have fewer than 10 instances each, while the meeting documentation (D) genre alone represents more than 25% of all annotated documents. The length of documents vary with genre, but the number of sentences per document reported in Table II do not reflect the original document length, since sentences were discarded for the sentence alignment project.

Table II gives basic statistics per genre in both languages: average sentence length, number of singletons (words occuring only once in the corpus) per document, and token-to-type ratio (or average word frequency). The biography/résumé genre (A) for instance has shorter sentences than average, while meeting documentation (D) has much longer sentences. Genres based on more formal or official documents tend to reuse more vocabulary with high token to type ratio (D, E, H), while informal genres exhibit more variation in vocabulary (B, C, F). The biography/résumé genre also has a very low token to type ratio, which can be explained by its small size and by the frequent occurrence of named-entities. As can be expected, the same trends are observed in Chinese and English.

It is hard to predict a priori how computer-aided translation tools are affected by these properties: it depends on the translation technology used, as well as on the properties of the existing translations used as a translation memory or training corpus. A statistical machine translation system might make fewer reordering errors in shorter sentences. However, the short sentences of genre A might still be hard to translate correctly if they contain many rare words that have not been seen in the training data. In a translation memory, short sentences might yield incorrect translations due to lack of context, while long sentences with many repetitions (such as in genre E) might be easier to translate if they are well covered by the examples in the memory.

In the following experiments, the large, non-annotated corpus is used to feed the translation memory and to train the statistical machine translation engine. The smaller, genre-annotated UN2000 corpus is used to test the impact of the corpus structure on the document translation quality. In particular, we check the impact of the genre on the translation quality.

V. Experiment set-up

We consider the two types of CAT tools described in Section II: translation memories (TM) and statistical machine translation (SMT). The TM and SMT systems are built using the exact same training data. We use the 33k documents from the UN-9399 corpus. This results in a training set of more than 3 Million sentence pairs, which is on par with the large-scale conditions considered in machine translation benchmarks.

The UN-2000 section of the corpus is used to evaluate the performance of the CAT tools. In order to speed up experimentation, we do not use the entire genre-annotated corpus as a test set, but only a subset of 51 randomly selected documents, or a total of 2942 sentences which is roughly the standard size for SMT test sets.

We will evaluate translation quality on different partitions of the corpus: (1) genres, according the manual annotation described in Section IV, and (2) document clusters learned automatically as described in Section III. Note that translations are always produced using a single CAT tool, i.e., a single translation memory or a single machine translation system trained on the entire training set. The resulting CAT tool is then applied to all test documents. Genre and cluster information does not inform the translation process, and is only used at evaluation time.

We use BLEU score [7] to automatically compare TM and SMT output with reference translations. Although some recent metrics have been shown to correlate slightly better with human judgments of translation quality, BLEU remains the de-facto standard metric in statistical MT, both for system optimization and for evaluation [14], [1]. BLEU is defined as the geometric mean of n-gram precision scores for $n = 1..4$, augmented with a length penalty, so that higher BLEU scores represent better translation quality. We did experiment with other metrics such as Word Error Rate [15], but do not include them because they yield the same trends as BLEU.

In addition to the overall BLEU score obtained for each partition of the UN-2000 corpus, we will compute BLEU scores for each document in the corpus. When comparing the translation quality of different systems on the same test set, corpus-level BLEU score is known to correlate better with human judgments of translation quality than document-level scores. However, document-level BLEU scores are still interesting in our setting since we are not interested in comparing different SMT systems on a fixed data set, but in comparing translation quality of a fixed SMT system across different data sets. Following recent work on translation quality estimation [16], we study the distribution of documents over BLEU quartiles rather than on the absolute BLEU values alone. We rank all documents in the UN-2000 test set according to their BLEU score, and report for each genre the percentage of documents that are ranked in each quartile.

VI. Impact of Genre Structure on CAT

In this section, we evaluate standard CAT tools based on genre distinctions.

TABLE III
BLEU scores for computer-aided translation, by genre.

Genre	# sent	TM BLEU	SMT BLEU
A	3	54.25	66.99
B	17	19.92	42.44
C	50	13.75	24.56
D	292	19.39	35.94
E	892	50.66	52.30
F	21	58.72	58.37
G	52	39.91	53.60
H	1635	23.94	32.28
All:	2942	27.29	36.97

TABLE IV
Document-level BLEU scores by genre. BLEU quartiles are defined by ranking documents across all genres, and are ordered from best (1st) to worst (4th)

	BLEU by doc			docs in each BLEU quartile (%)			
Genre	Mean	Min	Max	1st	2nd	3rd	4th
A	66.99	66.99	66.99	1	0	0	0
B	42.44	42.44	42.44	0	1	0	0
C	32.32	17.06	51.30	0	0.25	0.25	0.5
D	44.11	22.68	65.71	0.307	0.153	0.384	0.153
E	44.53	26.12	57.32	0.285	0.285	0.428	0
F	54.62	42.44	61.80	0.666	0.333	0	0
G	55.24	47.41	63.08	0.5	0.5	0	0
H	44.37	22.76	70.45	0.230	0.307	0.307	0.153

A. Genres capture some useful distinctions for CAT

Table III summarizes the BLEU scores obtained on each genre subset of the UN-2000 test set. It shows that CAT tools do very well on some genres such as E and F. This was perhaps expected for genre E, as we have seen in Section IV that it is one of the most formal and repetitive genres. Interestingly, genre F which is more informal and less repetitive aso gets a high BLEU score even though it might be considered hard to translate a priori. Genre A also shows a high BLEU score, but as this is the smallest genre, these scores are probably not reliable. The worst BLEU scores with both CAT tools are obtained for genre C.

Statistics for BLEU scores per documents are reported in Table IV. While the mean document-BLEU do not yield the exact same ranking as the genre-based BLEU, the top and worst genres remain the same. The distribution of documents in each BLEU quartile suggests that translations of documents in genres F and G are more reliable than those in genres E or C, since all documents in genres F and G are ranked within the best two BLEU quartiles. In contrast, half of the documents in genre C are in the worst BLEU quartile, confirming that its translation quality is not as good as those of the other genres.

B. Genres capture non-trivial distinctions

Could we have guessed which genres are easier to translate simply by comparing test documents with the training data? Unfortunately genre annotation is not available for the training corpus, so we cannot directly compare the genre distributions in the training and test data. However, we can compare the Chinese documents in each genre with the entire training corpus using Out-Of-Vocabulary (OOV) rate and source language model perplexity per word (PPL). Given a sequence of

TABLE V
LANGUAGE MODEL EVALUATION PER GENRE: OUT OF VOCABULARY WORD RATE (OOV) AND PERPLEXITY (PPL) FOR CHINESE AND ENGLISH 4-GRAM LANGUAGE MODELS TRAINED ON THE UN93-99 CORPUS

Genre	Chinese		English	
	OOV	PPL	OOV	PPL
A	10.61	336.76	4.28	191.51
B	0.17	107.35	0.21	80.64
C	1.57	126.58	1.27	93.81
D	0.79	106.72	0.54	74.63
E	0.87	77.28	0.50	57.95
F	2.04	163.70	1.25	110.61
G	0.72	186.93	1.07	124.68
H	0.71	156.16	0.74	109.65

words $S = (w_1, .., w_s)$ (e.g, a sentence, document, or set of documents), they are defined as follows:

- the out-of-vocabulary rate, OOV, represents the ratio of test tokens that are unseen in the UN93-99 corpus:

$$OOV = \frac{\sum_{i=1}^{s}(1 - \mathbb{1}_{UN93-99}(w_i))}{s}$$

- the perplexity per word, PPL, represents the amount by which the n-gram language model LM[3] reduces uncertainty about the next word:

$$\log(PPL) = -\frac{1}{s}\sum_{i=1}^{s}\log P_{LM}(w_i|w_{i-n+1}^{i-1})$$

While the genre with the lowest PPL is the one that gets the higher BLEU score, OOV and PPL alone are not sufficient to predict translation performance. In particular, genre F yields one of the highest BLEU scores despite the high perplexity of the Chinese language model on this genre. This suggests that predicting translation quality is not straightforward and that genre distinctions do capture useful distinctions that cannot be predicted by the similarity of test and training documents alone.

C. SMT and TM behave differently across genres

Comparing results across CAT tools, Table III shows that there is a bigger variability in behaviour with TM than with SMT: the performance on the three worst genres is much lower for the TM. In fact, it is on these genres that the SMT outperforms the TM most, at least according to the BLEU metric. Overall, it is clear that the SMT consistently does at least as well as the TM, and sometimes much better.

One key to explaining this difference lies in the distinct ways in which the two tools function. A TM performs well when its *coverage* is good: there are many segments in the text to translate that match or are very similar to the memorized training data. On genres where the coverage is lower, the best the TM can do is return the target side of poorly-matching segments. On the contrary, the SMT system is able to generate translations that have never been seen in the training corpus. SMT will obviously perform better when the test set is close to its training data, but when it is not, SMT at least attempts to produce a prediction that contains translations of the source segment, often at the expense of grammaticality, instead of just outputting a grammatical, but poorly-related segment.

Another factor is that SMT and TM are affected differently by alignment errors in the training data. Since our training corpus was obtained by automatically aligning sentence pairs (which is standard procedure in SMT), there is a significant number of sentence alignment errors. Due to the way the SMT system is estimated, these alignment errors tend to be "averaged out" by the statistical estimation process. On the other hand, the TM will just return a misaligned target segment as long as the source segment matches. TM is therefore less robust to alignment errors than SMT.

D. Impact on translation workflow

This evaluation of translation quality suggests that organizing documents by genre is useful in order to decide how to route documents for translation. Our results suggest that different strategies might be needed depending on genres. For instance, we can have a two-stage strategy where CAT tools alone are used to translate the best performing genres (E and F). SMT can be used either alone or in combination with TM. For genres with lower BLEU scores and more documents in the lower BLEU quartiles, SMT will probably require human post-edition in order to be usable.

VII. IMPACT OF LEARNED CORPUS STRUCTURE ON CAT

After evaluating translation quality by genre, we now perform a similar evaluation after automatically partitioning the training and test documents in an unsupervised fashion. Clusters are obtained automatically using the combination of unsupervised topic modeling and clustering described in Section III. We set the number of clusters $K = 8$, in order to have the same number of clusters and genres, and $T = 80$ after experimenting with various numbers of topics.

While it is unclear how to evaluate quality of topics intrinsically, manual inspection reveals that even with a relatively high number of topics the distinctions captured seem reasonably well motivated. We ranked words in each LDA topic according to the absolute value of their contribution to the Kullback Leibler divergence between each topic distribution considered and the word distribution in the entire UN9399 corpus. In order to give the reader a sense of the topics captured, we report here the top 10 words for the first few topics learned:

- **topic 1:** general, committee, resolution, assembly, united, council, secretary, draft, report, session
- **topic 2:** government, security, council, republic, special, military, refugees, united, humanitarian, situation
- **topic 3:** staff, period, general, space, cost, united, amount, costs, cent, mission
- **topic 4:** development, countries, women, economic, social, developing, international, national, trade, world
- **topic 5:** nations, united, programme, activities, general, development, information, system, organizations, support
- **topic 6:** rights, human, international, united, convention, states, nations, special, commission, children

[3]In our experiments we use $n = 5$, i.e. 5-gram language model.

TABLE VI
BLEU SCORES FOR COMPUTER-AIDED TRANSLATION, BY CLUSTER.

Cluster	# sent	TM BLEU	SMT BLEU
1	221	54.97	54.61
2	195	12.51	33.18
3	35	60.68	54.25
4	7	0.00	0.00
5	101	62.98	61.82
6	181	14.30	34.42
7	427	16.82	36.48
8	1775	24.09	31.65
All:	2942	27.29	36.97

TABLE VII
DOCUMENT-LEVEL BLEU SCORES BY CLUSTER. BLEU QUARTILES ARE DEFINED BY RANKING DOCUMENTS ACROSS ALL GENRES, AND ARE ORDERED FROM BEST (Q1) TO WORST (Q4). CLUSTER 4 IS TOO SMALL TO COMPUTE STATISTICS.

	BLEU by doc			% docs per BLEU quartile			
Cluster	Mean	Min	Max	q1	q2	q3	q4
1	54.61	54.61	54.61	1	0	0	0
2	33.30	24.39	42.44	0	0.2	0.6	0.2
3	48.39	25.38	61.80	0.2	0.6	0	0.2
4	-	-	-	-	-	-	-
5	60.15	53.75	66.99	0.714	0.285	0	0
6	34.07	33.25	34.89	0	0	1	0
7	41.88	30.96	54.68	0.166	0.333	0.5	0
8	42.85	17.06	70.45	0.277	0.222	0.277	0.222

- **topic 7:** article, state, law, international, court, convention, states, legal, paragraph, commission

A. Learned clusters capture useful distinctions for CAT

The BLEU scores per cluster summarized in Table VI are overall fairly similar to what we observed on the genre-based partitioning in Section VII.

CAT tools perform very well on some clusters (e.g., clusters 1, 3 and 5), less on others (2,4,6,8). As a result, clusters do identify useful distinctions for the purpose of routing documents. Just like with genres, the Translation Memory performs poorly on some clusters, where the SMT performance is also lower, but much better than that of the TM. Both TM and SMT do poorly on cluster 4, which happens to be very small and mostly composed of "documents" which contain list of numbers and other non-words which should probably not be "translated", in the usual sense of the word. If anything, it seems that clusters reinforce what we saw on genres: on the best performing clusters, the TM does even better than the SMT, while on the weaker clusters SMT performance is much less impacted.

As for genres, we study the document-level BLEU scores for each cluster in Table VII. These scores also highlight the good translation quality of cluster 5, which has by far the highest mean score, and more than 70% of documents ranked within the top BLEU quartile.

B. Clusters capture different distinctions than genres

Automatically learned clusters are quite different from genres. The cluster distribution is more balanced than the genre distribution overall. In addition, the distribution of test documents across genres for each cluster shows that there is not a one-to-one mapping between clusters and genres.

TABLE VIII
TOPIC-BASED CLUSTERS AND GENRES RESULT IN VERY DIFFERENT PARTITIONS OF THE UN2000 DOCUMENTS, AS QUANTIFIED BY PURITY, RAND INDEX (RI), PRECISION, RECALL AND F_1 SCORE

purity	RI	p	r	F_1
71.75	74.06	34.31	34.59	34.45

In order to quantify these differences, we use clustering evaluation metrics to compare automatically learned clusters with the manual genre annotation on the UN2000 corpus. Since clusters capture other distinctions than genres by design, note that we are repurposing the evaluation metrics to represent the distance between clusters and genres, rather than the quality of the clusters. We compute external clustering evaluation metrics as defined in [17]. Purity simply maps each cluster to its most frequent genre, and represents the percentage of correctly assigned documents. The remaining metrics view clustering as a series of decisions about each pair of documents. A decision is correct if two documents assigned to the same cluster are also assigned to the same genre. Rand Index represents the percentage of decisions that are correct (i.e. accuracy), while Precision/Recall/F-Measure separate the impact of false positives and false negatives. The relatively high purity and Rand Index scores reported in Table VIII can be explained by the imbalanced genre distribution which results in assigning the most frequent genre to most clusters. The lower precision and recall scores show that the clusters yield distinctions that are very different from genres.

This yields interesting differences between genre-based and cluster-based translation quality. While SMT yields much better BLEU scores than TM for all but one genre, the difference in BLEU scores between TM and SMT within clusters is more varied. In particular, for the 3 clusters that yield the highest BLEU scores, the TM does better than the SMT.

One key difference is that clusters are informed by the training documents since they are learned using the SMT/TM training corpus. In contrast the genres are defined indepedently from the training corpus.

C. Impact on translation workflow

These results seem to suggest a possible strategy for applying CAT tools, which could go as:

- For clusters 1, 3 and 5, the high BLEU scores obtained with SMT and the even higher scores obtained with TM suggest that a combination of TM and SMT [5] should provide results that need little or no post-edition (depending on the end user's quality requirement);
- Clusters 2, 6, 7 and 8 should be suitable for SMT with human postediting. These clusters get lower BLEU scores, a higher percentage of documents in the worst BLEU quartiles and a bigger difference between SMT and TM performance.

Taken together, this suggests that SMT can yield reasonable translation quality for some documents, but that it cannot be trusted for all documents without human verification.

VIII. Conclusion

We showed that translation quality for heterogeneous documents from the same UN source varies widely across genres and topic-based clusters, for two very different computer-aided translation tools: translation memories and statistical machine translation. This suggests that when translating large heterogeneous text collections such as those built from social media data, different translation strategies should be followed depending on the genres or clusters of documents to translate. SMT and TM can be used in combination for a subset of the documents, but both genres and cluster distinctions identify documents where translations probably require human post-edition in order to be useful.

This analysis departs from previous work in machine translation evaluation which focuses on comparing the performance of different systems on a fixed test set [1, for instance]. In contrast we have compared the performance of a single TM and a single SMT across different types of documents.

We plan to extend this work along several directions. First, we would like to further investigate unsupervised document clustering, and study whether topic-based clusters and genre distinctions provide complementary information for predicting translation quality. Second, we will leverage the multilingual nature of the UN corpus and evaluate whether our findings on Chinese-English translation hold for translation between other languages such as Arabic-English. Third, we will study whether genre and cluster corpus structure can be leveraged to improve the performance of CAT tools for specific document types, as in domain adaptation [18], [19]. Finally, we would like to experiment with document routing in actual translation workflows involving human translators and post-editors for translation tasks representative of the needs of their clients.

Acknowledgment

The authors would like to thank John S. White and Timothy B. Allison for making data available for this project; Massih-Reza Amini, Michel Simard, and the Portage team at NRC for helpful comments and suggestions; and the three anonymous reviewers for their insightful feedback.

References

[1] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland, July 2011, pp. 22–64.

[2] "NIST open machine translation (OpenMT) evaluation," http://www.itl.nist.gov/iad/mig/tests/mt/, last visited April 23, 2012.

[3] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X*, 2005.

[4] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 11, pp. 1506–1518, 2006. [Online]. Available: http://dx.doi.org/10.1002/asi.20427

[5] M. Simard and P. Isabelle, "Phrase-based machine translation in a computer-assisted translation environment," in *Proceedings of the twelfth Machine Translation Summit*, Ottawa, Ontario, Canada, August 2009, pp. 120–127.

[6] C. Lin and F. Och, "Orange: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, pp. 501–es.

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.

[8] G. Foster, B. Chen, E. Joanis, H. Johnson, R. Kuhn, and S. Larkin, "PORTAGE in the NIST 2009 MT Evaluation," NRC-CNRC, Tech. Rep., 2009.

[9] P. Koehn, *Statistical Machine Translation*. Cambridge University Press, 2009.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, March 2003. [Online]. Available: http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

[11] M. Steyvers and T. Griffiths, "Probabilistic topic models," In Handbook of Latent Semantic Analysis, Hillsdale, NJ, 2007.

[12] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[13] Y.-M. Kim, "Document clustering in a learned concept space," Ph.D. dissertation, University Pierre and Marie Curie, 2010.

[14] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, Uppsala, Sweden, July 2010, pp. 17–53, revised August 2010.

[15] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Machine Translation," in *Proceedings of COLING'96*. Copenhagen, Denmark, 1996, pp. 836–841.

[16] R. Soricut and A. Echihabi, "TrustRank: Inducing Trust in Automatic Translations via Ranking," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 612–621. [Online]. Available: http://www.aclweb.org/anthology/P10-1063

[17] C. D. Manning, P. Raghavan, and H. Schuetze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[18] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 128–135. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0717

[19] G. Foster, C. Goutte, and R. Kuhn, "Discriminative instance weighting for domain adaptation in statistical machine translation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Cambridge, MA, October 2010, pp. 451–459. [Online]. Available: http://www.aclweb.org/anthology/D10-1044

NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

The impact of sentence alignment errors on phrase-based machine translation performance

Goutte, Cyril; Carpuat, Marine; Foster, George

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas, 2012-11

NRC Publications Record / Notice d'Archives des publications de CNRC:
https://nrc-publications.canada.ca/eng/view/object/?id=6aeda7ee-6f72-466f-9c7a-56c71e481d52
https://publications-cnrc.canada.ca/fra/voir/objet/?id=6aeda7ee-6f72-466f-9c7a-56c71e481d52

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at
PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada
Conseil national de recherches Canada

Canada

The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance

Cyril Goutte **Marine Carpuat** **George Foster**
Interactive Language Technology
National Research Council Canada
Gatineau, QC, J8X 3X7
Cyril.Goutte@nrc.ca Marine.Carpuat@nrc.ca George.Foster@nrc.ca

Abstract

When parallel or comparable corpora are harvested from the web, there is typically a tradeoff between the size and quality of the data. In order to improve quality, corpus collection efforts often attempt to fix or remove misaligned sentence pairs. But, at the same time, Statistical Machine Translation (SMT) systems are widely assumed to be relatively robust to sentence alignment errors. However, there is little empirical evidence to support and characterize this robustness. This contribution investigates the impact of *sentence* alignment errors on a typical phrase-based SMT system. We confirm that SMT systems are highly tolerant to noise, and that performance only degrades seriously at very high noise levels. Our findings suggest that when collecting larger, noisy parallel data for training phrase-based SMT, cleaning up by trying to detect and remove incorrect alignments can actually degrade performance. Although fixing errors, when applicable, is a preferable strategy to removal, its benefits only become apparent for fairly high misalignment rates. We provide several explanations to support these findings.

1 Introduction

Parallel or comparable corpora are routinely harvested from the web or other large resources using automatic or semi-automatic methods (Tiedemann, 2009; Callison-Burch et al., 2009; Fung et al., 2010). Although this enables rapid collection of large corpora, it also requires fairly automated systems, which process large amounts of data with little human supervision. In particular, corpora are segmented into sentences which are then aligned using automatic sentence alignment techniques, e.g. (Moore, 2002). Despite the good performance of state-of-the art automatic sentence alignment, the size of the corpora and the cascading of earlier document or paragraph matching steps tend to result in many misaligned sentence pairs. We estimate in section 3.1 that three of the corpora used in the WMT evaluation[1] contain between 1.2% and 13.1% misaligned sentence pairs.

It is often argued that SMT systems, due to their statistical nature, are relatively robust to sentence alignment errors. However, there is to our knowledge little empirical support for this belief (Khadivi and Ney, 2005). In this paper, we attempt to address this by analysing the impact of sentence misalignment rate on SMT output quality. We provide three main contributions:

1. We describe a sampling approach for estimating alignment noise reliably with limited human effort, and provide resulting estimates on common corpora;

2. We measure the robustness of a typical phrase-based MT system to increasing levels of misalignment errors by simulating these errors in a high-quality corpus;

3. We suggest a strategy for training MT systems from noisy data coming for example from parallel or comparable corpora, relying on the two previous contributions.

[1] http://statmt.org/wmt12/translation-task.html

We emphasize that this work deals with *sentence alignment errors* as opposed to word alignment errors, although we later discuss the interaction between them.

Our investigations require a large, high quality, sentence aligned corpus, into which we add increasing amounts of random misalignments (sections 2-3). Section 4 presents the experimental results obtained by training a state-of-the-art phrase-based SMT (PBMT) system on the degraded corpora and measuring the impact on its performance. We show that indeed, PBMT systems are surprisingly robust to alignment noise, and that performance is actually higher on the noisy corpus than on a clean version of the corpus where alignment errors have been filtered out. Section 5 discusses these findings.

2 Data

In order to investigate the impact of alignment error on SMT performance, we need a parallel corpus that is *large enough* to be representative of the large data conditions needed to train state-of-the-art SMT systems, and *clean enough* to let us control the level of sentence-alignment noise. Unfortunately, as we demonstrate in the next section, most widely available large corpora have moderate-to-high misalignment rates, making it impossible to obtain a "clean" reference for the SMT performance. The Europarl corpus appears to have high sentence alignment quality, but it is fairly small by current standards, at least on well-studied language pairs. We acquired a large corpus of French-English parallel data from the Canadian Hansard, including proceedings from the House of Commons and from committees.[2] Using careful alignment, we obtained a total of 8,194,055 parallel sentences. We reserved subsets of 16,589 and 17,114 sentences in order to sample development and test sets, respectively, leaving up to 8,160,352 sentences for training the PBMT system. Section 3.1 shows that the estimated misalignment rate for this corpus is 0.5%, and section 3.2 describes how we gradually introduce increasing amounts of alignment error into the corpus for the purpose of our experiments.

[2]This corpus is available on request.

3 Method

We first introduce a sampling method for estimating the baseline level of alignment error in a parallel corpus, and apply it to the Hansard corpus, as well as several others for comparison. We then artificially introduce random alignment errors into the Hansard, as described below. We also briefly describe the PBMT system used in the experiments.

3.1 Estimating Sentence Alignment Error

We model the estimation of alignment errors using a simple binomial model. In this model, the bilingual corpus containing well-aligned and misaligned sentence pairs can be viewed as a large "urn" containing black and white balls. By sampling sentence pairs and evaluating whether they are correctly aligned or not, we draw balls from the urn and look at their colour. The outcome of this experiment is used to estimate the rate of misalignment in the corpus, just as we would estimate the proportion of white balls from our draw.

Let S be the number of sentence pairs that we sample, out of which "m" are misaligned, and $S-m$ correctly aligned. Given the (unknown) misalignment rate μ, the distribution of the number m of misaligned pairs is given by a binomial

$$P(m|\mu, S) = \frac{S!}{m!(S-m)!}\mu^m(1-\mu)^{S-m}.$$

A natural choice for the prior distribution on μ is a symmetric Beta, $p(\mu) = \frac{\Gamma(2\lambda)}{\Gamma(\lambda)^2}\mu^{\lambda-1}(1-\mu)^{\lambda-1}$. With $\lambda = 1$ this is a uniform distribution, while $\lambda = 1/2$ is the non-informative prior (Box and Tiao, 1973), which we will use here. Bayes' formula yields the posterior distribution for μ:

$$p(\mu|S, m) \propto \mu^{m-\frac{1}{2}}(1-\mu)^{S-m-\frac{1}{2}},$$

which is again a Beta distribution, with parameters $m + \frac{1}{2}$ and $S - m + \frac{1}{2}$.

From the posterior, we can derive a number of interesting estimates such as the expected misalignment rate, $\widehat{\mu} = (m + \frac{1}{2})/(S + 1)$, which is a smoothed version of the empirical estimate. More importantly, we can use the posterior distribution to derive confidence intervals and guarantees on the maximum misalignment rate.

Corpus	S	m	$\widehat{\mu}$ (%)	95% CI
Europarl	300	3	1.2	[0; 2.3]
UN	300	8	2.8	[0; 4.5]
Giga	300	39	13.1	[0; 16.4]
Hansard	300	1	0.5	[0; 1.3]

Table 1: Estimated expected misalignment rate ($\widehat{\mu}$) for four MT corpora. S is the number of sentence pairs evaluated, and m the number of incorrectly aligned ones. The confidence interval is the one-sided 95% interval.

In order to illustrate this, we consider three corpora from among the official WMT dataset, in addition to the Hansard described above:

- Europarl: 1,328,360 sentence pairs;
- United Nations: 12,886,831 sentence pairs;
- Giga (10^9) corpus: 22,520,400 sentence pairs;
- Hansard: 8,160,352 sentence pairs.

We sample a small number of sentence pairs (usually 300) and manually evaluate the correctness of the alignments. The results are given in Table 1. Depending on the corpus, between 1 and 39 pairs were found to be misaligned, resulting in expected misalignment rates between 0.5% and 13%.

The differences between these corpora is illustrated on Figure 1 where we plot the posterior distribution resulting from our evaluation. We see that the estimated misalignment rates as well as the uncertainty on this estimate (the spread of the posterior distribution) vary widely. Note that the expected misalignment rate does not coincide with the location of the highest probability (mode) of the distribution, which is normal for a skewed distribution.

Giga is the largest corpus, but also has the highest misalignment rate, at an estimated 13%, which corresponds to close to 3 million incorrect sentence pairs in this corpus (more than twice the entire Europarl corpus). The UN corpus has a lower misalignment rate, estimated below 3%. Europarl is even cleaner, with $\widehat{\mu} = 1.2\%$, but also much smaller. Finally the Hansard has an estimated misalignment rate of around 0.5%. The last column shows that, for the Hansard corpus, we can say with 95% confidence that the misalignment rate is below 1.3%. The estimated 0.5% misalignment corresponds to around

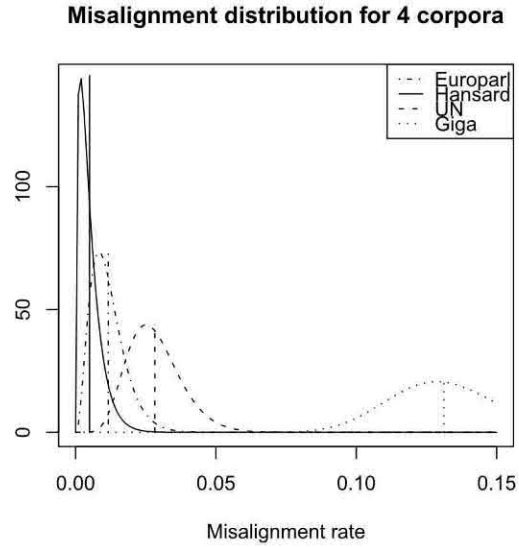

Figure 1: Estimated distribution of misalignment rates.

40k incorrectly aligned sentence pairs out of this 8.16M sentence corpus. This motivates our use of that corpus as a basis for our experiments: it is relatively large and very "clean".

The overall message of this section is that using a sampling-based approach, it is possible to obtain a fairly reliable bounded estimate of the misalignment quality of a corpus with relatively modest effort.

3.2 Introducing Alignment Errors

Starting from the large, high-quality Hansard corpus, we gradually introduce random alignment errors, by increments of 10%. We randomly sample a number of sentence pairs corresponding to the target error level,[3] and do a permutation of the target (English) side. For example, for 10% noise, we sample around 775,000 sentence pairs, then the target side of the first pair in the sample is assigned to the source side of the second pair, the target of the second to the third source, etc. We also ensure that each perturbation is strictly included in a larger perturbation, ie the 20%-noise misalignments contain the 10%-noise misalignments, etc. To average results over the choice of alignment errors, we sample 6 random samples at each of 10%, 20%, ... 90% misalignment rate, hence a total of 54 noisy corpora, plus the original, "clean" one.

As we introduce noise by sentence permuta-

[3] Minus 0.5% to account for the baseline misalignment rate.

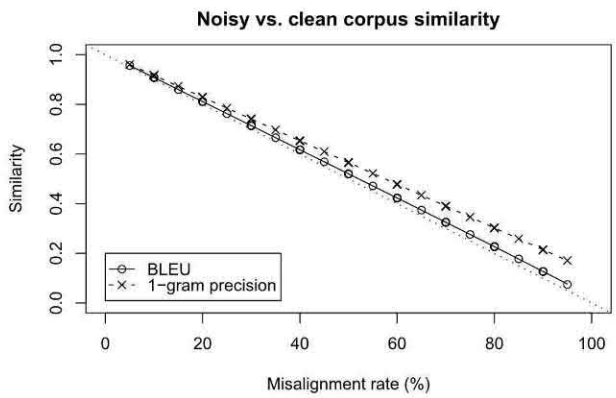

Figure 2: Similarity (BLEU and 1-gram precision) of noisy corpora vs. clean reference.

tion within a well-defined domain, some words or phrases may still align to source words. We quantify this by computing the overlap between the noisy and original versions of the target side at each noise level, using BLEU and 1-gram precision. Figure 2 shows that the overlap between noisy and clean corpora decreases linearly and roughly matches the percentage of clean sentences in the corpora. This suggests that there are only few matching words between permuted and original target sentences, hence little chance of extracting correct phrase pairs from incorrect alignments. This is discussed further in section 5.

As additional alignment errors are introduced artificially, we know exactly which pairs are misaligned, apart from the 0.5% baseline errors. To support further experiments, we produce "cleaned" versions of the corpora at each noise level, where we remove the artificially introduced errors, leaving only the unmodified sentence pairs. These have the same baseline misalignment rate of 0.5% but are smaller in size (90.5%, 80.5% . . . 10.5% of the full corpus). Although we can't remove the final 0.5% of misaligned sentence pairs, for convenience we call this "perfect filtering" below.

3.3 Training the SMT model

For each sample at each perturbation level, we train a standard PBMT model and estimate its performance on the reference test corpus. We use a typical PBMT system which has achieved competitive results in recent NIST and WMT evaluations (Larkin et al., 2010). We use the following feature functions in the log-linear model:

- 4-gram language model with Kneser-Ney smoothing (1 feature);
- relative-frequency and lexical translation model probabilities in both directions (4 features);
- lexicalized distortion (6 features); and
- word count (1 feature).

The parameters of the log-linear model are tuned by optimizing BLEU on the development set using MIRA (Chiang et al., 2008).[4] Phrase extraction is done by aligning the corpus at the word level using both HMM and IBM2 models, using the union of phrases extracted from these separate alignments for the phrase table, with a maximum phrase length of 7 tokens. Phrase pairs were filtered so that the top 30 translations for each source phrase were retained. The translation performance was measured using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

4 Results

We show how SMT output degrades with increasing alignment noise. We see that, surprisingly, even relatively high levels of noise have little impact on translation performance. We then compare the robustness of PBMT systems to that of Translation Memories, a commom computer-aided translation tool.

4.1 Impact on translation performance

Figure 3 shows how translation performance, as estimated by BLEU (circles), degrades when the number of misaligned sentence pairs increases. Not surprisingly, increasing the noise level produces a general decrease in performance. Although there are variations depending on the samples, the smoothed curve (solid line) is strictly decreasing as expected. What may be more surprising is how little the performance is affected as alignment error approaches relatively high levels. After adding 30% alignment errors, the average BLEU score drops from the 37.59 obtained on the "clean" corpus, down to 37.31, less

[4] MERT gives qualitatively similar results.

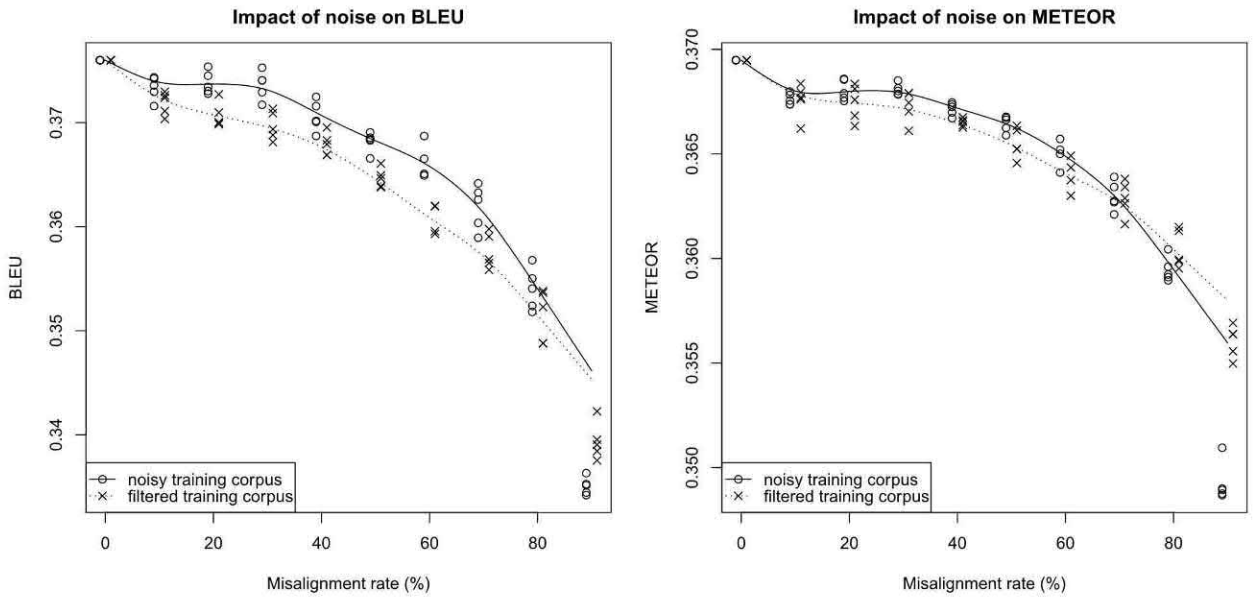

Figure 3: Impact of sentence misalignment error on SMT performance estimated with BLEU (left) and METEOR (right). Results are for training on full corpora (o) and when misalignment errors are filtered out (x). Curves are smoothed using a local linear smoother (`locpoly` in R).

than 0.3 BLEU points below. Translation performance degrades faster after that, but only takes a large hit when the misalignment rate grows beyond 60-70%, ie far more incorrect alignments than correct ones, a situation that should be very easy to detect (and hopefully rare) in practice. Note that the average BLEU score at 70% noise is still 36.13, less than 1.5 BLEU points below the "clean" BLEU.

In order to show that these results are not an artefact of using BLEU for estimating the translation quality, we also produce curves showing the impact of misalignment noise on METEOR (Banerjee and Lavie, 2005), another popular metric. The right plot in Figure 3 shows these results. We see that although the metrics are different, the general picture is quite similar: low to moderate noise has little to no impact on the performance estimated by METEOR, apart from the initial variability due to the fact that we have only one training set at 0.5% error. Performance starts to really degrade from around 30% noise, and gets much worse after 60-70% noise.

4.2 Comparison with perfect filtering

To put this into perspective, we perform another set of experiments on training corpora where we filter out the misaligned pairs that we introduced earlier. This results in high quality but smaller corpora of 90%–10% of the original corpus size. The performance of the PBMT systems trained on these "filtered" corpora is plotted as crosses and dotted line in Figure 3. The surprising outcome of this experiment is that the performance on the filtered corpus is no better than when misaligned sentences are kept in the training data. In fact, this "perfect filtering" produces a small but consistent *decrease* in performance until very high noise levels are reached. One explanation for this is that the increase in quality in the *translation model* that we expect to result from the filtering is insufficient to compensate for the decrease caused by the reduced amount of data available for training the *language model*.

In order to test that hypothesis, we trained a hybrid model at each noise level, using the entire corpus for training the language model, and the filtered (cleaner but smaller) corpus for training the translation and distortion models. The language model uses only the target side of the corpus, and is invariant through the permutations used to introduce the noise in our experiments. The results plotted in Figure 4 completely validate our explanation: the use of a language model trained on a larger corpus greatly improves over the "filtered" performance: The hybrid (dashed) curve is always above the "filtered" (dotted) curve. Note also that until around 30%

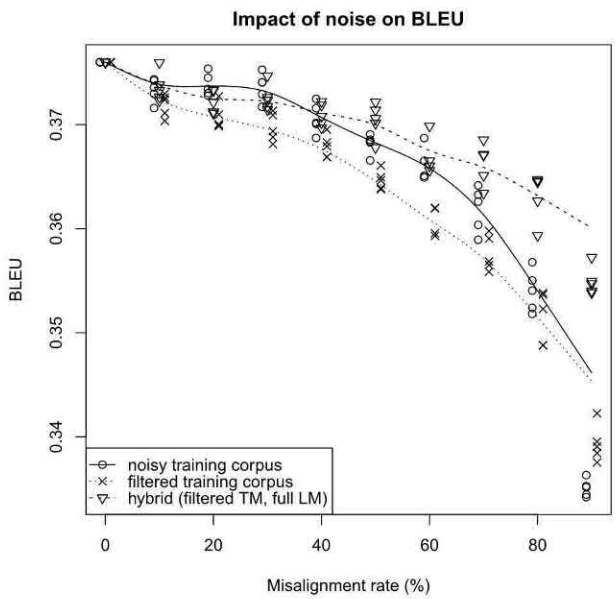

Figure 4: BLEU score vs. misalignment error, for PBMT systems trained on full corpora (o), filtered corpora (x) and hybrid approach (triangles). For clarity, we introduced a slight horizontal offset for markers.

noise, performance of the "noisy" (solid) and "hybrid" (dashed) models is very close, showing that the translation model does not suffer much from being trained on a corpus containing up to 30% sentence alignment errors. Above 40% noise, the hybrid model dominates the other two quite clearly.

The "perfect filtering", removing material from both TM and LM training, may seem like an odd choice given the previous analysis. This is however the typical way large corpora are built and made available. For example, among the large corpora used for the WMT workshop shared tasks, most of them were harvested (semi-)automatically; none includes monolingual versions that incorporate filtered out material.

4.3 Comparison with Translation Memory

We contrast the impact of noise on PBMT with its impact on the output of a simple translation memory (TM). Translation memories are a Computer-Aided Translation tool widely used by translators. Although they are only used in practice to provide translations to test segments that are highly similar to the content of a bilingual corpus, they may be generalized to provide what is essentially a one-nearest-neighbour translation prediction. Our TM implementation searches the parallel training corpus for the source sentence with maximum similarity[5] to each test sentence, and outputs the corresponding target sentence. TM quality is much worse than SMT: the TM applied to the clean corpus yields a BLEU score of 13.64. It is also much more sensitive to sentence alignment errors: at 30% and 90% noise levels, the average scores for the TM are respectively 9.75 and 1.97 (relative decrease of 28% and 86%). This confirms the remarkable robustness of SMT, for which BLEU scores degrade much more gracefully.

5 Discussion

Although it is well known that PBMT systems are robust to noise, our results indicate that this robustness holds to a remarkable extent, in fact to levels of noise that are far higher than usually found in parallel corpora. Three questions are relevant: what is the explanation for this phenomenon; can we expect these results to generalize to other settings and to non-synthetic noise; and what are the practical ramifications for training PBMT systems?

5.1 Analysis of SMT Robustness

To explain our results, we consider the effect of alignment noise on the phrase table. For a typical entry, $P(t|s)$ is peaked on a few values of t. Other incorrect target phrases appear erroneously in the phrase table due to sentence and word alignment errors. As we introduce more alignment errors, the number of target segments incorrectly associated with a source segment grows, and the estimated probability of the correct translations drops correspondingly. However, due to the random nature of the alignment errors, incorrect translations keep a low probability: $P(t|s)$ gets "broader" and "flatter", but as long as there are enough good alignments to collect statistics, the most likely t are still correct and translation is only moderately affected. Note also that the phrase extraction heuristics may also help reject incorrect translations as they will only consider word alignments that satisfy a number of regularity constraints.

We support our explanation by inspecting the part of the phrase table that is used for producing trans-

[5] We use a smoothed BLEU score as the similarity metric.

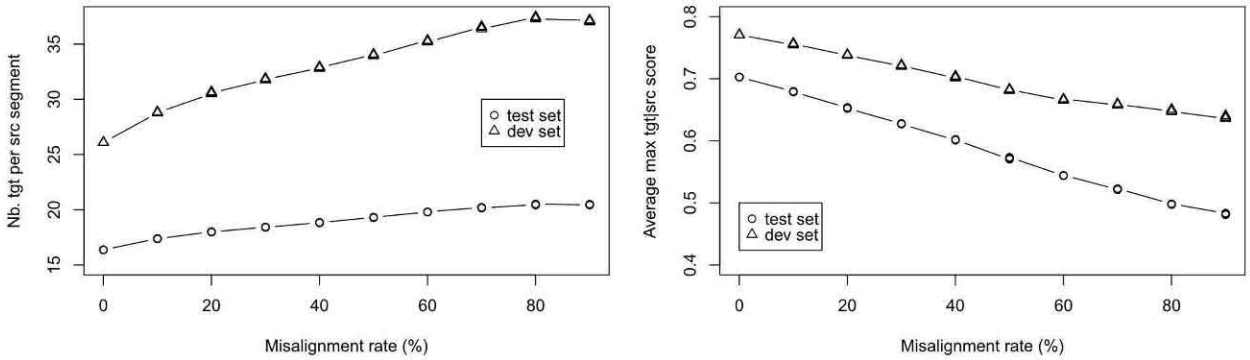

Figure 5: Average # target segment (left) and max. posterior probability (right) over source segments in phrase table.

lations on our test and dev sets, and computing two quantities that characterize the posterior $P(t|s)$: average number of target segments per source segment, and average maximum value of the posterior probability. Figure 5 shows that, as the number of misaligned sentences increases, the posterior indeed gets "broader" (there are more translations t for each s) and "flatter" (the probability of the most probable translation gets lower).

From a somewhat different perspective, the presence of sentence alignment errors in the training corpus may be seen as the addition of noise to what is essentially the input of the MT model estimation procedure. In machine learning, it has long been known that the presence of noise in the input data has an effect similar to regularization (Bishop, 1995). Tuned to the right level, it may therefore potentially yield improvements in performance. Although it is impractical to tune the appropriate level of sentence misalignment in a bilingual corpus, this, together with the smoothing effect illustrated above, helps shed further light on the robustness of PBMT to alignment errors.

5.2 Simulated Noise vs. Real Noise

A concern with all experiments that rely on simulation is how relevant the simulated conditions are to actually observed, real conditions. We first note that the noise we consider in this paper affects translation quality only—it assumes well-formed source and target sentences. Clearly, many real sources of noise will not have this profile, but these are arguably easier to detect and address, since detection can be based on monolingual properties (which will necessarily affect translation quality). However, we make no claims here about SMT robustness to this kind of noise.

The kind of noise we model can typically arise through automatic alignment procedures for parallel or comparable corpora. Our model makes two assumptions that do not always characterize real data. First, it assumes that errors are always "maximally bad", as demonstrated in figure 2: it is very unlikely that any valid phrase pairs will be extracted from mis-aligned sentence pairs. A substantial proportion of errors in real parallel and comparable corpora will not be this bad, and will permit extraction of some valid pairs from properly-aligned sentence fragments. Clearly these errors will not damage translation performance as much as our simulated errors. Our results can thus be seen as estimates of *minimum* noise robustness (when errors are counted at the sentence level).

A second assumption we have made is that noise will be uniformly distributed. This is a key assumption, as our analysis in the previous section shows. Clearly, real errors will sometimes exhibit systematic tendencies that violate this assumption. However, if these are frequent, they will be easy to detect and correct; and if they are infrequent, they will be inconsequential. The middle ground—"malicious" errors capable of tricking an SMT system into producing bad translations—seems implausible. Certainly we found no traces of any such effect in our manual alignment evaluations.

5.3 Training with Noise

The results in this paper, and the foregoing discussion, suggest a possible strategy for dealing with alignment noise when training an SMT system. The first step, as usual, is to manually inspect and address any obvious sources of noise. Since real corpora vary enormously, this step is difficult to automate. Next, estimate alignment error using the procedure in section 3.1. If this is greater than approximately 30%, discard low-confidence pairs until it is below 30%. (We assume the existence of sentence-pair confidence scores as a typical by-product of the alignment process.) When discarding sentence pairs, retain the target sentences for language model training.

We emphasize that this procedure is of course highly tentative. It will need to be adjusted for many different factors, such as language pair, the importance of small performance differences to the application, the reliability of confidence scores, etc. However, for large noisy corpora, or for small noisy corpora used as part of a larger training set, it offers significant potential speed and convenience advantages over the alternative of re-training from scratch and measuring performance at different noise levels.

6 Conclusion

We analysed the impact of sentence alignment errors on SMT quality. We first described a method for quickly estimating alignment noise by sampling, and provide estimates on common corpora. Then, through simulation, we analyzed the robustness of phrase-based MT to alignment errors.

Our results showed that phrase-based MT is highly robust to alignment errors: performance is hardly affected when the misalignment rate is below 30%, and introducing 50% alignment error brings performance down less than 1 BLEU point. We suggest that this may be thanks to the phrase table extraction and estimation procedure. Our results are limited to one corpus, language pair and SMT system, but suggest that efforts spent on elaborate procedures for filtering out sentence alignment errors from automatically harvested corpora may bring little payoff, or even degrade performance. More specifically, the increase in quality achieved by filtering out incorrect alignments may not offset the decrease resulting from lower corpus size.

These findings can inform strategies for training MT systems with noisy data. For instance, we suggest handling corpora with low alignment quality by filtering bilingual pairs so that alignment error is below 30% while keeping all target side segments for training the language model. This seems especially promising for sentence pairs extracted from comparable corpora, which we will investigate in future work.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

C. M. Bishop. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

G. E. P. Box and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. Wiley.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Fourth Workshop on Statistical Machine Translation*, pages 1–28.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation. In *EMNLP*, pages 224–233.

P. Fung, E. Prochasson, and S. Shi. 2010. Trillions of comparable documents. In *LREC Workshop on Building and Using Comparable Corpora*, May.

S. Khadivi and H. Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *Natural Language Processing and Information Systems*, LNCS 3513, pages 263–274. Springer.

S. Larkin, B. Chen, G. Foster, U. Germann, E. Joanis, J. H. Johnson, and R. Kuhn. 2010. Lessons from NRC's portage system at WMT 2010. In *5th Workshop on Statistical Machine Translation*.

R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In S. D. Richardson, editor, *AMTA*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144. Springer.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the ACL*.

J. Tiedemann. 2009. News from OPUS—a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, pages 237–248.

NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Machine translation: benefits and advantages of statistical machine translation and NRC's Portage

National Research Council of Canada. Information and Communication Technologies

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

https://doi.org/10.4224/21274926

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=72d5a618-0cd3-4b68-b56f-41178020c9d0
https://publications-cnrc.canada.ca/fra/voir/objet/?id=72d5a618-0cd3-4b68-b56f-41178020c9d0

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada Conseil national de recherches Canada

Canada

NRC·CNRC Information and Communications Technologies

Machine Translation: Benefits and Advantages of Statistical Machine Translation and NRC's *Portage*

April 2015

National Research Council Canada | Conseil national de recherches Canada

Canada

Page: 19 of 362
A2020-0026

Introduction

Machine translation (MT) – or automatic translation, as it is sometimes called – automatically transforms a text from one natural language into what is ideally a semantically equivalent and grammatically correct text in another natural language using a computer. From the point of view of an average user, an MT system is just a black box: you type in some text in a source language, or direct the system to a file, and seconds later you obtain a translated version of the text in the target language.

But, to dig a little deeper, different types of machine translation – statistical and rule-based – can have an impact on the results. Since 2003, the world-class research team at the National Research Council of Canada has been developing a proprietary statistical MT (SMT) system, called *Portage*.[1] This paper provides some background about SMT, and contrasts it with the other major current in machine translation: rule-based MT (RBMT). For most of this young field's history, RBMT was the dominant paradigm in MT. It can generally be traced back to 1948 when a memo sent out by Warren Weaver to a small group of scientific colleagues suggested that computers might be used for translation in the same way they had been used for code breaking during the Second World War. Later, as Chomskyan linguistics grew more dominant in the mid-1960's, MT came to adopt similar types of declarative rule formalisms.

System developers in that period tended to be linguists and computer scientists whose objective was to program the machine to emulate what they understood a human translator does: analyse the source text into some sort of meaningful representation and map that abstract representation into an equivalent target-language text. This required handcrafting large-scale computational grammars and dictionaries that contained many hundreds, if not thousands of linguistic rules – a challenging, time-consuming and costly endeavour. The resulting systems were not particularly robust and often failed to produce an output, largely because it is

[1] Actually, the Canadian government has had an active interest in MT since 1963, financing research in various universities and public labs over the years. *Portage* is the latest product of that long-standing commitment.

extremely difficult to write an exhaustive grammar and a complete dictionary for an object as complex and evolving as a natural language, not to mention the contrastive grammars and dictionaries required to map one language into another.

In 1988, a group at the IBM T.J. Watson Research Centre proposed statistical machine translation. Some of this group had previously been involved in the successful development of statistical speech recognition systems – another difficult problem that had long stymied the rule-based approach.

In contrast to RBMT, SMT systems have no declarative grammar rules or dictionary entries. Rather than relying on linguists' intuitions about language, these systems instead draw their linguistic knowledge directly from large bodies of previously translated text. Machine learning algorithms are applied to this text in order to automatically estimate the probability of some word or phrase *X* in the source language should be translated as *Y* (or W, or Z…) in the target language. An SMT system is composed of *two* major linguistic components which are both probabilistic: a translation model, which roughly corresponds to the bilingual dictionary in an RBMT system; and a language model, which ensures the fluency of output text by maximizing the probability of each word W of the translation given the choices already made for the *n* preceding words.

There are numerous other statistical modules, as well as a core algorithm called the decoder, which performs the actual search for an optimal target text in applying these probabilistic repositories of linguistic knowledge to the source text. Referring back to our preliminary definition above, one could say the translation model is responsible for the semantic fidelity of the translation, while the language model is responsible for its grammaticality.

SMT advantages

SMT offers a number of important advantages over RBMT, the first being greater **versatility**, or polyvalence. Under the rule-based approach, whenever a new language *T* has to be added, computational linguists specializing in that language have to be hired, along with bilingual lexicologists qualified in the pair *S* > *T* – if any such can be found.

In SMT, on the other hand, the same machine learning algorithms can be applied to virtually any language pair. Someone with an English > French SMT system who decides to tackle English > Spanish (or Portuguese > Bulgarian, and so on) can do so without delay – provided, of course, that the requisite training materials are available. And these, as mentioned above, are simply large bilingual volumes of well-translated texts.[2] Consequently, SMT reduces the time, effort and cost of developing new MT systems by orders of magnitude. A system that would take months and many hundreds of man-hours to develop under the rule-based approach can now (literally) be generated overnight.

Another important advantage that SMT systems have over their RBMT counterparts is their **robustness**. In SMT, every source language sentence has a multitude of possible target language renderings, although some are usually much more probable than others. As a result, SMT systems never fail to produce an output and, in this sense, they are naturally more robust than the older RBMT systems, which often depended on a complete grammatical analysis of the source sentence. Of course, this robustness would be of little interest if the quality of all the generated translations were poor.

[2] How large a bilingual corpus is required to train an SMT system? People who work in the field tend to say the larger the corpus, the better. However, there is a minimal size for a training corpus, below which the acquired probabilities will not be reliable. A figure that is often cited is a hundred thousand sentence pairs, or roughly a million words.

However, it has been conclusively demonstrated in open international competitions like those organized by the American military and the National Institute of Standards and Technology (NIST) that SMT systems tend to produce **better-quality** translations than their RBMT counterparts. The best MT systems available today are all statistical; SMT is state of the art.

NRC's *Portage* system

Statistical systems completely dominate the MT landscape today. Some of the better known of these systems, like *Google Translate* and *Microsoft Translator*, can be accessed for free over the Internet, or are available for very modest user fees. Other open-source SMT systems (most of which are based on *Moses*[3]) are being marketed by vendors who promise to tailor them to the particular needs of their clients. With all of these options, someone who is considering MT may initially wonder why they should bother looking at *Portage*, which is not offered as freeware, but there are a number of good reasons to do so.

Among statistical MT systems, *Portage* is one of the very best available in the world today. This was clearly demonstrated by the system's showing at recent NIST competitions, where *Portage* finished first among all systems translating from Chinese to English, and second among all systems translating from Arabic to English – no easy feat! NRC's *Portage* team has twice been invited to participate in large and prestigious projects funded by DARPA, the US Defense Department's Advanced Research Projects Agency, alongside such highly regarded labs as Raytheon BBN and RWTH (Germany). As well, the group has actively collaborated with major commercial partners like *Systran*, a company that has dominated the MT business for decades.

[3] *Moses* is a full-blown, open source SMT system originally developed under the leadership of Philipp Koehn as an academic alternative to the proprietary systems developed by large corporations like Google and Microsoft. Supported by funding from the European Commission, *Moses* has recently been extended to numerous European language pairs.

Many potential MT users have legitimate concerns over the security of their texts and are reluctant to send them out of the country for translation, or over the Internet into the cloud. Fewer still are prepared to see their texts used to improve another company's MT engine, over which they have no rights.[4] *Portage* provides a straight-forward solution to all of these concerns. The system is normally installed on a local server, within the user's secure premises. No one else has access to it, or to the texts it processes. And as we will discuss, *Portage* can be easily integrated within a user's existing document processing environment.

Opportunities to benefit from *Portage*

From the early origins of machine translation right through to the end of the Cold War, military and intelligence communities were the principal users of MT, and also provided the primary source of funding for MT research and development. The vast volumes of material that these services needed to process far exceeded the productive capacities of all available human translators -- MT offered the only possible hope.

Furthermore, geopolitical considerations determined the major foreign languages that MT systems were developed for. It was no accident, in other words, that for the first public demonstration of machine translation that was given in New York in 1954, the language direction was from Russian into English. Nor was it fortuitous that the American government heavily subsidized the development of a Vietnamese-to-English MT system in the 1960's, during the Vietnam War, and a Farsi-to-English system in the years before and after the fall of the Shah of Iran.[5] In all such cases, MT was seen as responding to the urgent needs of national security.

---

[4] This is the case with *Microsoft Translator*, although the company insists that no other client can recover and reconstitute the texts a user has uploaded onto its translation servers.
[5] In the case of both Vietnamese and Farsi, the principal beneficiary of these US government investments was the firm of Logos Corporation.

In the United States, military and intelligence communities continue to make extensive use of MT. If anything, their reliance on machine translation has dramatically increased in recent years, since they now monitor not just written texts but spoken communications, videos and social media as well. If nothing else, the attacks of September 11 served to sharpen the interest of the American Defence Department in machine translation, prompting it to reinvest heavily in MT R&D and organize the international competitions that were alluded to above.
And of course, Arabic suddenly became a principal focus of attention, with Chinese not far behind. Funding agencies have emphasized the importance of rapid system development, on the basis of smaller and more varied training materials. It is safe to say that much of the remarkable progress that has recently been achieved in machine translation is attributable to the large investments in the technology made by the military and intelligence communities.

In this context, it is important to note that MT is primarily being used for **information-gathering** purposes. Even if the quality of the MT output was absolutely perfect (which it is not), no intelligence service could possibly afford the time to read through such volumes of translated materials. Instead, MT serves as a crucial component of a larger chain designed for **triage**, helping to sift through massive quantities of foreign-language texts with a view to locating those of interest, which may then be translated by humans. And in this role, today's MT systems function very well.

Interestingly, the military and intelligence agencies are no longer the only ones interested in gathering information that appears in a foreign language. As a result of the increasing globalization of trade and commerce, businesses today are also eager to know what people are saying about their products and services in foreign-language markets. In recent years, a whole new industry has emerged that focuses on **business intelligence**, i.e. on the gathering and analysis of large amounts of data, particularly on the increasingly important social media, which is intended to help businesses adjust and hone their commercial strategies.

A few years ago, it was primarily large multinationals that were actively involved in so-called 'foreign' markets. Today, more and more SMEs are venturing abroad, and they too have a pressing need to comprehend what is being said about them in a range of foreign languages. Firms specializing in business intelligence are therefore looking to machine translation to help them respond to this demand for large-scale multilingual text processing and analysis. In fact, NRC's *Portage* team has already participated in some exploratory work in cross-linguistic sentiment analysis with the Canadian firm MediaMiser.[6]

Machine translation for dissemination purposes

As we mentioned above, in their information gathering activities, the intelligence and business communities are using MT essentially for **assimilation** purposes, to help people read and comprehend part of the content of a foreign language text. But what about MT for **dissemination** purposes, where the system is used to generate and eventually publish a text in a foreign language? This, after all, was the original aim and ultimate goal of machine translation.

Since MT's inception, there have been many attempts to employ the technology in this way, both in the public and private sectors. In the United States, companies like Systran, Weidner and Logos (to name just a few) were founded in the second half of the last century with just this goal. In Europe, with all the official languages of the EU creating an enormous demand for translation, the European Commission invested substantial effort and large sums of money in adapting a commercial MT system to help it meet its translation needs, before launching its own ambitious MT development program.[7]

---

[6] Sentiment analysis refers to the process of computationally categorizing opinions expressed in a piece of text.

[7] The program in question was called Eurotra, and it ran from 1978 to 1992. To this day, the EC continues to finance numerous R&D projects in MT and MT-related technology.

In Canada, the government's Translation Bureau mounted extensive trials of several commercial MT systems in order to determine if the technology could help it better meet a spiralling demand for its services and hopefully reduce its costs. Viewed retrospectively, however, the great majority of these efforts came to naught. Until recently, the successful implementations of MT for dissemination purposes have been relatively few and far between.

Although reliable figures are difficult to obtain, there can be little doubt that MT's share of the world translation market has until now been very modest, not to say minimal. And the reason for this is quite simple: the quality of the raw MT output has not generally been good enough to allow for its direct publication or cost-effective exploitation. Even when the MT system is coupled with a human post-editor who systematically revises, corrects and improves the computer's output, this arrangement will not prove cost-effective unless the machine's first draft achieves a certain level of quality. Otherwise, the post-editor finds herself spending more time correcting the machine's proposals than it would take to translate the text from scratch.

Canada's own *Météo* system, a system specifically designed for the translation of the weather forecasts issued daily by Environment Canada, is among the few exceptional cases where the raw machine output does attain close-to-publishable quality. The reason this system succeeds, however, is that it is designed specifically for weather bulletins, a very particular and narrow sublanguage. Language service providers (LSP's), for their part, are normally called upon to handle a much wider variety of texts, most of which are far more challenging than weather bulletins, and until recently there has been no compelling business case for them to adopt machine translation in order to help them produce the high-quality texts their clients require. And the same is true for most in-house translation services that are part of large corporations.

The advent of the new statistical paradigm in MT is in the process of changing all that. As mentioned above, SMT systems have been shown, objectively and

convincingly, to produce better-quality translations than their rule-based counterparts. Furthermore, the ease with which **specialized SMT systems** can be trained now makes it possible to develop systems that are tailored to the particular text type, style and terminology of any number of clients or client departments. And here too, it has been shown that these specialized systems can produce better-quality translations than such general-purpose systems as *Google Translate*.

Just as important as the advances in the technology, people's attitude and expectations with regard to MT have also undergone a significant change in recent years. Not very long ago, many who turned to machine translation expected that these computerized systems would allow them to dispense with costly, slow-moving human translators and obtain a high-quality translation, literally at the push of a button. After all, if computers helped us put a man on the moon and are now capable of defeating the world's best chess masters, why can't they automate the age-old process of language translation?[8]

Today, due in large part to the public's greater exposure to this technology, through easily accessible websites like *Google Translate* and *Bing Translator*, we no longer entertain such naïve views. Human translators are not about to go the way of hand weavers. That said, the improved quality of the translations produced by the new breed of statistical MT systems does seem to have reached a point where a **productive partnership** with human post-editors may finally be cost-effective. And this, in conjunction with the ever-increasing demand for translation and intense market pressures to reduce costs and turnaround time, is leading many LSP's and in-house translation services to take a fresh look at MT.

In the previous paragraph, we mentioned the public's increased exposure to machine translation and a consequent change in its attitude to the technology. Among translators too, there has been an important shift in attitudes. It used to be the case that many professional translators feared that MT would actually put them

[8] Why indeed? The question warrants a more fully developed answer than we have space for here.

out of a job. Far fewer translators believe that today, for they have seen firsthand what the technology can and cannot do, and they know that the demand for translation continues to spiral upward, far exceeding the collective capacity of the profession.

Instead, they are coming to view MT as a welcome assistant, which may well help them increase their productivity and, in certain cases, actually relieve them of the need to tackle the most repetitive and least interesting texts.[9] Translation service managers are also changing their attitude and expectations with regard to machine translation. In the past, many of those who were eager to test the technology did so in the hope that MT would save them money by replacing all the other tools and resources that their translators had previously been using. Not so today. Now, more and more translation managers understand that MT is best implemented as an additional component in the translator's varied arsenal, supplementing rather than replacing the other tools the translator has come to rely on and use productively.

Most popular among these are undoubtedly the translation memory (TM) systems that automatically detect and recover the translation of previously encountered sentences. At first, many translators resisted the introduction of these systems; now, the overwhelming majority use them on a day-to-day basis -- and would be very unhappy to relinquish them. But what about the many sentences that haven't been previously encountered, for which a TM system will have nothing to propose? An effective way of combining the two technologies in this situation is to have the MT system insert a machine translation in the TM editor, in what would otherwise be an empty target cell.[10]

Needless to say, the translator is free to accept, modify, or reject the machine's proposal, just as with the proposals that come from the translation memory. But if the MT system has been trained on an adequate corpus that corresponds to the

---

[9] Indeed, this was the original impetus that led to the development of the Météo system: translator ennui with boringly repetitive weather bulletins.
[10] Or a copy of the source segment in the target cell.

domain of the source text, chances are that the machine's proposal will contain at least some elements that the translator will be able to profitably recycle.

This way of combining MT and TM is all the more natural given that the core of the training material used to generate an SMT system is often drawn from the very same database of past translations as the translation memory. Another important point that TM's and SMT have in common is that both improve with use. As new, human-approved translations are added to the TM database, that system should be able to find matches for more and more sentences; and the SMT system will also have more data on which to calculate and improve its translation probabilities and coverage.

The impressive recent advances in SMT, together with this kind of productive partnership between different translation technologies are leading an increasing number of translation services and language service providers not just to consider machine translation, but to actually implement it. According to the January 2015 newsletter published by TAUS (the Translation Automation Users' Society), 40% of all translators are already making use of machine translation in one way or another.

Jost Zetzsche, in the 243rd issue of his highly respected Tool Box Journal, cites data from Memsource, a cloud-based TM system with more than thirty thousand registered users, suggesting that over 50% of their clients regularly consult the free online MT systems offered by *Google Translate* and *Microsoft Translator*. And in Canada, two of this country's largest LSP's are now using *Portage* on a day-to-day basis, because they have found that the system allows them to increase productivity, thereby saving time and money, without compromising quality or the private nature of their clients' texts.

A host of additional applications

So far, we have focused on two ways of using machine translation which could be considered to be at opposite poles of the application spectrum: on one hand, raw MT, used principally by the intelligence and business communities for information-gathering purposes; and, on the other, revised MT, used principally by translation services and LSP's for publication purposes. But between these two poles, there is a wide variety of other possible applications for MT, some of which we will briefly mention here.

The phenomenal growth of the social media, along with the steadily increasing popularity of e-commerce, has indirectly created an enormous new demand for translation. Millions of people every day post comments and user reviews on social media and business websites, which others, who don't necessarily understand the language of the post, nevertheless want to read. They can, of course, cut-and-paste that user-generated content into a free online MT system like *Google Translate.* But some companies have found that they can provide their users with better-quality translations by actively partnering with an MT developer in order to create a semi-customized system for their particular type of texts; this is what TripAdvisor has done with SDL. Other popular online sites, such as eBay, have invested in the development of their own proprietary MT technology.

Now in most such cases, the lifespan of the posted texts is very short, and so there is little point in having the machine translations carefully revised. Hence, what the users generally obtain is raw MT output; and most often, it is enough to allow them to make a decision as to whether they want to patronize a given hotel or restaurant, or purchase a given product. In the eBay scenario, moreover, the users may actually interact with the vendor in a conversation that is partially mediated by machine translation.

A slightly different application of unedited machine translation has been implemented at Microsoft, where the company is constantly updating enormous

online knowledge bases that support its wide array of products. Human translation in this context is clearly out of the question, and so Microsoft has applied its own MT system to the task. The content of the knowledge bases is drafted in English, but Microsoft clients can now query them in seven different languages (at last count), and hopefully find answers to their questions in their mother tongue. The company has also surveyed users on their satisfaction with the answers they obtained and, interestingly, there appears to be little difference in the satisfaction rate of those who query the knowledge bases in English and those who obtain their answers via machine translation.

Finally, between raw, unedited machine translation and fully edited MT output that aims for the same level of quality as human translation, various intermediate levels of post-editing are of course possible. One well-known compromise is commonly referred to as 'light post-editing'. What this means is that the reviser limits themselves only to those corrections that will interfere with understanding on the part of the reader. Hence, lightly post-edited MT output may contain minor grammatical errors, or stylistic infelicities – as long as they won't prevent the reader from grasping the meaning and all the essential content of the translated text. Lightly post-edited MT is often used for in-house communications or, in one well-known case, for internal memos and preliminary drafts of documents at the European Commission.

Needless to say, NRC's *Portage* could easily be adapted to any and all of the above scenarios.

Summary

Machine translation has progressed remarkably in recent years, particularly since the advent of the new statistical paradigm. Among SMT systems, NRC's *Portage* is among the very best in the world, as repeatedly demonstrated in various international competitions. What is more, the system is currently being used to cost-effective advantage by major language service providers in Canada, while being tested by businesses in other domains, as well as by the military and intelligence services in the United States.

The NRC is eager to extend the use of *Portage* to a broader range of partners as there are significant opportunities for the system to provide benefits in multiple areas. *Portage* is not only state of the art as far as the MT industry is concerned, but also less costly than most of its major competitors.

Organizations interested in finding out more about *Portage* –for practical demonstration or to inquire about licensing fees and collaboration opportunities – are invited to contact NRC to learn more:

Pierre Charron, Client Relationship Leader
Email: pierre.charron@nrc-cnrc.gc.ca
Tel.: +1 613 990-0336

NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Traduction automatique: avantages et retombées de la traduction automatique statistique et le logiciel Portage du CNRC

Conseil national de recherches du Canada. Technologies de l'information et des communications

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

https://doi.org/10.4224/21274927

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=91f8b1a9-5052-47e7-9c94-033efe904396
https://publications-cnrc.canada.ca/fra/voir/objet/?id=91f8b1a9-5052-47e7-9c94-033efe904396

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada Conseil national de recherches Canada

Canada

CNRC·NRC Technologies de l'information et des communications

Traduction automatique

Avantages et retombées de la traduction automatique statistique et le logiciel *Portage* du CNRC

avril 2015

Conseil national de recherches Canada
National Research Council Canada

Canada

Introduction

La traduction automatique (TA) est une façon de transposer au moyen d'un ordinateur un texte écrit dans une langue naturelle en un texte d'une seconde langue naturelle qui sera idéalement son équivalent sémantique et grammatical. Pour le commun des mortels, le système de TA se résume à un encadré noir à l'intérieur duquel on dactylographie le texte dans sa langue d'origine ou indique à l'ordinateur quel fichier il doit traiter. Quelques secondes plus tard, la machine en propose une version dans la langue souhaitée.

Toutefois, il suffit de creuser un peu plus pour constater que le système de traduction automatique peut, selon sa nature – par apprentissage statistique ou à base de règles – influer sur les résultats obtenus. Depuis 2003, une équipe de spécialistes de calibre international du Conseil national de recherches du Canada perfectionne un système breveté de traduction automatique statistique (TAS) baptisé *Portage*[1]. L'article que voici remet la TAS en contexte et compare cette technologie à l'autre grande méthode de traduction automatique en usage : la traduction automatique à base de règles (TABR). Durant la majeure partie de sa courte vie, la traduction automatique avait pour paradigme principal la TABR, dont la genèse peut être retracée à l'année 1948 quand, dans une note envoyée à quelques collègues, le mathématicien Warren Weaver suggéra que l'on pourrait utiliser l'ordinateur pour réaliser des traductions, un peu comme on y avait recouru durant la Deuxième Guerre mondiale pour déchiffrer des codes. Subséquemment, quand la linguistique chomskyenne s'imposa, vers le milieu des années 1960, la traduction automatique en vint à épouser des règles de formalisation déclarative analogues.

Durant cette période, les systèmes de TA étaient mis au point par des linguistes et des informaticiens qui s'efforçaient de programmer l'ordinateur afin qu'il imite le fonctionnement d'un traducteur tel qu'ils se l'imaginaient, c'est-à-dire analyser le

[1] L'État canadien s'intéresse activement à la traduction automatique depuis 1963. Au fil des ans, il a financé la recherche dans ce domaine dans divers laboratoires publics et universités. *Portage* est le produit le plus récent de cette longue démarche.

texte de départ pour s'en faire une quelconque représentation utile, puis transposer cette représentation abstraite en un texte équivalent de la langue d'arrivée. Pour y parvenir, ces chercheurs devaient produire d'impressionnants dictionnaires et précis grammaticaux pour l'ordinateur, tâche manuelle titanesque, aussi laborieuse qu'onéreuse. Les systèmes qui en émergèrent n'étaient pas particulièrement robustes : ils échouaient souvent à produire un résultat quelconque parce qu'il s'avère extrêmement difficile de rédiger une grammaire exhaustive et un dictionnaire complet pour quelque chose d'aussi complexe, en perpétuelle évolution, qu'est une langue naturelle. À ces difficultés, il faut encore ajouter celles tout aussi redoutables de compiler des dictionnaires bilingues et des grammaires contrastives capables d'établir un pont adéquat entre les deux langues.

En 1988, un groupe du T.J. Watson Research Centre d'IBM avança l'idée que la traduction automatique pourrait s'effectuer par apprentissage statistique. Quelques membres du groupe avaient participé antérieurement à la création de systèmes statistiques de reconnaissance de la parole – autre problème complexe tenant depuis longtemps en échec l'approche basée sur des règles.

Contrairement à la TABR, la TAS ne repose pas sur des règles grammaticales de formulation ni sur les entrées d'un dictionnaire. Au lieu de se fier aux intuitions du linguiste concernant un langage, ces systèmes puisent leurs connaissances linguistiques directement dans un vaste corpus de traductions. L'application d'algorithmes d'apprentissage machine aux textes permet d'estimer automatiquement la probabilité qu'un mot ou une phrase *X* dans la langue d'origine se traduise par le mot ou la phrase *Y* (ou W, ou Z…) dans la langue d'arrivée. Le système de TAS comporte deux grands éléments linguistiques, tous deux de nature probabiliste : un module traductionnel qui correspond grosso modo au dictionnaire bilingue du système TABR et un module linguistique qui garantit la fluidité du texte résultant en maximisant la probabilité de chaque terme W de la traduction, selon les choix retenus pour les *n* termes précédents.

De nombreux autres modules statistiques et un algorithme de base appelé « décodeur » effectuent les recherches afin de restituer le meilleur texte possible dans la langue cible, par application de ces ensembles probabilistes de connaissances linguistiques au texte d'origine. Pour revenir à la définition initiale donnée plus haut, on pourrait dire en quelque sorte que le module traductionnel veille à la fidélité sémantique du texte final, et le module linguistique, à sa rectitude grammaticale.

Avantages de la TAS

Comparativement à la TABR, la TAS présente plusieurs avantages majeurs, le premier étant sa **polyvalence**. En effet, avec la traduction automatique basée sur des règles, chaque fois que l'on souhaite ajouter une nouvelle langue *T*, il est nécessaire de mobiliser des linguistes informaticiens spécialisés dans cette langue et des lexicologues bilingues qui connaissent bien la paire *S* > *T*, pour autant qu'on puisse en trouver.

La TAS, en revanche, applique les mêmes algorithmes d'apprentissage machine à n'importe quelle paire de langues, ou presque. Une personne possédant un système de TAS anglais > français qui décide de s'attaquer à une traduction de l'anglais à l'espagnol (ou du portugais au bulgare, et ainsi de suite) peut le faire sur-le-champ, à condition – évidemment – qu'elle dispose du matériel voulu pour réaliser l'apprentissage machine. Par là, on entend simplement un nombre suffisant de documents correctement traduits entre les deux langues concernées[2]. Par conséquent, la TAS diminue le temps, les efforts et l'argent investis dans le développement de nouveaux systèmes de traduction automatique de plusieurs ordres de grandeur. Ce qui exigerait des mois et de nombreuses centaines

[2] Combien de textes bilingues faut-il pour former un système de TAS? Les spécialistes ont tendance à dire que, plus le corpus est grand, meilleurs seront les résultats. Néanmoins, un corpus de taille minimale s'avère nécessaire au départ pour que les probabilités qui en émanent soient fiables. On cite souvent comme chiffre une centaine de milliers de phrases dans les deux langues, soit approximativement un million de mots.

d'heures-personnes de travail avec la traduction automatique à base de règles peut désormais être réalisé (littéralement) du jour au lendemain.

Un autre avantage appréciable de la TAS, par rapport à la TAR, est sa **robustesse**. Avec la TAS, chaque phrase dans la langue d'origine peut être restituée de multiples façons dans la langue d'arrivée, bien que certaines propositions soient habituellement beaucoup plus probables que d'autres. En conséquence, les systèmes de TAS produisent toujours un résultat. Sur ce plan, ils sont donc naturellement plus robustes que les systèmes de TAR plus anciens qui, souvent, dépendent d'une analyse grammaticale complète de la phrase d'origine.

Évidemment, cette robustesse présente peu d'intérêt si la qualité des traductions obtenues laisse trop à désirer. Or, on a prouvé, et de manière concluante, dans les concours internationaux ouverts comme ceux organisés par l'armée des États-Unis et le National Institute of Standards and Technology (NIST) que les systèmes de TAS ont tendance à engendrer des traductions de **qualité supérieure** à celles obtenues avec leurs contreparties TAR. Les meilleurs systèmes de TA actuels s'appuient tous sur l'apprentissage statistique, et la TAS est à la base même du **système *Portage* du CNRC.**

Aujourd'hui, les systèmes statistiques dominent totalement le panorama de la traduction automatique. Quelques-uns parmi les plus connus, comme *Google Translate* et *Microsoft Translator*, sont disponibles gratuitement sur Internet, ou peuvent être exploités pour une somme très modique. D'autres, de source ouverte (la plupart reposent sur *Moses*[3]), sont offerts par des distributeurs qui les adaptent aux besoins de leurs clients. Face à ces multiples possibilités, celui qui songe à la traduction automatique serait en droit de se demander « Pourquoi me donner la

[3] *Moses* est un système de TAS complet, créé initialement sous la direction de Philipp Koehn. Le système devait constituer l'équivalent universitaire des systèmes brevetés par les grandes sociétés comme Google et Microsoft. Soutenu financièrement par la Commission européenne, *Moses* a récemment été élargi pour accepter de nombreuses paires de langues en usage en Europe.

peine d'adopter le système *Portage*? », qui n'est pas un gratuiciel. Pourtant les raisons ne manquent pas.

Portage est l'un des meilleurs systèmes de TA par apprentissage statistique actuellement disponibles dans le monde. Les résultats obtenus par ce système lors des récents concours du NIST l'ont prouvé de manière éclatante. En effet, *Portage* a coiffé tous les autres systèmes au poteau pour les traductions du chinois à l'anglais, et s'est classé bon deuxième pour celles de l'arabe à l'anglais – un véritable exploit! L'équipe *Portage* du CNRC a été invitée à deux reprises à participer à d'ambitieux et prestigieux projets financés par la DARPA, la Defense Advanced Research Projects Agency, du gouvernement des États-Unis, de concert avec des laboratoires renommés tels Raytheon BBN et RWTH (Allemagne). Parallèlement, le groupe a collaboré activement avec d'importants partenaires commerciaux comme *Systran*, entreprise qui brille au sommet, sur le marché de la traduction automatique, depuis des décennies.

Beaucoup d'utilisateurs éventuels de la traduction automatique redoutent – et la crainte est légitime – que la sécurité de leurs documents soit compromise, et hésitent soit à expédier ceux-ci hors du pays pour les faire traduire, soit à les télécharger dans le nuage par Internet. Plus rares encore sont ceux qui accepteraient de voir leurs documents employés pour perfectionner le moteur TA d'une autre entreprise, sur lequel ils n'exerceraient aucun droit[4]. *Portage* offre une solution simple à ces préoccupations. Normalement, le système est installé sur un serveur local, dans les locaux sécurisés de l'utilisateur. Personne d'autre n'y a accès, pas plus qu'aux documents traités. Comme nous le verrons ci-dessous, l'utilisateur pourra facilement intégrer le logiciel à sa plateforme de traitement des documents existante.

---

[4] C'est notamment le cas de *Microsoft Translator*, bien que Microsoft insiste sur le fait que les textes téléchargés sur ses serveurs de traduction ne puissent être récupérés ni reconstitués par qui que ce soit.

Possibilités commerciales associées à *Portage*

De ses origines à la fin de la Guerre froide, la traduction automatique était surtout employée par l'armée et les services de renseignement qui finançaient aussi principalement la recherche et le développement dans ce domaine. En effet, la masse de documents traitée par ces services dépassait, et de loin, la capacité des traducteurs disponibles. Le seul espoir de remédier à pareille situation consistait à automatiser la traduction.

Des considérations d'ordre géopolitique ont par ailleurs fixé les principales langues étrangères pour lesquelles ont été développés les systèmes de TA. En d'autres termes, ce n'est pas un hasard si la première illustration de traduction automatique donnée à New York, en 1954, portait sur une conversion du russe à l'anglais. Les sommes faramineuses injectées par le gouvernement américain dans l'élaboration d'un système de TA du vietnamien à l'anglais durant les années 1960, en pleine guerre au Vietnam, puis du farsi à l'anglais, dans les années qui ont précédé et suivi la destitution du shah d'Iran ne sont pas non plus accidentelles[5]. Dans ces cas de figure, la traduction automatique devait répondre à un besoin pressant au niveau de la sécurité nationale.

L'armée et les services de renseignement continuent d'exploiter largement la traduction automatique aux États-Unis. Depuis quelques années, leur dépendance envers cette technologie s'est même accrue de façon draconienne, car en plus de ce qui s'écrit, on surveille maintenant ce qui se dit, ce qui se filme et ce qui se bavarde sur les médias sociaux. Le moins que l'on puisse dire est que les attentats du 11 septembre 2001 ont avivé l'intérêt du département de la Défense américaine pour la traduction automatique, l'incitant à injecter de lourdes sommes dans la R-D pertinente et à mettre sur pied les concours internationaux mentionnés précédemment.

---

[5] Logos Corporation a été la principale entreprise à profiter des sommes investies par les États-Unis dans du vietnamien et du farsi.

Pour des motifs analogues, l'arabe est subitement devenu le centre d'attraction, suivi de près par le chinois. Les organismes de financement ont insisté sur la rapidité avec laquelle les systèmes devaient absolument être développés, à partir d'une documentation moins volumineuse et plus variée. On peut donc avancer sans grand risque de se tromper qu'une bonne partie des progrès remarquables réalisés récemment dans le domaine découlent des fonds engloutis par l'armée et les services de renseignement dans cette technologie.

Il importe de souligner que, dans un tel contexte, la traduction automatique sert principalement à **glaner de l'information**. Même si les traductions obtenues s'avéraient parfaites (ce qui n'est certes pas le cas), aucun service de renseignement n'aurait le temps de parcourir une documentation aussi volumineuse. La traduction automatique n'est alors qu'un maillon dans une plus vaste chaîne ayant pour but de **passer au crible** l'information, bref de trier une quantité colossale de documents en langue étrangère afin de dénicher ceux qui présentent le plus d'intérêt, pour ensuite faire traduire ceux-ci par des humains. Dans une telle optique, les systèmes de TA existants fonctionnent à merveille.

Fait digne de mention, l'armée et les services de renseignement ne sont plus les seuls à s'intéresser à la collecte des informations présentes dans les documents en langue étrangère. Consécutivement à la mondialisation du commerce, les entreprises souhaitent de plus en plus savoir ce que les gens pensent de leurs produits et de leurs services sur les marchés de la planète. Depuis quelques années, une nouvelle industrie du **renseignement commercial** a donc vu le jour. Cette industrie vit de la collecte et de l'analyse de masses de données volumineuses, principalement récoltées dans les médias sociaux, dont l'importance ne cesse de croître, et qui aident, soutient-on, les entreprises à adapter et à peaufiner leurs stratégies commerciales.

Il y a quelques années, les grandes multinationales étaient pratiquement les seules à faire affaire sur les marchés qu'on dit « étrangers ». À présent, de plus en plus de PME s'y aventurent à leur tour, et elles aussi désirent savoir rapidement ce qu'on

pense d'elles dans d'autres langues. Les sociétés spécialisées dans le renseignement commercial sont contraintes de se tourner vers la traduction automatique pour satisfaire la demande relative au traitement et à l'analyse de textes multilingues sur une grande échelle. De fait, l'équipe *Portage* du CNRC a réalisé quelques travaux exploratoires sur l'analyse translinguistique des sentiments avec l'entreprise canadienne MediaMiser[6].

Application de la traduction automatique à la diffusion

Tel qu'il a été indiqué précédemment, lorsqu'ils auparavant l'information, les services de renseignement militaires et commerciaux se servent essentiellement de la traduction automatique aux fins d'**assimilation**, soit pour aider les gens à lire et à saisir en partie le contenu d'un texte rédigé dans une langue étrangère. Mais qu'en est-il de l'utilisation de la traduction automatique dans un but de **diffusion**, c'est-à-dire se servir du système pour produire et éventuellement publier un document dans une autre langue? Après tout, cela n'est-il pas le but premier et ultime de la traduction automatique?

Depuis sa naissance, on a à maintes reprises tenté d'employer la traduction automatique à cette fin, au public comme au privé. Aux États-Unis, des entreprises comme Systran, Weidner et Logos (pour ne mentionner qu'elles) ont expressément été fondées dans ce but durant la seconde moitié du siècle dernier. Outre Atlantique, où les multiples langues officielles de l'Union européenne ont engendré une demande énorme du côté de la traduction, la Commission européenne a investi considérablement d'argent et d'efforts dans l'adaptation d'un système commercial de traduction automatique susceptible de l'aider à subvenir à ses besoins avant de se lancer dans un ambitieux programme de développement bien à elle[7].

---

[6] Par « analyse des sentiments », on entend la catégorisation des opinions exprimées dans un document au moyen d'un ordinateur.
[7] Le programme, baptisé Eurotra, a fonctionné de 1978 à 1992. La CE finance toujours de nombreux projets de R-D en traduction automatique et sur des technologies connexes.

Au Canada, le Bureau de la traduction a abondamment testé plusieurs systèmes de TA commerciaux pour déterminer si cette technologie l'aiderait à mieux composer avec l'escalade de la demande pour ses services, tout en réduisant parallèlement ses coûts. La plupart de ces tentatives n'ont rien donné. Jusqu'à tout récemment, la traduction automatique n'avait servi que très peu et de loin en loin à la diffusion.

Bien qu'il soit difficile d'obtenir des chiffres fiables, il est indéniable que la part du marché mondial de la traduction détenue par la traduction automatique était jusqu'à ce jour très modeste, pour ne pas dire infime. La raison pour cela est fort simple : de manière générale, la qualité des textes issus de la traduction automatique n'est pas assez bonne pour qu'on en autorise la publication ou pour qu'on les exploite de façon rentable. Même lorsqu'on combine la traduction automatique à une personne qui révisera, corrigera et améliorera méthodiquement le travail après coup, la rentabilité n'est pas au rendez-vous, à moins que le premier jet produit par la machine n'atteigne un certain degré de qualité. Sinon, corriger ce que propose l'ordinateur exige plus de temps qu'il en faudrait pour traduire le document d'emblée.

Le système *Météo* du Canada, spécifiquement conçu pour traduire les bulletins météorologiques quotidiens d'Environnement Canada, est l'un des rares cas où la traduction brute réalisée par ordinateur parvient à une qualité suffisante pour être diffusée. Néanmoins, le succès du système est attribuable au fait qu'il a précisément été créé pour les bulletins météo, qui emploient une terminologie très précise et particulière. Or, les langagiers doivent le plus souvent composer avec des textes fort variés, beaucoup plus ardus à traduire que des bulletins météorologiques pour la plupart, et rien, jusqu'à tout récemment, ne les incitait à lorgner du côté de la traduction automatique pour les aider à produire les documents de qualité réclamés par leurs clients. Cette vérité s'applique aussi à la majorité des services de traduction internes des grandes sociétés.

L'avènement du nouveau paradigme statistique en traduction automatique est sur le point de changer la donne. Comme on a pu le lire plus haut, les systèmes de TAS ont prouvé de façon objective et convaincante qu'ils sont capables de fournir de

meilleures traductions que celles des systèmes de TABR. Par ailleurs, la facilité avec laquelle on produit les **systèmes de TAS spécialisés** autorise désormais le développement de systèmes adaptés à des textes d'une nature, d'une terminologie et d'un style particuliers, peu importe le nombre de clients ou de ministères desservis. Là aussi, on a fait la preuve que ces systèmes spécialisés produisent des traductions de meilleure qualité que les systèmes généralistes du genre *Google Translate*.

L'attitude et les attentes de la population à l'endroit de la traduction automatique ont sensiblement changé eux aussi, ces dernières années, aspect aussi important que les progrès réalisés par la technologie eux-mêmes. Il y a peu encore, beaucoup de ceux qui se tournaient vers la traduction automatique étaient persuadés que l'ordinateur les aiderait à se départir de traducteurs lents et onéreux tout en récoltant un produit de qualité, simplement au déclic d'un bouton. Après tout, si l'ordinateur a permis à l'homme de poser le pied sur la lune et réussit à battre les plus grands maîtres aux échecs, pourquoi ne pourrait-il automatiser cette démarche antique qu'est le passage d'une langue à une autre[8]? Nous ne sommes plus si naïfs aujourd'hui, et on le doit en grande partie à l'expérience que les sites Web, aisément accessibles, comme *Google Translate* et *Bing Translator* ont fait vivre au public.

Non, le traducteur ne disparaîtra pas aussi vite que le tisserand d'autrefois. Cela dit, la qualité des textes obtenus avec la nouvelle génération de systèmes de traduction automatique par apprentissage statistique semble avoir atteint le point où l'on peut désormais envisager un **partenariat productif** entre la machine et un correcteur humain. Cette évolution, et la demande qui ne cesse de grandir, de même que les fortes pressions exercées par le marché en vue d'une réduction des coûts et d'un raccourcissement des délais, incitent de nombreux langagiers et services de traduction internes à voir la traduction automatique d'un œil neuf.

---

[8] Oui, effectivement, pourquoi pas? La question mériterait une réponse plus élaborée que l'autorise malheureusement l'espace dont nous disposons.

Le paragraphe précédent mentionnait une exposition de plus en plus grande du public à la traduction automatique, avec le changement d'attitude que cela implique. Pareil changement s'observe également chez les traducteurs. Auparavant, beaucoup de membres de cette profession craignaient de voir disparaître leur gagne-pain à cause de la traduction automatique. Ils sont nettement moins nombreux aujourd'hui, car ils ont vu de leurs propres yeux ce que cette technologie peut et ne peut réaliser, et ils savent que la demande, en poursuivant son escalade, dépasse considérablement la capacité collective de leur profession.

À présent, les traducteurs voient plutôt dans la traduction automatique un instrument inespéré qui leur permettrait d'accroître leur rendement et, dans certains cas, les soulagerait de la fastidieuse corvée que représentent les textes les plus répétitifs et les moins intéressants[9]. Ceux qui gèrent des services de traduction ne regardent plus davantage la traduction automatique du même œil. Dans le passé, beaucoup s'étaient empressés de tester la technologie dans l'espoir qu'elle amènerait des économies en remplaçant les autres outils et ressources employés jusque-là par les traducteurs. Plus maintenant. Un nombre croissant de ces gestionnaires ont compris que la traduction automatique ajoute une corde à l'arc du traducteur, qu'elle augmente son arsenal au lieu de se substituer aux outils sur lesquels il a appris à compter et dont il se sert efficacement.

Parmi les technologies les plus populaires figure sans aucun doute la mémoire de traduction, système qui détecte et récupère automatiquement les phrases déjà traduites. Si beaucoup de traducteurs se sont d'abord opposés à l'introduction de tels systèmes, la vaste majorité y recourt maintenant tous les jours et ne s'en priverait plus que de mauvaise grâce. Mais qu'en est-il des nombreuses phrases jamais encore rencontrées, pour lesquelles la mémoire de traduction n'a rien à proposer? Une façon efficace de combiner les deux technologies en pareil cas consiste, pour le système de TA, à incorporer une traduction automatique à l'éditeur

---

[9] À dire vrai, il faut y voir le motif à l'origine du système Météo : l'ennui qui assommait ceux contraints de traduire des bulletins répétitifs à mourir.

de la mémoire de traduction, dans une cellule cible qui sans cela resterait vide[10]. Inutile d'ajouter que le traducteur peut, à sa guise, accepter, modifier ou rejeter ce que l'ordinateur lui propose, comme il le fait déjà avec les propositions soumises par la mémoire de traduction. Si le système de TA est basé sur un corpus adéquat, correspondant au domaine du texte d'origine, il y a alors de fortes chances que la version suggérée renferme au moins quelques éléments que le traducteur réutilisera à son avantage.

L'amalgame de la traduction automatique et de la mémoire de traduction est d'autant plus naturel que les données d'apprentissage idéales pour « entraîner » le système de TAS sont précisément celles contenues dans la mémoire de traduction. Un autre point important que la mémoire de traduction et la TAS ont en commun est que ces deux systèmes s'améliorent avec le temps. En effet, avec l'addition de nouvelles traductions sanctionnées par l'être humain à la base de données, le système de mémoire de traduction devrait trouver de plus en plus de phrases qui se recoupent; et en même temps, le système de TAS disposera de données plus nombreuses pour mieux estimer les probabilités de traduction et mieux couvrir le domaine pertinent.

Les progrès récents – impressionnants – réalisés en TAS et une association aussi productive des technologies traductionnelles expliquent pourquoi un nombre croissant de services de traduction et de langagiers ne font pas qu'envisager, mais adoptent la traduction automatique. Selon le bulletin de janvier 2015 de la TAUS (Translation Automation Users' Society), quarante pour cent des traducteurs recourent déjà à la traduction automatique d'une manière ou d'une autre. Dans le 243e numéro du très respecté *Tool Box Journal*, Jost Zetzsche cite les données de Memsource, un système de TA en nuage auquel se sont inscrits au-delà de trente mille utilisateurs. Ces données suggèrent que plus de la moitié des clients de Memsource consultent régulièrement les systèmes de TA en ligne gratuits offerts par *Google Translate* et *Microsoft Translator*. Au Canada, deux des plus grands

[10] Ou la copie du segment d'origine dans la cellule cible.

cabinets de traduction du pays recourent désormais quotidiennement à *Portage* après avoir constaté que le système augmentait leur productivité, bref leur épargnait temps et argent sans que la qualité ou la confidentialité des textes de leurs clients en souffrent pour autant.

Une pléthore d'applications

Jusqu'à présent, nous n'avons parlé que de deux usages possibles, et diamétralement opposés, de la traduction automatique : à une extrémité du spectre, la traduction automatique brute, employée essentiellement par les services de renseignement militaires et commerciaux pour glaner de l'information, et, à l'autre bout, la traduction automatique révisée, surtout utilisée par les services de traduction et les langagiers aux fins de diffusion. Entre ces deux pôles cependant se situe une vaste gamme d'applications que nous nous bornerons ici à mentionner.

L'essor phénoménal des médias sociaux et l'engouement pour le cybercommerce ont indirectement engendré une nouvelle demande, colossale, en traduction. En effet, chaque jour, des millions de gens publient commentaires et analyses sur les médias sociaux ou le site Web des entreprises, dans une langue peut-être incompréhensible à ceux qui aimeraient en prendre connaissance. Bien sûr, ces derniers peuvent copier-coller le contenu affiché par les utilisateurs dans un système de TA en ligne gratuit tel *Google Translate*, mais des sociétés ont constaté qu'elles pourraient offrir à leurs visiteurs une traduction plus digne de foi en s'associant à un développeur pour créer un système à moitié spécialisé, adapté à la nature particulière de leurs documents. TripAdvisor a choisi cette solution avec SDL. D'autres sites en ligne populaires, tel eBay, ont pour leur part investi dans le développement d'une technologie de traduction automatique qui leur est propre. Actuellement, les textes affichés n'ont le plus souvent qu'une vie très brève, et il y a peu d'utilité à réviser avec minutie les textes restitués mécaniquement. L'utilisateur n'obtient donc généralement qu'une traduction automatique brute. Celle-ci lui suffit habituellement pour décider s'il fréquentera ou pas tel hôtel ou tel restaurant, ou achètera ou pas un produit quelconque. Avec la solution retenue par eBay

néanmoins, l'utilisateur qui le désire interagit avec le fournisseur en entamant une conversation nourrie en partie par la traduction automatique.

Microsoft a instauré une application légèrement différente de traduction automatique sans vérification en vertu de laquelle l'entreprise actualise en permanence les énormes bases de connaissances en ligne qui appuient la multitude de produits offerts. Recourir à des traducteurs pour cela serait manifestement impensable. C'est pourquoi Microsoft a mis son propre système de TA à la tâche. Bien que le contenu des bases de connaissances soit initialement rédigé en anglais, la clientèle de Microsoft peut le consulter dans sept langues (au dernier recensement) et espérer trouver la réponse à ses questions dans sa langue maternelle. L'entreprise a sondé ses clients pour connaître leur degré de satisfaction. Fait intéressant, ceux qui effectuent une recherche dans la base de connaissances en anglais semblent aussi satisfaits que ceux dont les réponses ont été traduites mécaniquement.

Enfin, on peut évidemment atteindre des niveaux de révision qui se situent entre, d'une part, la traduction automatique brute, sans correction, et, d'autre part, la traduction automatique révisée visant un degré de qualité similaire à celui d'une traduction humaine. Compromis courant, la « postédition légère » suppose que le réviseur corrige uniquement ce qui empêcherait l'utilisateur de saisir le sens du texte. Les traductions de ce genre seront donc émaillées de fautes de grammaire mineures et d'impropriétés stylistiques pourvu que le lecteur comprenne l'essentiel de ce qu'il lit. On recourt souvent à la traduction automatique avec postédition légère pour les communications internes comme le fait, pour ne citer que ce cas, la Commission européenne avec ses notes de service internes et l'ébauche de ses documents.

Inutile de dire que le logiciel *Portage* du CNRC s'adapte aisément à chacune des situations préalablement décrites.

Sommaire

La traduction automatique a connu des progrès remarquables ces dernières années, surtout depuis l'avènement du nouveau paradigme statistique. Le logiciel *Portage* du CNRC figure parmi les meilleurs systèmes de TAS au monde, ainsi qu'on a pu le constater à maintes reprises dans divers concours internationaux. Des importants langagiers du Canada y recourent déjà pour accroître leur rentabilité, alors que des entreprises d'autres secteurs en font présentement l'essai, au même titre que l'armée et les services de renseignement des États-Unis.

Le CNRC souhaite ardemment étendre l'utilisation de *Portage* à un plus grand nombre de partenaires, et le système pourrait engendrer des avantages appréciables dans plusieurs secteurs. *Portage* n'est pas que le ne plus ultra de la traduction automatique, c'est aussi un système moins coûteux que la plupart de ceux offerts par les principaux concurrents du CNRC.

Les gens qui aimeraient en apprendre davantage sur *Portage* – assister à une démonstration pratique ou en savoir plus sur les droits associés aux licences d'exploitation et les possibilités de collaboration, par exemple – sont invitées à communiquer avec le CNRC.

Pierre Charron, chef, Relations avec les clients
Courriel : pierre.charron@nrc-cnrc.gc.ca
Tél. : +1 613-990-0336

A BLEU Comparison of Portage *vs.* Google Translate on Four Domains

Roland Kuhn, Darlene Stewart, & Rabib Islam – Oct. 28, 2015.

Introduction

Commercial clients of the NRC's Portage machine translation software typically train a version of Portage on one or more translation memories, each memory representing a particular domain (*e.g.*, a particular client or topic or type of document). Each trained version of the system is then deployed to generate initial translations for new documents in that domain, which are then post-edited by human translators.

By contrast, some translators use an online machine translation system such as Google Translate (GT) or Microsoft's Bing to generate initial translations. This practice may pose a threat, real or imagined, to the security of the documents being translated. Organizations that license Portage from NRC have frequently stated that one of the reasons they prefer to use Portage rather than an online system is because when one trains and deploys versions of Portage inside an organization, no-one outside the organization (not even employees of NRC) will ever see either the documents on which these Portage versions are trained, or the new documents they will translate.

This document will not discuss the security issues which make Portage preferable, in some situations, to online machine translation systems. It is concerned only with machine translation quality. It describes how we measured the performance of Portage against the most widely used online machine translation system, GT, on four different domains, to explore the question of when Portage yields better translations and when GT does.

Each of the two systems has its own areas of strength:

> When a version of Portage trained on previously translated sentences from a specific domain is used to translate new documents from the same domain, it has the advantage of **specificity**. It will have memorized the specific ways in which words and phrases from this domain are typically translated.
> The GT system for a specific language pair – e.g., English to French (denoted here by "English > French") – is trained on a huge amount of bilingual data from all over the Internet, several orders of magnitude more than a typical translation memory. It therefore has much broader coverage than any version of Portage any of NRC's clients are ever likely to train. GT has the advantage of **generality**.

It is difficult to assess which system has the advantage in terms of **algorithmic quality**. GT often enters international evaluations of machine translation quality, and always performs extremely well, but it never participates in the "constrained" tracks of such evaluations, where the training data is pre-specified (limited to a much smaller amount of data than a GT system for a given language pair is trained on). It is thus impossible to say how much of GT's good performance is due to good algorithms, and how much to enormous quantities of training data. However, Google employs a large number – perhaps a majority – of the most respected machine translation experts on the planet, so its algorithms are likely to be very good. Portage often ranks first or close to first in "constrained" tracks of international algorithms, so its algorithms are also good.

The conclusion from our experiments is that when Portage is used as recommended by NRC – to translate texts from the same domain as the given version of Portage has been trained on – it will often yield translations that are of dramatically better quality than those that will be obtained from Google Translate. On the other hand, Google Translate will yield better translations for out-of-domain or mainly out-of-domain texts.

Experiments

Roughly 80% of the translations performed by Portage systems installed by NRC's clients are English > French, so the experiments we carried out were on this language pair. The easiest way of measuring quality differences is BLEU. This metric inputs a set of sentences in the source language and high-quality translations for them produced by human translators; it measures the similarity between the outputs produced by a machine translation system and the corresponding human translation. The higher the BLEU score on a given set of bilingual data, the better the quality of the machine translation is likely to be. BLEU doesn't correlate perfectly with human judgments of translation quality, but it correlates quite well. The rule of thumb is that BLEU differences of about 1.0 and above are perceptible to people. That is, if system A yields outputs on a test set that have a BLEU score of 35.0, and system B yields outputs on the same test set that have a BLEU score of 35.3, human translators might not see a quality difference – they might even prefer the translations from system A. But if system A yields outputs with 35.0 BLEU, and system B yields outputs with 36.0 BLEU, it is extremely likely that human experts will judge system B to be producing higher-quality outputs.

The experiments explore the "specificity" *vs.* "generality" question by considering four different scenarios involving four different domains. In all experiments, the Portage system is in the "recommended" configuration for the next release. In addition to the bilingual training data (taken from a specific Government of Canada department web site for each of the three in-domain scenarios, and from Version 10 of the WMT News Commentary for the NC out-of-domain scenario), this configuration always contains a background "gc.ca" language model and a neural net model (NNJM), both trained on many different Government of Canada web sites. In all four scenarios, the version of Portage that's deployed is trained on a realistic amount of data (i.e., a number of sentence pairs that could plausibly be found in a translation memory).

Three of the scenarios are in-domain: the training data and the test sentences were drawn from the same domain (efforts were made to ensure that the test sentences didn't occur in the training data). The three in-domain scenarios varied in both the amount and the source of the training data: there was a "small" (EC), "medium" (PCH), and "large" (HC) in-domain scenario.

The fourth scenario (NC) is different from the others, because it is to some extent an out-of-domain scenario. Because of the background "gc.ca" language model and the NNJM, both trained on Government of Canada web data, all four versions of Portage tested here have a Government of Canada bias, and probably a bias towards Canadian over international French. In the NC scenario, the test data come from francophone, non-government sites outside Canada. There is only a small amount of in-domain training data, which is probably not sufficient to cover the new vocabulary and phraseology in the test set.

Before we performed the experiments, we believed that GT would outperform Portage in the NC scenario – given GT's vastly greater amount of training data, any other result would be surprising. We were not sure which system would perform better in the in-domain scenarios.

The GT experiments were performed in mid-October 2015. Note that GT is a moving target: the GT outputs would probably be different if they were obtained six months earlier or six months later.

Scenario	**Data source**	**#training sentence pairs**	**#training Eng. tokens**	**#training Fre. tokens**
In-domain, "small" (EC)	Environment Canada (EC)	251,598	3,878,355	4,753,869
In-domain, "medium" (PCH)	Patrimoine Canada Heritage (PCH)	566,027	10,043,937	12,028,577
In-domain, "large" (HC)	Health Canada (HC)	1,194,779	20,851,583	25,692,050
Partly out-of-domain, "small" (NC)	News Commentary v10 **(Test = newstest)**	199,183	5,063,260	5,997,583

Scenarios & training data

Test set	**#sent. pairs**	**#Eng. tokens**	**#Fre. tokens**
EC test1	1,907	28,452	34,698
EC test2	1,934	30,584	37,428
PCH test1	1,987	36,792	43,751
PCH test2	1,976	37,204	44,275
HC test1	1,976	34,888	43,371
HC test2	1,982	35,614	43,801
NC test1 (2011)	3,003	74,563	84,617
NC test2 (2013)	3,000	64,527	73,708

Test data

Results and Discussion

Results are shown in the table below. As expected, Google Translate (GT) outperformed Portage on the NC test (by 4.3 BLEU points). Surprisingly, Portage outperformed GT by an even larger margin of 5.4 – 6.8 BLEU points on the three in-domain tests: the "small" EC, "medium" PCH, and "large" HC tests.

This dramatic advantage for Portage over GT on the in-domain tests happens even though it is quite possible that GT is "cheating". The EC, PCH, and HC test sets are all drawn from Government of Canada web sites that Google has access to for training, so GT may have been trained on these test sets (the Portage results are "honest" ones because we tried to avoid choosing test sentences that were in the training data for Portage). It turns out not to matter that GT was possibly trained on test data - if that's the case, the specific examples GT was both trained and tested on were drowned out by all the billions/hundreds of millions of other English-French examples GT is trained on.

Scenario	System	test1	test2	Average	Delta
EC	Portage	38.81	36.94	37.88	+5.35
	Google	32.98	32.07	32.53	
PCH	Portage	41.03	42.18	41.61	+5.79
	Google	35.36	36.28	35.82	
HC	Portage	38.84	39.70	39.27	+6.84
	Google	32.50	32.37	32.44	
NC	Portage	28.92	27.46	28.19	-4.30
	Google	34.35	30.63	32.49	

BLEU comparison: Portage *vs.* Google Translate

One of us (R. Kuhn) has carried out a quick visual comparison of the Portage and GT outputs for the three in-domain scenarios. Generally, both systems perform well. However, certain examples help to suggest hypotheses for why Portage obtains a higher BLEU score than GT for these scenarios. Because these hypotheses are too speculative to be real science (yet), they have been relegated to the "Appendix" below.

The numbers above pertain to the next release of Portage, which none of NRC's clients have access to yet. What about the previous Portage release, the one our clients are currently using – is it also superior to the current version of Google Translate when tested on in-domain data?

We haven't performed experiments with the previous Portage release on the test sets used above, but we can answer the question. In similar tests, the previous Portage release scored about 1.4 BLEU lower than the next Portage release for the EC scenario, 1.8 BLEU lower for the PCH scenario, and 1.5 BLEU lower for the HC scenario. If we do the appropriate subtractions, we see that our clients are currently deploying a version of Portage that would score 4.0 BLEU points (5.4-1.4) higher in the EC scenario than GT would, 4.0 BLEU points (5.8 – 1.8) higher in

the PCH scenario than GT would, and 5.3 BLEU points (6.8 – 1.5) higher in the HC scenario than GT would.

Thus, when Portage is trained on translations from a given domain and then deployed to generate new translations for texts from that domain – the usage NRC recommends to its clients – it can yield not just slightly, but dramatically better translations than Google Translate. That is true not only for the next Portage release, but also for the one our clients are currently using.

Future Work

We are planning to carry out a human evaluation to see if humans have as a strong a preference for Portage over GT in in-domain scenarios as BLEU does. We may also do BLEU and human comparisons of Portage and GT in the opposite language direction, French > English.

Appendix: Examples & Hypotheses *(by R. Kuhn)*

I have looked at outputs for the three in-domain scenarios – EC, PCH, and HC – to understand why Portage has a much higher BLEU score than Google Translate (GT) for these scenarios. At first glance, this is a little surprising: by and large, people familiar with machine translation would say that both systems are performing relatively well, and might think they should score about the same. However, I believe I have identified some of the differences between the two systems that cause Portage to score higher. It seems likely that these systematic differences correlate with productivity – i.e., that a post-editor working with Portage outputs for these scenarios would expend less effort in producing a good translation than if he or she was post-editing Google Translate.

A type of difference between the two systems that was **not** surprising was cases where both Google and Portage translated an English expression into perfectly correct French, but the French expression chosen by Portage was more typical of Government of Canada documents (the EC, PCH, and HC training & test data are taken from Government of Canada websites). GT for English > French is trained on bilingual documents from many different countries and organizations, so it is entirely natural that from time to time "Government of Canada" usage is "drowned out" by translation preferences from elsewhere. Here is an example from **HC test2** ("SRC" means the source English sentence, "REF" means the reference translation provided by a human translator):

```
SRC: Please include notes for each budget line item to explain how the
total amount requested was calculated.
REF: Chaque ligne budgétaire doit comporter des annotations sur le
calcul menant au total.
GOOGLE: S'il vous plaît inclure des notes pour chaque élément de la
ligne budgétaire pour expliquer comment le montant total demandé a été
calculé.
PORTAGE: Veuillez inclure des notes pour chaque ligne du budget point
d'expliquer comment le montant total demandé a été calculé.
```

Here, both "`S'il vous plaît inclure`" and "`Veuillez inclure`" are good translations of "`Please include`". However, in the EC, PCH, and HC data I've looked at, the preferred French translations for English expressions of the form "Please VERB_INFINITIVE" are of the "Veuillez …" form rather than the "S'il vous plaît …" form. Perhaps this is true in

general across the Government of Canada. This is the kind of difference I expected to see in the data: GT's translation isn't wrong, it just isn't worded in the way typically seen in a given domain. It is obviously preferable to have outputs that reflect preferred translation usage for the domain a post-editor is working in – e.g., changing "S'il vous plaît …" to "Veuillez …" every now and then costs time – so from the productivity point of view, Portage has an advantage here even though GT's translation is also correct.

Other differences are more surprising. A list of these differences is followed by examples illustrating them:

Holes in GT's vocabulary. Both GT and Portage handle out-of-vocabulary ("OOV") items by passing them through to the French translation unchanged. That is, if the system has not seen an English word in its bilingual training data, it is copied to the French side. I would not have guessed that GT would not know how to translate "stacks", "reporting", or "highlight" (see below).

Missing words. For some reason, GT has decided to omit the French translation of an English word (this is not an OOV problem – if it were, the English word would appear unchanged in the French translation).

Problems with French word order. GT prefers to stick with the original, English word order more often than Portage does, and this sometimes leads to bad translations, especially of compound nouns.

Part of speech problems. GT sometimes seems to get the part of speech of an English word wrong (in one of the examples below, it translates the verb "lies" as if it were the noun meaning "untruths"). It also seems to prefer to translate English words to French words with the same part of speech – nouns to nouns, verbs to verbs, etc. – which sometimes yields unidiomatic expressions in French.

Prepositions. Portage seems to translate these correctly more often than GT does.

Perhaps, slightly better handling of syntax by Portage. It is difficult to be sure – both systems make plenty of syntactic errors – but it is my impression that Portage makes somewhat fewer such errors. This is especially true of agreement errors.

I'm not convinced that all these differences in behaviour are due to differences in the training data for Portage and GT. A few years ago, an NRC researcher (Colin Cherry) devised a particularly sophisticated model for handling word order changes during translation; that model is incorporated in Portage. Portage doesn't explicitly model syntax, but it does explicitly model the most likely sequence of parts of speech in the target language (in these experiments, French). Though I wouldn't dare say this in a scientific paper – that would require thorough quantitative analysis of the data - it is possible that Portage has better algorithms than GT for handling word order and parts of speech. The interaction of these algorithms would probably result in better, though far from perfect, handling of syntax.

Here are some examples to illustrate these points.

<u>From EC test1</u>

```
SRC: TABLE A12-5: 2003 Greenhouse Gas Emissions Summary for Prince
Edward Island
REF: TABLEAU A12-5 : Résumé des émissions de gaz à effet de serre pour
l'Île-du-Prince-Édouard, 2003
GOOGLE: TABLEAU A12-5 : 2003 à effet de serre des émissions de gaz
résumé de l' Île-du-Prince-Édouard
```

PORTAGE: TABLEAU A12-5 : Émissions de gaz à effet de serre pour l'Île-du-Prince-Édouard, 2003
Comment: This is an example of Google's problem with French word order.

SRC: Open combustion
REF: Combustion à l'air libre
GOOGLE: Ouvrir combustion
PORTAGE: Combustion à ciel ouvert
Comment: Google sometimes gets parts of speech wrong - here, it translates "open" as if it were an imperative verb.

SRC: Business Numbers (BNs) can be found on all forms issued to a business by the Canada Customs and Revenue Agency.
REF: Ce numéro d'entreprise (NE) est le numéro qu'utilise l'Agence canadienne des douanes et du revenu sur tous les formulaires qu'elle envoie à une même entreprise.
GOOGLE: Le numéro d' entreprise (NE) peuvent être trouvés sur tous les formulaires émis à une entreprise par l'Agence des douanes et du Canada.
PORTAGE: Les numéros d'entreprise (SBST) peuvent être trouvées sur tous les formulaires qu'elle envoie à une même entreprise par l'Agence des douanes et du revenu du Canada.
Comment: Domain-specific knowledge helps Portage here - it knows that "Canada Customs and Revenue Agency" can be translated as "l'Agence des douanes et du revenu du Canada".

SRC: Hexavalent Chromium Compounds
REF: Composés du chrome hexavalent
GOOGLE: Chrome hexavalent
PORTAGE: Composés du chrome hexavalent
Comment: This is one of the cases where Google fails to translate an English word - here, "compounds".

SRC: Reporting of CACs from Stacks & Reporting Physical Parameters
REF: Déclaration des PCA pour les cheminées at les paramètres physiques
GOOGLE: La déclaration des PCA de stacks & reporting paramètres physiques
PORTAGE: La déclaration des PCA provenant des cheminées et des paramètres physiques de déclaration
Comment: Google has an out-of-vocabulary (OOV) problem - it doesn't know how to translate either "stacks" or "reporting", and thus passes these words through into the French translation unchanged.

SRC: Pollution prevention focuses on avoiding the creation of pollutants rather than trying to manage them after they have been created.
REF: La prévention de la pollution consiste à éviter de créer des polluants plutôt que d'essayer de les gérer après coup.

GOOGLE: Prévention de la pollution se concentre sur évitant la création de polluants plutôt que d'essayer de les gérer après qu'ils ont été créés.
PORTAGE: La prévention de la pollution vise à éviter la production de polluants plutôt que de chercher à s'en débarrasser après coup.
Comment: Google doesn't seem to be as knowledgeable about French syntax as Portage is - "se concentre sur évitant" is ungrammatical.

SRC: ARET aims to achieve virtual elimination of emissions of these 30 persistent, bioaccumulative toxic substances and reduced emissions of the other 87 toxic substances to levels insufficient to cause harm.
REF: ARET vise à atteindre: l'élimination virtuelle des émissions de 30 substances toxiques persistantes, bioaccumulables et toxiques; la réduction des émissions de 87 autres substances toxiques àdes niveaux tels qu'ils ne pourront pas causer de dommages.
GOOGLE: ARET vise à atteindre la quasi-élimination des émissions de ces 30 substances toxiques persistantes, bioaccumulables et une réduction des émissions de 87 autres substances toxiques à des niveaux suffisants pour causer des dommages.
PORTAGE: ARET vise la quasi-élimination des émissions de ces 30 substances toxiques persistantes et bioaccumulables et la diminution des émissions de 87 autres substances toxiques à des niveaux insuffisants pour causer des dommages.
Comment: Google has reversed the meaning of part of the sentence by translating "insufficient" as "suffisants" instead of "insuffisants". This error is hard to explain.

SRC: CCME Water Quality Guideline
REF: Recommandation pour la qualité des eaux du CCME
GOOGLE: CCME directrice qualité de l'eau
PORTAGE: Recommandations pour la qualité de l'eau du CCME
Comment: Again, Google has a problem with French word order.

<u>From HC test2</u>
SRC: NOTICE TO THE MEDIA March 2, 2005
REF: AVIS AUX MÉDIAS Le 2 mars 2005
GOOGLE: 2 AVIS AUX MÉDIAS mars 2005
PORTAGE: AVIS AUX MÉDIAS le 2 mars 2005
Comment: Google has a French word order problem.

SRC: Members are only protected when the advice given lies within the mandate of the Committee.
REF: Les membres sont protégés uniquement lorsque l'avis donné s'inscrit dans le cadre du mandat du conseil.
GOOGLE: Les membres ne sont protégés que si les conseils prodigués mensonges dans le mandat de la commission.
PORTAGE: Les membres ne sont protégés que si les conseils prodigués s'inscrivent dans le cadre du mandat du comité.
Comment: Google has a part of speech problem here - it has translated "lies" as a noun ("mensonges" means "untruths").

SRC: Honoraria is offered to members who request it.
REF: Des honoraires seront versés aux membres qui en font la demande.
GOOGLE: Honoraires est offert aux membres qui le demandent.
PORTAGE: On offre des honoraires aux membres qui en font la demande.
Comment: Google has an agreement problem here - "Honoraires" is singular and should be followed by a singular word form.

SRC: 2) Amount of research funding to that industry.
REF: 2) Niveau de financement de la recherche pour cette industrie.
GOOGLE: 2) Montant de financement de la recherche à cette industrie.
PORTAGE: 2) Le montant du financement de la recherche pour l'industrie.
Comment: Google frequently has problems with prepositions - here, it translates "to" over-literally as "à".

SRC: The best way to maintain circulation is provision of adequate fluids and electrolytes.
REF: Le meilleur moyen de maintenir la circulation est de fournir une quantité adéquate de liquide et d'électrolytes.
GOOGLE: La meilleure façon de maintenir la circulation est la fourniture de fluides et les électrolytes adéquats.
PORTAGE: La meilleure façon de maintenir la circulation est d'administrer une quantité suffisante de liquide et d'électrolytes.
Comment: Google has a part of speech problem here - it has translated "provision", an English noun, as a French noun, "fourniture", giving an unidiomatic result.

SRC: The 35-54 year group is expected to increase by 41% over the same time period (Statistics Canada, The Daily, January 13, 1998).
REF: On prévoit que le groupe de 35-54 ans enregistrera une hausse de 41 p. 100 au cours de la même période (Statistique Canada, Le Quotidien, 13 janvier 1998).
GOOGLE: Est prévu les 35-54 ans devrait augmenter de 41% sur la même période (Statistique Canada, Le Quotidien, le 13 janvier, 1998).
PORTAGE: Le groupe 35-54 ans devrait augmenter de 41% au cours de la même période (Statistique Canada, Le Quotidien, le 13 janvier 1998).
Comment: Google fails to translate an English word - here, "group".

SRC: What will the indicators be?
REF: Quelles seront les indicateurs?
GOOGLE: Quels seront les indicateurs être?
PORTAGE: Quels seront les indicateurs?
Comment: Google's translation is too literal here & thus unidiomatic - it should not have translated "be".

<u>From PCH test1</u>
SRC: Based on current trends, by 2016 one fifth of Canada's population will be visible minorities.
REF: Si l'on se fie à la tendance actuelle, d'ici 2016, le cinquième de la population canadienne sera composé de représentants des minorités visibles.

GOOGLE: Selon les tendances actuelles, d' ici 2016, un cinquième de la population du Canada aura minorités visibles.
PORTAGE: En se fondant sur les tendances actuelles, d'ici 2016 un cinquième de la population du Canada seront des membres des minorités visibles.
Comment: Google's syntactic error here is hard to explain – why did it translate "will be" as "aura" ("will have")?

SRC: Highlight of presentations and discussions - html
REF: Faits saillants des présentations et discussions - Version html
GOOGLE: Highlight de présentations et de discussions - html
PORTAGE: Faits saillants des présentations et discussions - html
Comment: Google has an out-of-vocabulary (OOV) problem – it doesn't know how to translate "highlight", and thus passes the word through into the French translation unchanged.

TASATJ-2 : Traduction automatique au Service administratif des tribunaux judiciaires

Rapport final

Michel Simard
Marc Tessier
Traitement de texte multilingue
CNRC

Résumé

Le présent document décrit les travaux effectués durant la deuxième année de collaboration entre le SATJ et le CNRC. Au cours de cette deuxième année, nous avions initialement prévu travailler sur 1) l'amélioration des systèmes de TA ; 2) l'automatisation du processus d'entraînement ; 3) la mise en production des systèmes ; et 4) le routage automatique des données. Toutefois, vu les circonstances difficiles qui ont marqué l'année 2016, nous avons choisi de nous limiter aux deux premiers objectifs.

Les travaux portant sur l'amélioration des systèmes de TA ont permis d'obtenir des gains significatifs, aussi bien en évaluation automatique qu'en évaluation manuelle. Toutefois, ces gains restent modestes, et n'ont pu être obtenus qu'au prix d'une augmentation substantielle de la mémoire et des temps de calcul requis par les systèmes. On peut probablement affirmer que la qualité obtenue se situe aux limites de ce qu'il est raisonnable d'attendre à l'heure actuelle des systèmes de TA statistiques.

Par ailleurs, nous considérons avoir effectué des progrès importants sur le plan de l'automatisation de l'entraînement des systèmes. D'une part, avec l'utilisation du produit AlignFactory, la gestion des corpus d'entraînement et des mémoires de traduction est grandement simplifiée, et ne requiert plus l'intervention d'un expert. D'autre part, partant de ces données, la création des systèmes PORTAGE II est presque entièrement automatisée.

1. Introduction

Au cours de l'année 2015-16, le CNRC et le Service administratif des tribunaux juridiques (SATJ) ont exploré la possibilité d'utiliser au SATJ la technologie de traduction

automatique (TA) PORTAGE II du CNRC, pour produire un premier jet de traduction, devant être révisé (ou post-édité) avant publication. Les conclusions de cette première étude ont été très positives : Bien que la qualité des traductions automatiques soit assez inégale (un phénomène normal), globalement, les sorties de PORTAGE II présentaient une qualité suffisante pour envisager un processus de post-édition systématique. En outre, nous avons établi qu'il serait possible pour le SATJ d'envisager l'utilisation de l'environnement de traduction assistée par ordinateur (TAO) Matecat.com et des mémoires de traduction MyMemory, conjointement avec les systèmes PORTAGE II.

Partant de ce constat positif et de l'intérêt du SATJ pour une approche technologique de cette nature, nous avons convenu de poursuivre le projet, avec des activités qui devaient s'étaler sur quatre phases :

Phase 1 : Amélioration des systèmes de TA

Phase 2 : Automatisation du processus d'entraînement

Phase 3 : Mise en production

Phase 4 : Automatisation du routage

Suite au décès de Lucie Langlois, instigatrice et coordonnatrice du projet au SATJ, les choses ont ralenti considérablement. Par ailleurs, le dépôt par le Commissaire aux langues officielles d'un rapport au Parlement sur les activités du SATJ ont jeté beaucoup d'incertitude quant aux débouchés du projet. Dans ce contexte incertain, nous avons donc décidé de nous limiter aux Phases 1 et 2. Le présent document se veut un bilan du travail effectué au cours de la deuxième année du projet. Nous présentons d'abord dans la Section 2 le contexte dans lequel s'inscrit la collaboration entre le SATJ et le CNRC, et résumons les travaux de la première année . Nous présentons ensuite les travaux de l'année courante : nous décrivons à la section 3.2 les travaux visant spécifiquement à améliorer la qualité des traductions, et à la section 3.1 notre travail sur l'automatisation du processus de construction des systèmes de TA. Nous présentons enfin à la section 3.3 les résultats obtenus. Nous terminons avec nos recommandations pour la suite du projet.

2. Contexte : Collaboration SATJ - CNRC

Le Service administratif des tribunaux judiciaires (SATJ) est un organisme du Gouvernement du Canada. Il offre des services à la Cour d'appel fédérale (CAF), la Cour fédérale (CF), la Cour d'appel de la cour martiale du Canada (CACM) et la Cour canadienne de l'impôt (CCI). En vertu de la Loi sur les langues officielles, ces cours sont tenues de publier leurs décisions en français et en anglais, les deux versions ayant même poids au regard de la loi. C'est le SATJ qui a charge d'assurer la traduction des décisions émanant de ces cours, qui représentent près de dix millions de mots par année, principalement de l'anglais vers le français.

Jusqu'à récemment, toutes ces traductions étaient effectuées en sous-traitance, seule la révision finale étant faite à l'interne (par des jurilinguistes). Afin de réduire à la fois les coûts et les délais liés à la traduction, le SATJ a envisagé effectuer une plus grande partie de ces traductions à l'interne, et se tourner vers les outils d'aide à la traduction

Tableau 1 – Statistiques sur les corpus de textes du SATJ

		Sous-corpus				
		CACM	CF	CAF	CCI	**Total**
Documents	appariés	142	25 260	6521	8799	**40 722**
	Orphelins EN	1	1438	278	423	**2140**
	Orphelins FR	0	465	70	138	**673**
Segments appariés	($\times 1000$)	28	3939	888	1761	**6616**
Mots EN	($\times 10^6$)	0,6	89	17	35	**141,6**
Mots FR	($\times 10^6$)	0,6	103	19	41	**163,6**

("traduction assistée par ordinateur", ou TAO), tels que la traduction automatique (TA) et les mémoires de traduction (MT).

C'est dans ce contexte que Lucie Langlois, directrice générale des services intégrés et de la planification au SATJ, s'est adressé à l'équipe de Traitement de texte multilingue du CNRC, afin de l'aider à atteindre ses objectifs. Ensemble, ces deux organismes ont entamé un projet de collaboration, dont l'objectif était de mesurer le potentiel des outils de TAO pour le SATJ. Nous résumons ci-dessous les grandes lignes de la première étape de cette collaboration, qui s'est déroulée entre l'automne 2015 et l'été 2016.

À partir des archives en format HTML des quatre cours desservies par le SATJ, nous avons d'abord constitué un corpus bilingue d'environ 6,6 millions de paires de segments, soit 140 millions de mots anglais et 160 millions de mots français. Pour ce faire, nous avons eu recours à des méthodes d'alignement de segments sophistiquées, ainsi qu'à des méthodes de filtrage ad hoc pour éliminer les segments ne présentant pas le profil linguistique souhaité. Ce processus nous a permis d'obtenir un ensemble de données d'entraînement tout-à-fait adéquat pour les systèmes de traduction automatique statistiques. (Voir le Tableau 1)

En utilisant ces données, nous avons alors pu produire et évaluer des systèmes de TA reposant sur la technologie PORTAGE II du CNRC : nous avons ainsi obtenu des systèmes spécialisés pour chacune des quatre cours que dessert le SATJ, pour les deux directions, EN→FR et FR→EN. Les systèmes obtenus pour les quatre cours présentent des performances très encourageantes, avec des scores BLEU qui se situent entre 45 et 50 pour la direction EN→FR, et qui approchent les 55 pour la direction FR→EN. (Voir le Tableau 2 ; voir également l'Annexe B pour une présentation du score BLEU).

Nous avons également procédé à des expériences visant à mesurer le potentiel applicatif des mémoires de traduction (MT) pour le SATJ. Ces expériences laissent présager un rendement plutôt modeste pour cette technologie. Nous avons néanmoins conclu qu'il est tout à l'avantage du SATJ de constituer et de garder à jour des archives de sa production sous la forme de mémoires de traduction, de façon à faciliter la récupération des données d'entraînement pour les ré-entraînement périodiques des systèmes de TA.

Par ailleurs, nous avons évalué la possibilité pour le SATJ d'accéder aux systèmes de traduction PORTAGE II par le biais de l'environnement de traduction Web MateCat

Tableau 2 – Performance des systèmes de TA du SATJ, mesurée au moyen de la métrique BLEU

Test	Corpus	EN→FR	FR→EN
Test 1	CACM	45.4	46.5
	CF	50.3	52.6
	CAF	51.7	54.4
	CCI	47.8	52.0
Test 2	CACM	44.1	46.1
	CF	46.5	49.1
	CAF	49.0	51.3
	CCI	43.1	47.6

(voir l'Annexe C). Bien que le mode d'intégration des moteurs de traduction dans MateCat soit sous-optimal pour PORTAGE II, l'intégration résultante s'avère tout-à-fait adéquate pour une utilisation modérée des systèmes.

Nous avons déployé les systèmes de TA sur un serveur Web et à travers l'intégration dans Matecat, afin de permettre à nos partenaires du SATJ d'en faire l'essai. En pratique, le SATJ a d'abord procédé à une évaluation de la qualité linguistique des traductions produites (Macklovitch, 2016a). Cette évaluation a relevé un certain nombre de lacunes des systèmes PORTAGE II, notamment en ce qui a trait à la traduction des expressions numériques, de certaines entités nommées et des accords en genre et en nombre. Par la suite, le SATJ a effectué un essai opérationnel, impliquant deux traducteurs stagiaires et des réviseurs (Macklovitch, 2016b). Cet essai s'est avéré très encourageant : dans son rapport, Elliott Macklovitch concluait :

> [...] the results of our trial strongly suggest that CAS could substantially reduce its current translation backlog by introducing Portage into its translation workflow. How best to go about this – whether CAS should set up its own internal translation service or send the texts pre-translated by Portage to an outside language service provider – is another question. What we have been able to demonstrate, however, is that the machine translation technology developed by the NRC has the potential to significantly improve a situation that up until now has appeared intractable

3. Travaux de la deuxième année

Comme nous l'avons expliqué en introduction, compte tenu des circonstances difficiles dans lesquelles s'est déroulée la deuxième année de cette collaboration, nous avons choisi de nous limiter aux Phases 1 et 2 du projet, qui traitent respectivement de l'amélioration des moteurs de traduction et l'automatisation de la procédure d'entraînement de ces systèmes. Nous élaborons sur ces différentes tâches ci-dessous.

3.1. Automatisation des procédures

Les systèmes PORTAGE II produits pendant la première année du projet l'ont été dans des conditions "de laboratoire", c'est-à-dire par les chercheurs experts du CNRC, qui ont produit des composantes individuelles, et ont ensuite expérimenté les meilleures façons de les combiner. Ce faisant, nous nous sommes passablement éloignés de la procédure standard d'entraînement des systèmes de TA, telle que normalement déployée chez les clients commerciaux de PORTAGE II. De plus, le développement de ces systèmes s'est fait durant une période de transition, durant laquelle la version commerciale de PORTAGE II subissait des changements majeurs.

Si l'actuel projet de collaboration entre le SATJ et le CNRC doit déboucher un jour sur un déploiement opérationnel au SATJ, il est essentiel d'automatiser le plus possible la procédure d'entraînement. Cette automatisation est nécessaire entre autres choses parce que, pour bénéficier des avantages de la traduction automatique statistique, il est nécessaire de ré-entraîner périodiquement les systèmes. La fréquence d'entraînement est variable et dépend du contexte d'utilisation. Dans certains cas, il est souhaitable de ré-entraîner quotidiennement, alors que dans d'autres contextes, trois ou quatre mises-à-jour par année sont suffisantes. A priori, le SATJ tomberait sans doute plutôt dans cette dernière catégorie. Malgré tout, cette procédure doit nécessiter le moins d'interventions manuelles possible, surtout si elle doit être effectuée par le personnel du SATJ.

Nous avons donc effectué les changements nécessaires pour faciliter autant que possible l'entraînement. Nous décrivons ci-dessous les principales composantes de ce nouveau mode d'entraînement.

3.1.1. Prétraitement avec AlignFactory

Comme nous l'avons expliqué plus tôt (Section 2), nous avons initialement obtenu le corpus du SATJ sous la forme de documents HTML, dont nous avons extrait et aligné le corpus au moyen de différents outils ad hoc, certains extraits de la trousse PORTAGE II, d'autres des programmes d'utilisation générale. Cette façon de faire était non seulement compliquée à répliquer, elle donnait également des résultats plus ou moins satisfaisants, notamment en ce qui a trait à l'extraction du contenu textuel des fichiers HTML.

Nous avons donc fait le choix de remplacer la plus grande partie de cette chaîne de traitement par le produit AlignFactory, commercialisé par la firme Terminotix. Ce programme se charge d'apparier les fichiers, d'en extraire le contenu textuel, et d'en tirer des paires de segments alignés, directement sous la forme d'une mémoire de traduction. Outre le fait de combiner toutes ces opérations de façon propre, AlignFactory a l'avantage de pouvoir traiter un grand nombre de formats en plus du HTML, notamment les formats Microsoft Office et PDF.

Notre expérience avec ce produit a été très positive. L'utilisation du logiciel est beaucoup plus simple que notre chaîne de traitement antérieure, les résultats sont très satisfaisants, et le résultat peut être exploité directement par la plupart des systèmes de mémoire de traduction.

3.1.2. L'environnement *Framework* de PORTAGE II

Comme nous l'avons mentionné plus haut, les premiers systèmes de TA du SATJ ont été produits dans des conditions de laboratoire, avec tout ce que ça implique d'interventions manuelles. Nous avons maintenant entièrement transféré cette procédure dans l'environnement d'entraînement de PORTAGE II , que nous appelons le *Framework*. Cette opération était compliquée par le fait que les systèmes du SATJ sont en fait des systèmes combinés : pour chaque direction de traduction – anglais-français et français-anglais – nous produisons des systèmes spécialisés pour chacune des quatre cours – CACM, CF, CAF et CCI. Toutefois, chacun de ces systèmes est entraîné à partir de la totalité des données d'entraînement, en provenance des quatre cours. Cette éventualité n'était pas prévue dans le Framework, il a donc fallu mettre en place les mécanismes nécessaires.

Le résultat est une procédure intégrée qui, à partir des mémoires de traduction extraites par AlignFactory, produit huit systèmes de traduction spécialisés pour chacune des cours et des directions de traduction.

3.2. Amélioration de la qualité des traductions

Le rapport du SATJ sur la qualité linguistique des traductions (Macklovitch, 2016a) soulignait certains problèmes des systèmes de TA, notamment en ce qui a trait aux expressions numériques, aux accords grammaticaux en français, et au traitement des "entités nommées", c'est-à-dire les noms de personnes, de lieux, d'organismes, etc. Nous décrivons ci-dessous nos efforts pour atténuer ces problèmes.

3.2.1. Traitement des expressions numériques

Par défaut, les systèmes de TA PORTAGE II apprennent à traduire les expressions numériques – dates, montants, numéros de dossier ou de paragraphes, etc. – de la même façon que tout le reste : par l'analyse statistique de leurs occurrences en contexte, dans les données d'apprentissage. En pratique, cette façon de faire peut sembler hautement sous-estimale : après tout, par leur nature, il existe une infinité de telles expressions, et par ailleurs elles constituent un aspect de la langue qui semble se prêter particulièrement bien à des règles de traduction systématiques. C'est pourquoi la technologie PORTAGE II comprend un mécanisme de pré-traitement des expressions numériques, qui effectue leur traduction en marge du traitement normal. Ce mécanisme n'est pas activé par défaut : à l'expérience, nous nous sommes aperçus que la traduction des expressions numériques, bien que très systématique, est également très dépendante du domaine d'application. Voici quelques exemples :

— En anglais, dans l'écriture des nombres décimaux, on utilise le point "." pour séparer la partie entière de la partie décimale, alors qu'en français, on utilise la virgule ",". Lorsqu'on traduit des nombres, il convient donc de remplacer les points par des virgules. Le problème, c'est que dans le contexte du SATJ, la numérotation des sections, paragraphes, etc. emploie souvent une notation qui est ambiguë avec les nombres décimaux : "section 2.1", "paragraphe 18.2". Il est donc nécessaire de formuler des règles particulières pour ces textes.

— Pour les nombres plus grands que 10 000, l'anglais sépare les groupes de trois chiffres avec des virgules, alors qu'en français, on emploie plutôt un espace. Ici encore, on peut en déduire une règle de traduction simple, impliquant le remplacement des virgules par des espaces. Problème : dans les textes juridiques, lorsqu'un tel nombre s'avère être un numéro de brevet ("US Patent 1,234,567"), les virgules doivent être préservées en français.

Nous avons donc dû écrire un module de traitement des expressions numériques spécifique au SATJ. Lorsque ce module est activé, certaines parties du texte sont traduites au moyen de règles, alors que d'autres sont traduites au moyen des mécanismes statistiques propres à PORTAGE II. À l'expérience, nous nous sommes aperçus que certains types d'expressions souffraient de ce mode d'opération mixte, notamment les dates. Nous avons donc également incorporé dans le module un traitement intégral des dates.

Notons que ce module de traitement des expressions numériques a été déployé dès l'été 2016 sur les systèmes présentement en production. Nous n'avons pas effectué de mesure précise de sa performance, mais depuis sa mise en service, il est clair que les problèmes avec ce genre d'expressions rapportés dans le rapport d'Elliott Macklovitch ont essentiellement disparus.

3.2.2. Accords grammaticaux et syntaxe

Les stagiaires qui ont participé à l'essai de PORTAGE II au SATJ ont rapporté qu'ils rencontraient souvent des problèmes liés aux accords grammaticaux en français. Ça n'est pas une surprise : c'est un problème bien connu des systèmes statistiques (de traduction ou autre). Ce problème est d'autant plus criant lorsqu'il s'agit de dépendances non-bornées, c'est-à-dire lorsque les deux éléments qui doivent s'accorder (par exemple, un pronom et le nom qu'il remplace) sont éloignés dans la phrase, ou même lorsqu'ils apparaissent dans des phrases distinctes.

Il existe bien des mécanismes permettant de repérer et de corriger ce genre d'erreur. Malheureusement, leur utilisation est souvent très coûteuse sur le plan informatique, et leur efficacité est jusqu'à maintenant assez limitée.

Par ailleurs, la plus récente version de PORTAGE II permet d'utiliser des modules statistiques qui, dans certains cas, peuvent atténuer ce genre de problème. Une des composantes essentielles des systèmes de TA statistiques est le modèle de langue-cible (ou simplement modèle de langue – *language model* ou *LM* en anglais). Son rôle est de guider la production de la traduction, en favorisant les suites de mots les plus probables dans la langue cible. Les systèmes PORTAGE II peuvent incorporer plusieurs modèles de langues. Récemment, l'équipe PORTAGE II a mis au point plusieurs types de modèles qui peuvent être combinés avantageusement avec le modèle de langue standard.

modèles génériques : les modèles de langue standard sont créés à partir des données d'entraînement spécifiques au domaine d'utilisation. Par exemple, dans les systèmes PORTAGE II présentement en opération au SATJ, on utilise des modèles entraînés avec le corpus du SATJ. Récemment, nous avons constaté qu'il était souvent bénéfique d'utiliser, conjointement à ce modèle, un modèle de langue "générique",

créé à partir d'un très grand corpus, composé de textes provenant d'une large palettes de genres et de domaines différents. Ces modèles confèrent au système une connaissance de la langue "générale", en quelque sorte.

modèles de langue bruts (*Coarse language models*, ou CLM ; voir Stewart et collab. (2014)) : Les modèles de langue standard ont normalement une portée très courte, parce qu'il considèrent les probabilités de co-occurrence des mots dans une fenêtre limitée, typiquement 3 ou 4 mots. Les CLM arrivent à étendre substantiellement cette portée, en considérant des classes générales de mots plutôt que des mots spécifiques. De ce fait, ils peuvent parfois contribuer à résoudre les problèmes de dépendance non-bornée.

modèles neuronaux conjoints (*Neural Network Joint Models*, ou NNJM ; voir Devlin et collab. (2014)) : Tout comme les CLM, ce sont des modèles de langue. Comme leur nom le suggère, ces modèles s'appuient sur les récentes avancées en apprentissage profond (*deep learning*), et notamment sur la capacité de généralisation des réseaux neuronaux. À la différence des autres modèles de langue toutefois, ils disposent également d'une connaissance du contexte dans la langue source. L'expérience démontre qu'ils contribuent à améliorer la qualité des traductions, notamment en favorisant des choix lexicaux plus judicieux, tout en tenant compte des contraintes syntaxiques locales. Notons que la création des NNJM requiert l'utilisation d'un processeur graphique (*Graphics processing unit* – GPU), ce qui rend normalement prohibitif leur déploiement dans les systèmes commerciaux. Toutefois, le contexte particulier de notre collaboration avec le SATJ, en particulier le fait d'entraîner et d'héberger les systèmes sur les serveurs du CNRC, a rendu possible leur exploitation ici.

attributs dispersés (*sparse features* ; voir Watanabe et collab. (2007); Chiang et collab. (2009)). Il s'agit d'un mécanisme général permettant d'incorporer dans un modèle statistique une grande quantité (typiquement : des milliers) de tests très spécifiques, couvrant notamment des problèmes d'omission (le système omet de traduire un mot important de la source) et d'insertion sporadique (le système produit une phrase qui contient un mot dont la traduction n'apparaît pas dans la source). Ce mécanisme ne cible pas spécifiquement des problèmes syntaxiques, mais son utilisation a généralement un effet bénéfique sur la qualité des traductions produites.

3.2.3. Entités nommées et traitement de la casse

Les textes des différentes cours contiennent de nombreuses références à différents types d'entités : personnes, organismes, lieux, etc. Les traductions de celles-ci posent souvent problème, du fait que bon nombre d'entre elles n'apparaissent que dans un seul document. De ce fait, PORTAGE II ne connait souvent pas la traduction officielle de ces noms. Ceci engendre plusieurs problèmes. Dans le meilleur des cas, le nom est rendu de la même façon dans les deux langues, et la stratégie qui consiste à recopier dans la sortie ces "mots hors-vocabulaire", produit le résultat souhaité. Mais bien souvent, on assiste à des résultats déplorables. Par exemple, lorsque "Ms. Rice" se voit traduit par "Mme Riz", ou que "Courtyard Terrace Inc." devient "préau terrasse".