Machine Translation (2005) 19: 83–112
DOI 10.1007/s10590-005-2403-4

The Long-Term Forecast for Weather Bulletin Translation

PHILIPPE LANGLAIS, SIMONA GANDRABUR, THOMAS LEPLUS and GUY LAPALME
DIRO/RALI, Département d'informatique et de recherche opérationnelle, Université de Montréal, C.P. 6128, Montréal, H3C 3J7, Canada
E-mail: {felipe,gandrabu,leplus,lapalme}@IRO.UMontreal.CA

Abstract. Machine Translation (MT) is the focus of extensive scientific investigations driven by regular evaluation campaigns, but which are mostly oriented towards a somewhat particular task: translating news articles into English. In this paper, we investigate how well current MT approaches deal with a real-world task. We have rationally reconstructed one of the only MT systems in daily use which produces high-quality translation: the MÉTÉO system. We show how a combination of a sentence-based memory approach, a phrase-based statistical engine and a neural-network rescorer can give results comparable to those of the current system. We also explore another possible prospect for MT technology: the translation of weather alerts, which are currently being translated manually by translators at the Canadian Translation Bureau.

Key words: corpus-based MT, Translation Memory, statistical MT, bootstrapping, rescoring, MÉTÉO

1. Introduction

Machine Translation (MT) is a field nowadays strongly anchored in a paradigm of performance. Evaluation exercises such as those conducted within the TIDES project are flourishing, where the shared task usually consists in translating news article excerpts from a foreign language into English. While this is certainly a challenging issue, real life applications of MT in a production setting (i.e. without human revision) will likely be more targeted than newspaper articles.

More focused evaluation exercises do exist. Within the IWSLT Workshop (Akiba et al., 2004), the main objective was to provide an evaluation framework for spoken-language translation technologies. The shared task consisted in translating sentences from the Basic Travel Expression Corpus (BTEC) which gathers sentences believed to be useful for a tourist in a foreign country. In the *Verbmobil* project (Wahlster, 2000), transcriptions

of spontaneous speech from several narrow domains such as appointment scheduling were translated from German into English.

In this study, we focus on an even more concrete task and one of the greatest successes of MT. We specifically chose this task because there already exists a fully operational rule-based translation system designed for it, whose performance was carefully measured (Macklovitch, 1985), and because we had the chance to build a large bitext of previously published weather forecasts.

In the mid-1970s, a group of linguists and computer scientists at Université de Montréal (the TAUM group) developed an MT system to translate weather reports from English into French, which became known as TAUM-MÉTÉO. A general overview of this project can be found in Isabelle (1987) and several descriptions of the MÉTÉO system including historical notes can be found in Hutchins (1986), Chapter 13 and Hutchins and Somers (1992), Chapter 12. The system involves three major steps: dictionary look-up, syntactic analysis, and a light syntactic and morphological generation.

The transfer from English to French was encoded at the word level into three special-purpose lexicons: idioms (e.g. *blowing snow* ↔ *poudrerie*), locations (e.g. *Newfoundland* ↔ *Terre Neuve*) and a general dictionary containing syntactic and semantic features, such as (1),

(1) *amount* = N((F, MSR), *quantité*)

which means that *amount* translates into the feminine (F) measure (MSR) noun (N) *quantité*.

The syntactic stage is the result of a detailed analysis that was developed by hand at an early stage of the prototype. Chandioux (1988) reports that MÉTÉO-2, a subsequent system that became operational at Environment Canada, used 15 different grammars categorized into five major types from which the syntactic analysis chooses the most appropriate one.

The third and last step performs French word reordering (e.g. adjectives are placed after the noun they modify), preposition selection (e.g. ***à*** *Montréal* 'in Montreal', but ***en*** *Nouvelle-Écosse* 'in Nova Scotia', and ***au*** *Manitoba* 'in Manitoba') plus a few morphological adjustments (e.g. *le été* → *l'été* 'the summer').

The MÉTÉO system and its derivative successors has been in continuous use since 1984 translating up to 45,000 words a day. It runs under the supervision of professional translators from the Canadian Translation Bureau who may be occasionally prompted to correct machine output when the input English text cannot be parsed, often because of spelling errors in the original English text. MÉTÉO has often been called the most successful application of MT technology in history.

One of the reasons for the success of the MÉTÉO system is the nature of the problem itself: a specific domain, with very repetitive texts that are

FPCN18 CWUL 312130	*FPCN78 CWUL 312130*
SUMMARY FORECAST FOR WESTERN QUEBEC ISSUED BY ENVIRONMENT CANADA	*RESUME DES PREVISIONS POUR L'OUEST DU QUEBEC EMISES PAR ENVIRONNEMENT CANADA*
MONTREAL AT 4.30 PM EST MONDAY 31 DECEMBER 2001 FOR TUESDAY 01 JANUARY 2002. VARIABLE CLOUDINESS WITH FLURRIES. HIGH NEAR MINUS 7.	*MONTREAL 16H30 HNE LE LUNDI 31 DECEMBRE 2001 POUR MARDI LE 01 JANVIER 2002. CIEL VARIABLE AVEC AVERSES DE NEIGE. MAX PRES DE MOINS 7.*
END/LT	*FIN/TR*

Figure 1. An example of an English weather report and its French translation.

particularly unappealing for a human to translate (see for example the reports shown in Figure 1). Furthermore, the life of a weather report is, by nature, very short (approximately 6 hours), which make them an ideal candidate for automation (Grimaila and Chandioux (1992).

Given that recent corpus-based approaches to MT have proven their value in some contexts, we decided to see how well these approaches would fit in the context of weather-report translation. We obtained from Environment Canada 309,531 forecast reports in both French and English produced during 2002 and 2003.[1] We used this corpus as a source for developing different MT systems for translating weather reports and thus gave rebirth to one of the most successful MT systems in the history of the field.

We describe in Section 2 the METEO bitext we compiled for this study. Owing to the repetitiveness of this material, we first considered a very simple but efficient approach: a sentence-based translation memory which we describe in Section 3. We compared this to a more generic approach: a phrase-based statistical translation engine (Section 4). We also considered two other approaches that work on the output of other systems: a bootstrapping approach involving the multiple alignment of translations output by one or several engines (Section 5) and a neural network capable of rescoring the output of a native engine (Section 6).

We discuss in Section 8 the results of our experiments on the METEO task, and analyse the main errors produced by our best system. In Section 9, we report on experiments we conducted on a more challenging METEO task: the automatic translation of weather alerts issued almost daily by Environment Canada and which are currently being translated manually. We conclude with some final discussion in Section 10.

2. The MÉTÉO Corpus

The approaches we investigated in this study are all corpus-based; therefore, the first thing we did was to collect a MÉTÉO bitext, i.e. an aligned corpus of corresponding sentences in French and English weather reports. Like all work on real data, this conceptually simple task proved to be more complicated that we had initially envisioned. Indeed, it required about 1,500 lines of Perl code and a few weeks of monitoring; the details can be found in Leplus (2004).

2.1. THE RAW CORPUS

We received from Environment Canada files containing both French and English weather forecasts produced during 2002 and 2003. Both the source report, usually in English, and its translation, produced either by a human or by the current MÉTÉO system, appear in the same file. One file contains all reports issued for a single day. A report is a fairly short text, on average 304 words, in a telegraphic style: all letters are capitalized and non-accented and almost always without any punctuation except for a terminating period.

As can be seen in the example in Figure 1, a report usually starts with a code identifying the source which issued the report. For example, *FPCN18 CWUL 312130* indicates that the report was produced at 21.30 on the 31st day of the month; *CWUL* is a code corresponding to Montreal and the western area of Quebec. A report (almost always) ends with a closing markup: **END** or *FIN* according to the language of the report. If the author or the translator is a human, their initials are added after a slash following the markup. We used this signature to segment a file into its several weather forecasts and qualified as English or French a segment ending with **END** or *FIN* respectively. Given the fact that we started with a fairly large amount of data, we decided to discard any forecast that we could not identify with this process.

2.2. CREATING A BITEXT

To create a bitext from this selected material, we first automatically segmented the reports into words and sentences using an in-house tool that we did not try to adapt to the specificity of the weather forecasts. We then ran the Japa sentence aligner (Langlais, 1997) which took around two hours on a desktop workstation to align 4.2 million pairs of sentences, from which we then removed about 26,000 (roughly 0.6%) which were not one-to-one alignment pairs.

Table I. Main characteristics of the subcorpora used in this study in terms of number of pairs of sentences, English and French words and tokens. $|\text{sent}|_{e\neq}$ indicates the number of different English sentences in each corpus

Corpus	|pairs|	$\|\text{sent}\|_{e\neq}$	English		French	
			Tokens	Types	Tokens	Types
TRAIN	4 187 041	488 391	30 446 549	10 429	37 284 810	11 141
TRAIN_M	4 187 041	301 459	30 290 318	3 352	37 284 810	4 416
DEV	122 357	21 923	891 641	3 022	1 092 208	3 252
DEV_M	122 357	15 454	887 499	1 681	1 092 208	1 908
TEST	36 228	7 878	269 927	1 874	333 370	1 989
TEST_M	36 228	5 994	268 820	1 378	333 370	1 495
TEST-HARD_M	4 045	2 845	62 571	960	78 615	1 042

The sentences of the bitext are fairly short: on average 7.2 English words and 8.9 French words. Most sentences are repeated, only 8% of the English sentences being unique. About 90% of the sentences to be translated can be retrieved and matched with at most one edit operation, i.e. insertion, deletion or substitution of a word.

We divided this bitext into three non-overlapping sections as reported in Table I: TRAIN (January 2002 to October 2003) for training purposes, DEV (December 2003) for tuning the systems, and TEST (November 2003) for testing them. This way of splitting the bitext is slightly biassed against November and December texts, since the TRAIN corpus has half the amount compared to other months. However, it was deliberately chosen in order to recreate as much as possible the working environment of a system faced with the translation of new weather forecasts.

Arguably, a test corpus sampled uniformly over a one year period might be more representative. But at the same time, it is likely that by doing so, we would observe slightly better performances than the one we report. In any case, if we were to deliver a functional system, we would certainly retrain it on the full bitext.

We identified a few classes of tokens, hereafter "meta-tokens", that were worth treating as a single token: telephone numbers, months, days, time, numeric values, ranges of values and cardinal coordinates. We identified them by means of simple regular expressions. Ambiguous tokens were not handled by this process, such as the French word *EST* which could be either a cardinal point 'east' or a verb 'is'.[2] We postfix the name of a corpus by `M` to indicate that meta-tokenization has been performed on it. It is interesting to note the reduction in vocabulary size that this procedure achieves.

Finally, we distinguish among the sentences of the TEST corpus those that were not found verbatim in the TRAIN corpus. We call them the "hard sentences".

Note that this bitext as well as a few other resources we used in this study are available at **rali.iro.umontreal.ca/meteo**.

3. Memory-based Translation

Because of the highly repetitive nature of the MÉTÉO bitext, we started our investigation by seeing how well a translation memory would do on the translation of weather forecast sentences.

3.1. STRUCTURE OF THE MEMORY

Our memory $\mathcal{M}$ is a set of M entries p_i, each one consisting of e_i, the source sentence, and the set of its k_i translations f_i^j along with their cooccurence count with e_i in the training corpus n_i^j (2).

$$\begin{aligned} &\mathcal{M} = \{p_1, \ldots, p_M\} \\ &p_i = \left(e_i, \{(f_i^j, n_i^j)\}_{j \in [1,k_i]}\right) \quad \text{where } i \in [1, M] \text{ and } k_i \leq K \end{aligned} \tag{2}$$

In order to gain in coverage, we populated the memory with meta-tokenized sentences. The scenario for translating a new sentence e is the following. The source sentence is first preprocessed (in order to account for meta-tokens) into e'. Then, we seek in the memory the N source sentences that are at the shortest edit distance from e'. When there are more than N source sentences in the memory with an equal edit distance from e', we consider the most frequent ones (the approximate frequency of e_i is computed by summing the cooccurence counts n_i^j over j). Let $r = r_1, \ldots, r_N$ be the ranks of these closest entries in the memory. The ranked list of alternative translations, or "candidates" for e is called an "N-best list". It is obtained by ranking each target sentence of p_{r_n} according to a score which favors first the smallest edit distances (source side), then the relative frequency of a translation in its entry (recall that a given source sentence could have multiple translations, as is for instance the case in the example of Figure 2). These many translations will be combined by a technique described in Section 5. The selected material is then postprocessed to remove the meta-tokens introduced by the preprocessing stage. Figure 2 illustrates the overall process for the translation of one sentence.

We identified several parameters that could significantly affect the performance of the system. The main ones are its size (M) and the maximum number (K) of French translations retained for each English sentence.

MONDAY .. CLOUDY PERIODS IN THE MORNING WITH 30 PERCENT CHANCE OF FLURRIES EARLY IN THE MORN ING.

preprocessing

DAY1 .. CLOUDY PERIODS IN THE MORNING WITH _INT1_ PERCENT CHANCE OF FLURRIES EARLY IN THE MORNING.

translation memory

nearest source match (edit distance=3)

SRC **_DAY1_ .. BECOMING CLOUDY EARLY IN THE MORNING WITH _INT1_ PERCENT CHANCE OF FLURRIES IN THE MORNING.**

attested translations

TGT *__DAY1__ .. DEVENANT NUAGEUX TOT EN MATINEE AVEC POSSIBILITE DE __INT1__ POUR CENT D AVERSES DE NEIGE EN MATINEE.*

TGT *__DAY1__ .. NUAGEUX AVEC NEIGE PASSAGERE EN MATINEE AVEC POSSIBILITE DE __INT1__ POUR CENT DE NEIGE TOT LE MATIN*

selection

__DAY1__ .. DEVENANT NUAGEUX TOT EN MATINEE AVEC POSSIBILITE DE __INT1__ POUR CENT D AVERSES DE NEIGE EN MATINEE.

postprocessing

LUNDI .. DEVENANT NUAGEUX TOT EN MATINEE AVEC POSSIBILITE DE 30 POUR CENT D AVERSES DE NEIGE EN MATINEE.

Reference translation:

LUNDI .. PASSAGES NUAGEUX EN MATINEE AVEC 30 POUR CENT DE PROBABILITE D AVERSES DE NEIGE TOT EN MATINEE.

Figure 2. Illustration of a memory-based translation session. The source sentence is first preprocessed and the memory is queried. Here, the memory found only an approximate match with an edit distance of 3. The most frequent translation is then selected (in this case, the first one) and postprocessed.

The size of the translation memory affects our system in two ways. If we store only the few most frequent English sentences and their French translations, the time for the system to look for entries in the memory will be short. But, on the other hand, it is clear that the bigger the memory, the better our chances will be to find the exact sentences we want to translate (or ones within a short edit distance), even if these sentences were not frequent in the training corpus. We measured on the DEV corpus that the percentage of sentences to translate found directly in the memory grows logarithmically with the size of the memory until it reaches approximately 20,000. With the full memory, we can obtain a peek of 87% of sentences found verbatim in the memory.

For the setting of the second parameter (K), it is interesting to note that among the 488,792 different sentences found in our training corpus, 89.5% always have the same translation. This is probably because most of the data we received from Environment Canada is actually machine translated and has not been edited by human translators. Note however, that

post-edited or not, this is the material they published. We come back to this point in Section 8 where we analyse the performance of the best system we designed.

3.2. METRICS

Although, in general, there is no clear consensus over which automatic metric should be used to evaluate the quality of a translation engine, arguably the MÉTÉO task lends itself well to evaluation via Sentence Error Rate (SER), the percentage of produced translations that are identical to the reference translation. In Section 8 we put this metric in perspective, but during the development cycles we also found it useful to consider other metrics. We compute the Word Error Rate (WER) by normalizing by its length the edit distance between a candidate translation and a reference translation (the same weight was given to the three edit operations considered: insertion, deletion and substitution). We also report on the NIST (Doddington, 2002) and BLEU (Papineni et al., 2002) n-gram precision rates,[3] both computed by the script `mteval`.[4] For the computation of all these metrics, a single reference translation only was considered.

3.3. RESULTS

In Table II, we report on the performance of the translations produced on the TEST corpus by the translation memory. The best setting as measured on DEV was used here: $N = 5$, $K = 5$ and $M = 488,792$. We distinguish (left part) the performance measured on the full TEST corpus, and those measured only on the sentences that were not found verbatim in the training corpus (right part). We also report on two series of evaluations, one (labeled `memo`) where all the sentences are considered, and one ($\texttt{memo}_{\neq}$) where only one occurrence of a given sentence was translated. The former figures indicate the overall performance of the engine, while the latter are more indicative of the average quality the system achieves on the different sentences of the MÉTÉO task.

Not surprisingly, the performances measured on the full TEST corpus are much better than those measured on previously unseen source sentences. The difference is especially noticeable on the SER metric where more than 75% of the translations produced in the former case were verbatim to the reference ones, while less than 5% were in the latter case. It is interesting to note that, despite the simplicity of the approach, the translation memory already provides a viable solution to the task. It is likely that more advanced techniques of example-based MT (Carl and Way, 2003) would do better than that, as do the approaches we describe below.

Table II. Performance of the engine on the TEST corpus. The results in the left part of the table are those measured for the full corpus, while the right part concerns only those sentences of the TEST-HARD corpus. In this and subsequent tables, WER and SER are shown as percentages. The lines labeled `memo` report the performances measured on all the sentences, while the lines labeled $\texttt{memo}_{\neq}$ indicate the performance measured on one occurrence only of each sentence

	TEST				TEST-HARD			
	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
`memo`	5.53	23.73	10.9578	87.69	21.02	95.50	9.4048	68.37
$\texttt{memo}_{\neq}$	11.20	49.37	10.8610	78.36	22.58	96.73	9.2936	66.21

4. The SMT approach

The second approach we investigated was to build a phrase-based statistical engine, based on the PHARAOH decoder (Koehn, 2004). PHARAOH is a fast, carefully documented decoder which is easy to use.

4.1. THE SYSTEM

PHARAOH is a noisy channel decoder requiring a language model and an (inverted) translation table. If desired, weighting coefficients as well as a few pruning options can control the behavior of the engine. We split the TRAIN corpus in two subparts, TRAIN-T (4,180,000 pairs of sentences) for training the translation and the language models, and TRAIN-H (8,100 pairs) for tuning the different parameters of the engine.

We trained a Kneser–Ney smoothed trigram language model using the SRILM package (Stolcke, 2002). The perplexity of this model on DEV and TEST is respectively 4.94 and 3.83, which is very low compared to standard benchmarks (Zens and Ney, 2004).

To build our translation table, we first aligned our bitext at the word level. Following a common practice, we used the GIZA++ package (Och and Ney, 2004) to word-align our bitext in both directions (English-to-French and French-to-English).[5] We extended the set of word links that were present in both alignments by adding some links belonging to only one alignment direction, following the heuristics described in Koehn et al. (2003). From the resulting alignment $\mathcal{A}$, we collected the set of pairs of source and target sequences (f_a^b, e_i^j) from all regions $(a, b) \times (i, j)$ in the alignment matrix where none of the source words in f_a^b is aligned to a word not belonging to e_i^j and vice-versa (3).

$$(3) \qquad \begin{array}{l} \forall x \in [a, b],\ \forall y/(x, y) \in \mathcal{A},\ y \in [i, j] \\ \forall y \in [i, j],\ \forall x/(x, y) \in \mathcal{A},\ x \in [a, b] \end{array}$$

We did apply a few length-based heuristics to filter the parameters acquired in this way: (source or target) sequences of at most eight words were considered and we imposed that the length of the longest sequence in a pair was at most twice the length of its counterpart.[6] In so doing, we acquired a model of slightly less than 2 million parameters, a small excerpt of which is presented in Figure 3.

We considered two ways of scoring each parameter. The first is by relative frequency, that is, simply by counting the number of times a given pair (f, e) was seen aligned in the bitext, normalized by the number of times f was seen. The second score we used is the IBM model 1 conditional probability (Brown et al., 1993) (4).

$$(4) \qquad p(e_i^j \mid f_a^b) = (b-a)^{-j+i-1} \prod_{y=i}^{j} \sum_{x=a}^{b} p(e_y \mid f_x)$$

We can control the score PHARAOH optimizes to produce a translation. In our case, we tuned five coefficients: one for the language model, one for the built-in distortion model, two for the translation model (one per score) and one for the word penalty. We sought the best setting by uniformly sampling each parameter range with a small enough step size and picked the best configuration we measured on the TRAIN-H corpus. Starting with an SER of 35.5%, we ended up in this way with a rate of 26.2%. The better configurations were those with high language model and distortion weights, and a weight given to the relative frequency score of the phrase-based model higher than its IBM model 1 counterpart.

4.2. RESULTS

We present in Table III the performance of our phrase-based engine compared to the memory-based engine. Overall, the performance of the SMT

target sequence	source sequence	rel. freq.
TO STRONG SOUTH	*A FORTS DU SUD*	0.00273224
DEVELOP AHEAD OF	*SE LEVERONT A L AVANT D*	0.25
WILL DEVELOP AHEAD OF	*SE LEVERONT A L AVANT D*	0.75
ZERO IN THE AFTERNOON .	*DE ZERO EN APRES-MIDI .*	1

Figure 3. Excerpt of the parameters of the phrase-based model trained on the MÉTÉO bitext. The score is the relative frequency of the English sequence given the French one. The source and target nature of the sequence is with respect to the model.

Table III. Results of the phrase-based statistical engine on the TEST corpora. The memory-based translation results from Table II are reproduced here for comparison purposes

	TEST				TEST-HARD			
	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
memo	5.53	23.73	10.9578	87.69	21.02	95.50	9.4048	68.37
$\text{memo}_{\neq}$	11.20	49.37	10.8610	78.36	22.58	96.73	9.2936	66.21
smt	5.33	25.27	11.2683	88.52	11.24	57.77	11.0598	81.56
$\text{smt}_{\neq}$	8.97	44.96	11.6223	83.17	12.42	63.28	11.0259	79.67

engine is comparable to that obtained by the memory-based approach. While the SMT shows an average SER of 1.5 points higher than the memory, the balance shifts in favor of the SMT (an absolute decrease of 4.4 in SER) when we measure the systems using only one occurrence of each source sentence ($\text{smt}_{\neq}$).

Since the translation model collects pairs of phrases of up to eight words (which is close to the average METEO sentence length), this suggests that whenever a long sentence is seen in the training corpus, we should let the memory translate it, and leave it to the SMT otherwise. We will investigate such a combination in Section 7. It is also interesting to note that the decrease in performance measured on the hard sentences of the TEST corpus is not as drastic as it was with the translation memory; clearly, the SMT engine has more generalization power than our memory.

5. Bootstrapping experiments

Various approaches to MT have different properties, but none is likely to be perfect for a given task. It is therefore tempting to combine several engines with the hope of capitalising on their own merits. Frederking and Nirenburg (1994) first proposed combining the output of different black-box translation engines.

More recently, Bangalore et al. (2001) have shown, on a domain-dependent spoken dialog translation task, that combining the output of several off-the-shelf translation engines resulted in better performance than any one individual engine. Similar results were reported on a more general domain translation task in Bangalore et al. (2002). The key underlying idea of their work is to use the word alignment of the output of different translation engines in order to identify the locus of consensus, which in

turn, helps to produce better output, an operation the present authors call "bootstrapping".

5.1. SYSTEM

We implemented a similar idea to bootstrap the output of the memory-based and the phrase-based approaches we described above. Following Bangalore et al. (2002), we adapted the ClustalW multiple-string aligner first designed for biological sequence alignment (Thompson et al., 1994) to our domain.[7] An example of multiple-sequence alignment from the $N = 10$ best ranked translations output by the memory-based system on a single translation session for the input sentence (5a) with reference translation (5b) is given in Figure 4.

(5) a. **HIGH 12 EARLY THIS MORNING**
b. *MAXIMUM 12 TOT CE MATIN*

In (5), no candidate translation agreed with the reference on every single word, but as is often the case, most of them agree on some units such as *MAXIMUM DE 12* 'high of 12' or *TOT CE MATIN* 'early this morning'.

We then built a lattice out of this alignment that can generate both the produced translations as well as new ones. The lattice corresponding to the example in (5) is given in Figure 5. Using the CARMEL package (Knight and Al-Onaizan, 1999), we found a lowest-cost path in these automata in order to produce a final translation. The five lowest-cost consensus translations produced out of the ones reported in Figure 5 are indicated in (6). This example shows the tendency of the consensus translations to be more consistent with each other than the ones originally provided by the memory. This is also what we observed by casual inspection of the consensus translations we produced over the DEV corpus.

MAXIMUM	*DE*		*12*		*CE*	*MATIN*	.
MAXIMUM			*12*	*ATTEINT*	*CE*	*MATIN*	.
MAXIMUM	*DE*		*12*	*TOT*	*CET*	*APRES-MIDI*	.
MAXIMUM	*DE*	*PLUS*	*12*	*TOT*	*CE*	*MATIN*	.
NAPPES	*DE*	*BROUILLARD*		*TOT*	*CE*	*MATIN*	.
BRUMEUX	*PAR*	*ENDROITS*		*TOT*	*CE*	*MATIN*	.
MAXIMUM	*DE*		*12*		*EN*	*MATINEE*	.
BRUMEUX				*TOT*	*CE*	*MATIN*	.
MAXIMUM	*DE*	*PLUS*	*12*		*CE*	*MATIN*	.
MAXIMUM	*DE*	*MOINS*	*12*		*CE*	*MATIN*	.

Figure 4. Multiple-sequence alignment from the ten best-ranked translations provided by the memory-based system for the source sentence (5).

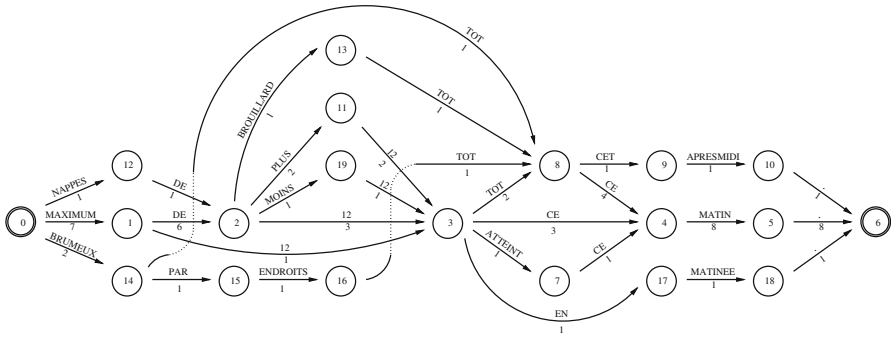

Figure 5. Lattice produced for the translations of Figure 4. The weights on the arcs are the frequency of a given transition. A non-smoothed local bigram language model is obtained by simply normalizing each node by the sum of the weights of the arcs leaving that node.

(6) a. *MAXIMUM DE PLUS 12 CE MATIN.*
b. *MAXIMUM DE 12 CE MATIN.*
c. *MAXIMUM DE PLUS 12 TOT CE MATIN.*
d. *MAXIMUM DE 12 TOT CE MATIN.*
e. *MAXIMUM DE TOT CE MATIN.*

5.2. RESULTS

We tried several variations on this idea. We first considered different ways of weighting an arc of the lattice, using various combinations of the native probability of the automaton and the probability provided by a language model trained on the full target side of the TRAIN material. None of the experiments we conducted with the language model yielded satisfactory results. This might be due to the fact that it is too general a model for discriminating between specific sentences. We finally scored each arc with the native counts obtained at construction time, giving it a credit inversely proportional to its rank in the N-best list where the transition sequence is observed.

We also investigated bootstrapping the translations drawn from the memory, from the SMT engine and from both of them, but observed positive results in the first case only. These are the performances reported in Table IV for the only source sentences of DEV that were not found verbatim in the TRAIN corpus. This is not the setting we used to evaluate the other approaches. The reason is that we conducted these experiments with several development cycles. Shortly after the first cycle (from which Table IV is drawn) we enhanced the SMT and rescoring approaches to a point that discouraged us from pursuing the bootstrapping strategy; even if the

Table IV. Results of the consensus approach on the output of the memory, for the sentences of the DEV corpus not seen verbatim in TRAIN

	WER	SER	NIST	BLEU
Memory	18.69	94.82	9.7853	66.56
+ consensus	18.97	85.53	9.9314	68.86

consensus improved the overall quality of the output of the memory in the first development cycle (a reduction of almost 10 points in SER). Nevertheless, if we were to improve upon this line of work, we would consider other ways of mixing the output of multiple engines, such as the ones recently described in Jayaraman and Lavie (2005) and van Zaanen and Somers (2005).

6. The rescoring approach

In recent years increased attention has been given to rescoring approaches for SMT (Gandrabur, and Foster, 2003; Bender et al., 2004; Blatz et al., 2004; Och et al., 2004; Zhao et al., 2004). For each source sentence, the base system produces an N-best list of translation alternatives, ranked in decreasing translation likelihood order as estimated by the base models. Rescoring consists in using additional information (or using existing information in different ways) to compute a new score for each candidate translation. The N-best list of candidates is then reranked according to the new score, in the hope of improving the accuracy of the resulting translation. One motivation behind rescoring is that it provides a simple framework for using additional sources of information data and feature functions that would be computationally expensive or difficult to integrate within the base SMT models and decoder.

6.1. DESCRIPTION

PHARAOH (Koehn, 2004) can output its search graph in a form which allows the CARMEL package (Knight and Al-Onaizan, 1999) to produce N-best lists. For each source sentence s, we built an N-best list of up to 1,000 different translation alternatives $\{t_j\}_{j\in[1,n\leq 10]}$ using the phrase-based model described above. Each translation alternative, represented as a vector v_j of feature functions and tagged as either correct $\oplus$ or wrong $\ominus$, constitutes a "rescoring example".

We experimented with two different tagging methods. The first is WER-based tagging, where a translation alternative was tagged as correct if and only if its WER with respect to the reference translation was below a fixed threshold (0 in our case). We refer to this approach as "rejection tagging", as it provides a discriminant framework that improved rejection of false translations.

A second approach consisted in tagging a translation as correct if it had the smallest WER rate among all alternatives within the N-best list. This approach, referred to as "reranking tagging", yields slightly better reranking results, as can be seen in Table V.

The rescoring model we used was a multi-layer perceptron (MLP) with a LogSoftMax activation function, trained by gradient descent with a negative-log-likelihood criterion. With this setup, the MLP is trained to estimate $p(\oplus|v_j)$, the conditional probability of correctness of each candidate translation t_j. We experimented with different numbers of hidden units within one single hidden layer and found the best results (on the validation set) with 25 hidden units. All MLP experiments were conducted using the open-source machine learning library Torch (Collobert et al., 2002).

Note that for each translation alternative t_j the base system actually can produce more than one decoding hypothesis h_j^i, depending on how it segmented t_j into chunks produced by the phrase-based model. Each such segmentation returns a different native probability estimate p_j^i.

The rescoring feature functions we used were as follows.

Table V. Results of the rescoring engine on the TEST corpus

	TEST				TEST-HARD			
	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
smt	5.33	25.27	11.2683	88.52	11.24	57.77	11.0598	81.56
smt+	5.27	25.52	11.3090	88.50	11.11	58.34	11.1157	81.44
reject	4.77	21.85	11.2471	90.06	10.87	56.51	11.1045	82.19
rerank	4.55	21.48	11.2449	90.56	10.87	56.44	11.0706	82.35
oracle	3.29	17.36	11.7187	94.22	4.88	33.08	11.8870	90.23
$\text{smt}_{\neq}$	8.97	44.96	11.6223	83.17	12.42	63.28	11.0259	79.67
$\text{smt+}_{\neq}$	8.91	45.15	11.6655	83.05	12.21	63.16	11.0924	79.70
$\text{reject}_{\neq}$	8.71	43.98	11.6527	83.65	12.03	62.61	11.0769	80.28
$\text{rerank}_{\neq}$	8.58	43.79	11.6279	83.98	12.06	62.65	11.0378	80.42
$\text{oracle}_{\neq}$	2.89	21.00	12.5421	92.55	5.36	37.67	11.9802	89.27

– The ratio of the length of s over the length of t_j: For a given pair of languages, this ratio is usually homogeneous.
– From the decoding hypothesis $h_j^r = \text{argmax}_i(p_j^i)$ that has the highest native score among hypotheses h_j^i corresponding to t_j, we retained the native score p_j^r as well as different statistics on chunk size. Longer chunks appear when the translation resembles a reference translation.
– The posterior probability estimate $c(t_j)$ captures the frequency of a translation t_j weighted by the native scores of all its corresponding decoding hypothesis h_j^i (7).

$$c(t_j) = \frac{\sum_i p_j^i}{\sum_j \sum_i p_j^i} \tag{7}$$

– This feature is more significant than the frequency or the native score taken in isolation and is a sound normalization that makes it independent of the sentence length.
– The score of the IBM model 1 and model 2 normalized by the length of t_j. These turned out to be the most significant features (model 2 slightly better than model 1). This is consistent with the findings of Och et al. (2004).

6.2. RESULTS

For training and validation of the MLP, we used examples extracted from DEV: out of a total 901,339 data examples, we kept 700,000 for training and the remaining 201,339 for validation purposes. The testing of the rescoring MLP was performed on the TEST corpus. Table V shows the gain in translation accuracy obtained by the rescoring layer. `smt` is the performance of the native SMT engine described earlier. `oracle` is the performance of the translations produced by assuming an oracle which selects out of the N-best list the translation with the lowest WER. `smt+` is the performance obtained when we add the DEV corpus that was used to train the rescorer to the pool of the training material. `reject` and `rerank` are the results obtained with the rescoring MLPs using the reject-tagging and the rerank-tagging approaches.

Rescoring improves the SER by almost four points over the native system, and performs better than the SMT engine trained as well on the DEV corpus (indeed `smt+` performs slightly worse than `smt`). Considering that retraining the full translation model is more demanding than just training the rescoring layer, this is an encouraging result.

Over the 5,994 different sentences of the TEST corpus, the native translations and the rescored ones differed 1,010 times (16.8%). Out of these

modified translations, 421 (41.6%) increased the overall WER, while 505 (50%) lowered it (the other modifications did not affect the WER). The most fruitful transformation introduced by the rescorer is to make explicit the translation of *DE PROBABILITE* 'chance' in sentences like (8c). Example (9c) similarly shows the translation of *AVERSES DE PLUIE* 'showers' rather than *PLUIE* 'rain'.

(8) a. **60 PERCENT CHANCE OF FLURRIES THIS EVENING**
b. *60 POUR CENT D AVERSES DE NEIGE CE SOIR*
c. *60 POUR CENT DE PROBABILITE D AVERSES DE NEIGE CE SOIR*

(9) a. **SHOWERS BEGINNING THIS EVENING AND ENDING OVERNIGHT.**
b. *PLUIE COMMENCANT CE SOIR ET CESSANT AU COURS DE LA NUIT.*
c. *AVERSES DE PLUIE DEBUTANT CE SOIR ET CESSANT AU COURS DE LA NUIT.*

7. Combination

We have shown that statistical phrase-based translation is better at predicting new sentences than the translation memory alone. This suggests a simple combination scheme of the translation memory and the rescored SMT engine: full sentences found in the memory (3,456 of the 5,994 different source sentences of the TEST corpus) are retrieved from the memory verbatim and the others are translated using the rescored phrase-based SMT system. This combined set-up does in fact yield the best results, as shown in Table VI.

Table VI. Performance on the TEST corpus of a simple combination of the translation memory system and the rescored phrase-based SMT engine

	TEST				TEST-HARD			
	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
memo	5.53	23.73	10.9578	87.69	21.02	95.50	9.4048	68.37
rerank	4.55	21.48	11.2449	90.56	10.87	56.44	11.0706	82.35
combo	4.40	19.37	11.4133	91.20	10.87	56.44	11.0706	82.35
$\text{memo}_{\neq}$	11.20	49.37	10.8610	78.36	22.58	96.73	9.2936	66.21
$\text{rerank}_{\neq}$	8.58	43.79	11.6279	83.98	12.06	62.65	11.0378	80.42
$\text{combo}_{\neq}$	6.75	34.93	11.8811	86.55	12.06	62.65	11.0378	80.42

8. Error Analysis

8.1. QUALITATIVE ANALYSIS

We analysed the output produced by our best system: the one which combines the output of the translation memory and the rescored SMT engine (`combo`). We considered those 1,645 different sentences[8] that the system did not translate identically to their reference counterparts and analyzed the most frequent errors the system committed.

We computed for this purpose a minimum edit distance alignment between reference and candidate translations and identified from these alignments rules that should be applied to transform the latter translation into the former one. As the classical edit distance between two sequences behaves poorly in the case of word reordering, some errors might not have been analyzed correctly by this process. Instead, we could have use an extended distance taking such as the one proposed in Leusch et al. (2003), where a block transposition edit operation is also considered. However, a casual inspection of the errors led us to conclude that this was not necessary.

Typical *errors* encountered are exemplified in (10–13). In each case the first line shows the source sentence, the second the reference translation, the third the automatic translation, the fourth the edit distance alignment (where S indicates a subsitution, I an insertion and D a deletion operation) and finally the errors made.

(10) **TODAY .. PERIODS OF RAIN ENDING NEAR NOON.**
*AUJOURD HUI PLUIE PASSAGERE CESSANT **EN MI-JOURNEE**.*
*AUJOURD HUI PLUIE PASSAGERE CESSANT **VERS MIDI**.*
= = = = = S S =
VERS MIDI⇝EN MI-JOURNEE

(11) **40 PERCENT CHANCE OF FLURRIES LATE THIS AFTERNOON.**
*40 POUR CENT **DE PROBABILITE** D AVERSES DE PLUIE CET APRES-MIDI.*
***POSSIBILITE DE** 40 POUR CENT D AVERSES DE PLUIE CET APRES-MIDI.*
I I = = = D D = = = = = = =
POSSIBILITE DE⇝ϕ, ϕ⇝DE PROBABILITE

(12) **WINDS INCREASING TO SOUTHEAST 30 WITH GUSTS TO 40 IN THE APPROACHES THIS AFTERNOON.**
*VENTS DEVENANT DU SUD-EST **A** 30 AVEC RAFALES A 40 **AUX** ABORDS CET APRES-MIDI.*
*VENTS DEVENANT DU SUD-EST **DE** 30 AVEC RAFALES A 40 **DANS LES** ABORDS CET APRES-MIDI.*
= = = = S = = = = = S I = = = =
DE⇝A, DANS LES⇝AUX

(13) **TEMPERATURE NEAR MINUS 29.**
TEMPERATURES PRES DE MOINS 29.
*TEMPERATURES **DE** PRES DE MOINS 29.*
= I = = = =
DE⇝ϕ

We collected a total of 549 different transformation rules (2,477 occurrences), the most frequent of which are reported in Table VII. Many of these errors turned out to be translations that were fully acceptable or easy to correct.

Some errors are clearly tokenization problems that could be handled very easily (such as *APRES-MIDI⇝*APRES MIDI 'afternoon' or *05H00⇝5H00* which likely occurred because some of the material was manually translated or post-edited). Many of the transformation rules relate to synonymic expressions (*DURANT⇝AU COURS DE* 'during'; *DIMANCHE MATIN⇝EN MATINEE DIMANCHE* 'Sunday morning'; *AU DEBUT DE LA MATINEE⇝TOT LE MATIN* 'early morning') and are therefore correct. The insertion (or deletion) of an article is also a frequent source of divergence, but we felt that in most of the cases, this was legitimate. The most frequent error is the use of *DE* for *A* or vice-versa (see example (12)). Only occasionally is the system patently wrong, as for instance the literal translation (14b) rather than (14c) produced for the source sentence (14a).

(14) a. **EXPECTED RAIN AMOUNT OF 30 MM .**
b. *PREVUE PLUIE ACCUMULATION DE 30 MM .*
c. *QUANTITE PREVUE DE 30 MM .*

Table VII. The 16 most frequent errors found on the translations produced by the `combo` system. These rules account for 50% of the mismatches between the candidate and the reference translations

Freq.	Rule	Freq.	Rule
212	*DE⇝A*	41	*DE⇝ϕ*
178	*A⇝DE*	36	*APRES-MIDI⇝APRES MIDI*
174	*ϕ⇝DE PROBABILITE*	34	*EN MATINEE⇝LE MATIN*
167	*ϕ⇝POSSIBILITE DE*	31	*ϕ⇝AVERSES DE*
72	*DURANT⇝AU COURS DE*	26	*DU⇝AU*
66	*VERS MIDI⇝EN MI-JOURNEE*	25	*D⇝DE L*
50	*ϕ⇝DE*	24	*ϕ⇝AVERSES*
46	*DES⇝ϕ*	23	*ϕ⇝DE PLUIE*

8.2. QUANTITATIVE ANALYSIS

According to the qualitative analysis we conducted, it seems clear that most translations produced are good ones. But we might still wonder how our systems compare to the actual performance of the MÉTÉO system. Unfortunately, we could not access the rule-based system directly, so we had to rely on its published outputs to infer its results; and even there, we had no indication of the level of revision done on the raw output of the system.

The only carefully described evaluation of MÉTÉO we could find is the one the Translation Bureau conducted on the MÉTÉO-2 system twenty years ago (Macklovitch, 1985). We have reason to believe that the system has not changed substantially since then except for an update of its dictionaries and its computing infrastructure. In this study, Macklovitch sampled a set of 1,257 sentences produced over a 24-hour period by Environment Canada. He counted the number of times the machine translation was identical to the final revised version. However, he took care to remove those errors that arose as a result of typos or clear omissions in the original source (English) text.

He found that only 11% of the sampled sentences were different from the revised ones. This evaluation setting roughly corresponds to the SER. Macklovitch also reports that a requirement for the MÉTÉO system then was that at least 80% of the sentences submitted to the system should be translated without any human post-editing.

While the corpus-based approaches we developed almost meet this last requirement, we must admit that the SERs we measured are still higher than the one Macklovitch measured on the MÉTÉO2 system. This might look at first a bit disappointing, but this comparison must be taken with a grain of salt. First, our evaluation was conducted over a much larger test set (36,228 sentences in our case). Second, we have already observed that the reference is not always consistent, nor is it the only possible translation. In fact, we observed that 7% of the translations of English sentences found in the memory did not match their single reference translation but had in fact been revised. Indeed, an informal evaluation carried on a random sample of translations produced by the translation memory that differed from the reference was conducted in Leplus et al. (2004) and revealed that 77% of these "bad" translations could be judged correct.

In any case, our goal in doing this study was not to outperform the hypothetical performance range of the MÉTÉO system. After all, it is unlikely that any of the approaches we investigated here could have been tested without the outputs of the MÉTÉO system. And even if we had

outperformed the current system, it would not mean that the Canadian Translation Bureau would migrate from it.

9. Prospect: Translating Weather Alerts

Given the success we had in translating MÉTÉO weather forecasts, we decided to challenge our systems with another type of weather bulletin: the weather alerts which are issued more sporadically by Environment Canada. These alerts are at present translated by humans because the current MÉTÉO system cannot deal with them. Given the urgency of this information for the public, the alerts must be broadcast rapidly and Environment Canada is looking for ways of speeding up the delivery of this information in both French and English. They have provided us with five years' worth of different types of weather alerts.

9.1. CORPUS

We created for this experiment a new bitext which we call hereafter the ALERTS bitext. The raw material came from three different sources of severe weather warnings, issued by Environment Canada over a period of five years. This raw material contained a total of 21,061,427 words out of which we only kept a fifth for our experiments.

Several reasons explain why we rejected so much raw material. First, the raw bulletins contain large quantities of irrelevant data, such as delimiters, repetitive key-words, and indicators of the times, dates, and places at which the warnings were issued. Also, we use only the bulletin core, a description of the weather phenomenon that triggered the warning written in natural language. The challenge in building the bitext therefore consisted in locating the bulletin cores and aligning the French with the corresponding English cores. Secondly, for processing convenience, we kept only the more recent bulletins (about two thirds of the total number) which had special delimiters for the bulletin cores and were easier to process automatically. Finally, we filtered out inconsistencies in the English–French aligned bulletin cores, such as empty French bulletins or bulletins containing both English and French.

Once again, preparing the data was more complicated than we initially anticipated, but we eventually ended up with a bitext whose main characteristics are reported in Table VIII. We distinguished the three different types of warnings: severe storm warnings (SW), tornado warnings (TW), and omnibus bulletins (OB) that are longer summaries of all weather warnings and watches over specific areas and periods of time. An example of an alert and its (manual) translation is given in Figure 6.

Table VIII. Main characteristics of the ALERTS bitext. *M* postfixes the meta-tokenized versions

	Number of e-sentences			Number of ≠ e-sentences			e-types all
	OB	SW	TW	OB	SW	TW	
TRAIN	78 005	21 002	0	51 357	5 750	0	6 454
DEV	497	504	665	444	192	251	1 466
TEST	2 089	2 209	0	1 532	653	0	3 183
TRAIN_M	78 005	21 002	0	50 715	5 190	0	5 808
DEV_M	497	504	665	444	181	247	1 340
TEST_M	2 089	2 209	0	1 526	600	0	2 124

A STRONG ARCTIC RIDGE OF HIGH PRESSURE WILL MAINTAIN VERY COLD CONDITIONS ACROSS SOUTHERN MANITOBA FOR THE NEXT 48 HOURS.	*UNE FORTE CRETE DE L ARCTIQUE MAINTIENDRA DU TEMPS TRES FROID SUR LE SUD DU MANITOBA AU COURS DES 48 PROCHAINES HEURES.*
TEMPERATURES NEAR MINUS 30 COMBINED WITH BRISK NORTH WEST WINDS NEAR 15KM / H WILL CONTINUE TO PRODUCE WIND CHILL VALUES IN EXCESS OF MINUS 40 THIS MORNING.	*L EFFET COMBINE DES TEMPERATURES DE PRES DE MOINS 30 ET DES VENTS VIFS DU NORD-OUEST DE PRES DE 15 KM / H VA CONTINUER A PROVOQUER UN REFROIDISSEMENT EOLIEN DEPASSANT MOINS 40 CE MATIN.*
FROSTBITE IS POSSIBLE WITHIN 10 MINUTES IN THESE CONDITIONS.	*DES ENGELURES SONT POSSIBLES EN MOINS DE 10 MINUTES DANS DE TELLES CONDITIONS.*
APPROPRIATE COLD WEATHER PRECAUTIONS ARE ADVISED.	*DES VETEMENTS APPROPRIES AU FROID SONT DE RIGUEUR.*
CONDITIONS WILL IMPROVE SLIGHTLY THIS AFTERNOON AS TEMPERATURES MODERATE, HOWEVER EXTREME WIND CHILLS IN EXCESS OF MINUS 40 WILL REDEVELOP TONIGHT AND PERSIST INTO THURSDAY MORNING.	*LA SITUATION S AMELIORERA LEGEREMENT CET APRES-MIDI CAR LES TEMPERATURES S ADOUCIRONT, CEPENDANT UN REFROIDISSEMENT EOLIEN EXTREME DEPASSANT MOINS 40 REPRENDRA CETTE NUIT ET PERSISTERA JEUDI MATIN.*

Figure 6. An example of a report from the OB subcorpus.

9.2. RESULTS

We translated this material with both the memory and the SMT engine (with and without rescoring). The results obtained on the different sentence types of the TEST material are reported in Table IX, while the overall

Table IX. Performance of the translation memory engine (memo), the SMT engine (smt) and the rescored SMT engine (reject). M indicates the MÉTÉO bitext; A denotes the ALERTS bitext, and M+A stands for a mix of both corpora (the MÉTÉO bitext plus 10 times the ALERTS one yielded the best performance on the DEV set)

	Train		Tuning		Metrics		
	tm	lm	tm	WER	SER	NIST	BLEU
SW							
$\texttt{memo}_{M,e\neq}$	M			75.41	100.00	1.5603	14.89
	A			40.96	84.33	6.2593	47.89
$\texttt{smt}_{M,e\neq}$	M	M	M	49.98	99.50	5.9505	37.80
	A	M	M	42.74	98.67	6.4270	42.84
	A	A	M	24.71	86.33	8.8281	65.56
	M+A	A	A	24.73	86.50	8.7795	66.69
$\texttt{reject}_{M,e\neq}$	M+A	A	A	24.47	86.17	8.8006	66.89
$\texttt{oracle}_{M,e\neq}$	M+A	A	A	16.39	72.00	9.7267	75.04
OB							
$\texttt{memo}_{M,e\neq}$	M			71.76	99.93	2.5895	19.35
	A			38.72	82.44	8.2735	52.78
$\texttt{smt}_{M,e\neq}$	M	M	M	46.48	99.15	7.3272	42.05
	A	M	M	40.10	98.36	7.8259	46.62
	A	A	M	26.63	91.87	9.9326	63.36
	M+A	A	A	27.03	92.53	9.7840	64.33
$\texttt{reject}_{M,e\neq}$	M+A	A	A	26.84	92.14	9.8176	64.54
$\texttt{oracle}_{M,e\neq}$	M+A	A	A	17.59	78.64	10.9293	72.87

performances are synthesized in Table X. These tables call for several comments.

First of all, the overall performance we achieved on the ALERTS bitext is much lower than that obtained on the MÉTÉO task. Whereas our MÉTÉO system translated roughly 80% of the sentences perfectly, the best SER measured on the ALERTS task is 41.6% for the SW corpus, and this climbs to 73.3% on the OB corpus. This might be explained by the less repetitive nature of the ALERTS bitext: 73.8% of the SW TEST sentences were seen verbatim in the TRAIN corpus of SW, but only 47.53 % of the sentences in the case of the OB warnings. This is also reflected by a larger vocabulary: once the ALERTS corpus has been meta-tokenized, the TRAIN has above 6,400 words, while the much larger TRAIN corpus of the MÉTÉO bitext has only 3,352 types.

Table X. Overall performance on the ALERTS task

	SW				OB			
	WER	SER	NIST	BLEU	WER	SER	NIST	BLEU
memo	14.92	43.46	7.6759	73.72	33.44	74.96	8.7239	53.71
smt	11.79	48.21	8.4136	76.93	29.12	91.53	9.6647	58.13
reject	11.78	48.98	8.4083	76.81	28.92	91.10	9.7028	58.37
combo	8.99	41.65	8.7448	82.10	20.93	73.38	10.5338	67.71

Second, we observe that, even if the translation memory performances are much worse than the SMT ones insofar as the WER and the precision n-gram metrics are concerned, the memory still produces the best SER. This explains why a combination of the memory and the rescored SMT engine yielded the best performance once again.

It is interesting to note that the MÉTÉO data we had already collected is of little use in translating the ALERTS material. This can be seen for instance by comparing the performance of the memory-based engine when the memory was populated with the MÉTÉO bitext (where the worst SER score was recorded), and when it was populated with the ALERTS one (with an SER of 84.3%). This can also be observed on the SMT experiments (row 2 of Table IX) where we varied the training corpora. Trained on the MÉTÉO material, the SMT engine achieves a BLEU score of 37.8, while a BLEU score of 66.7 is obtained when the training material was from the ALERTS bitext. Actually, we did find some use for the MÉTÉO bitext: by training the phrase-based model on both bitexts (duplicating 10 times the ALERTS material), we were able to improve the BLEU score slightly.

Nevertheless, a small *in-domain* corpus is far more valuable than a huge *out-of-domain* one. Of course, this is not a major discovery, but we have to recall that intuitively, we could have believed that the ALERTS material is quite close to that of MÉTÉO. Similar observations have been made for language modeling in Rosenfeld (2000). The author reports that tens of millions of words of out-of-domain text did not substantially improve the performance of an in-domain model trained on a few million words. We also have similar evidence for the MT case (Vogle et al., 2004).

Lastly, we observe that the rescoring layer slightly improves the SMT performance on the ALERTS task. The native system used for training the rescorer was the one labeled M+A in Table IX (last line of row 2). This improvement is consistent with our findings on the MÉTÉO corpus. Note however, that given the fairly small development set, we trained only a

rejection rescorer, training its neural network without hidden units, which is equivalent to training under a maximum entropy regime (Gandrabur and Foster, 2003). The fact that the rescorer still managed to improve upon the native system with such a small development set is encouraging. However, the improvement does not carry over when we measure the overall performance on the SW corpus (see Table X).

9.3. ERROR ANALYSIS

We applied the same procedure described in Section 8.1 and collected the different errors made by the `combo` system. The errors along with their occurrence counts on the TEST corpora are reported in Table XI.

A tenth of the different errors observed account for half the total differences between the output of the engine and the reference translation. The two most frequent errors arise in sentences such as (15). The first error concerns the translation of **WARNINGS AND WATCHES** into *AVERTISSEMENTS ET VEILLES*, while the reference translation is *ALERTES, AVERTISSEMENTS ET VEILLES*. The second error is actually a correct translation which the system produced for the source sequence **THIS BULLETIN**, while in the context, the human translator opted for *CELUI-CI* 'this one'. Other less frequent errors are often alternative ways of expressing the same concept, such as *PRODUIRE* for *CAUSER* 'cause', or *SE DEPLACE* for *VA* 'goes'.

(15) **A SUMMARY OF ALL WARNINGS AND WATCHES FOR SOUTHERN MANITOBA IS AVAILABLE IN THE OB@@:1 CWWG BULLETIN ISSUED IMMEDIATELY FOLLOWING THIS BULLETIN**
*UN RESUME DE TOUS LES **ALERTES**, AVERTISSEMENTS ET VEILLES POUR LE SUD DU MANITOBA EST DISPONIBLE DANS LE BULLETIN OB@@:1 CWWG EMIS IMMEDIATEMENT APRES **CELUI-CI***

Table XI. The ten most frequent modifications that should be done to transform the translations of the COMBO system into the reference one

Freq.	Transformation	Freq.	Transformation
114	$\phi \rightsquigarrow$ *ALERTES ,*	10	$\phi \rightsquigarrow$ *BULLETIN*
114	*CE BULLETIN* $\rightsquigarrow$ *CELUI-CI*	10	$\phi \rightsquigarrow$ *INTENSE*
56	*DES* $\rightsquigarrow$ *LES*	9	$\phi \rightsquigarrow$ *LE*
16	$\phi \rightsquigarrow$ *DE*	9	*FORMES* $\rightsquigarrow$ *DEVELOPPES*
11	*DE* $\rightsquigarrow$ *D*	9	*VERS LE* $\rightsquigarrow$ *AU*

UN RESUME DE TOUS LES AVERTISSEMENTS ET VEILLES POUR LE SUD DU MANITOBA EST DISPONIBLE DANS LE BULLETIN OB@@:1 CWWG EMIS IMMEDIATEMENT APRES ***CE BULLETIN***
= = = = = D D = = = = = = = = = = = = = = = = = = = S I
φ⇝ALERTES, CE BULLETIN⇝CELUI-CI

(16) **SEVERE THUNDERSTORMS ARE NOT LONGER THREATENING THE ABOVE REGIONS.**
LES *ORAGES VIOLENTS NE MENACENT PLUS LES REGIONS CI-DESSUS.*
DES *ORAGES VIOLENTS NE MENACENT PLUS LES REGIONS CI-DESSUS.*
S = = = = = = = = = =
DES⇝LES

10. Discussion

We have compared various ways of implementing corpus-based approaches for a well-defined real-life task: the translation of weather reports. The advantage of this particular application is that huge amounts of bitexts are available for this domain, and a commercially used rule-based MT system exists for the task.

We observed that a straightforward memory-based approach can already obtain good results, owing to the highly repetitive nature of the weather forecast domain. We found that a phrase-based SMT engine is even better suited to translate previously unseen sentences. We also registered further improvements after applying a rescoring layer. Finally, combining both systems yielded significant overall improvements.

We also examined another possible application of the developed technology to a more challenging task: the translation of weather alerts. Here, however, our approaches were unable to achieve the same level of success without further adaptation. Nevertheless, we did confirm that a combination of translation memory and a statistical phrase-based engine yielded the best performances. The lack of sufficient training data and the less repetitive nature of the material may account for these results.

Since we spent a fairly large amount of time preparing the bitexts we worked on, we thought it would be a good idea to make them freely available to the community. We therefore invite interested persons to consult the web page at **rali.iro.umontreal.ca/meteo**. XML version of both the MÉTÉO and the ALERTS bitexts are available as well as resources for processing them.

Even though we focussed in this study on the applicability of MT technologies for the MÉTÉO task, we must mention that some alternatives to MT have been proposed for weather reports, namely multilingual text generation directly from raw weather data: temperatures, winds, pressures etc. These generation systems also require that humans select templates to organize the report. Generating text in many languages from one source is quite appealing from a conceptual point of view and has been cited as one of the potential applications for natural language generation (Reiter and Dale, 2000); some systems have been developed (Kittredge et al., 1986; Goldberg et al., 1994) and tested in operational contexts. But thus far, none has been used in everyday production to the same level as that achieved by MT. One of the reasons in this domain is that meteorologists still prefer to write up their reports in natural language rather than selecting text structure templates.

Acknowledgements

We thank Rick Jones and Marc Besner from the Dorval office of Environment Canada who provided us with the corpus of bilingual weather reports. We are indebted to Elliot Macklovitch who provided us with articles describing the MÉTÉO system that we could not have found otherwise. We wish to thank the anonymous reviewers for their insightful comments and suggestions. We are also grateful to Fabrizio Gotti for his technical assistance. This work has been financially supported by grants from NSERC and FQRNT.

Notes

[1] The current reports are available on the web at **http://meteo.ec.gc.ca/forecast/textforecastf.html**. This site is continually being updated.

[2] All text appears in upper case in the weather bulletins, in the case of French, unaccented; where appropriate, accents are omitted in the text here.

[3] Actually, for the sake of readability, we report on 100×BLEU in this text.

[4] We used the version **v11a** of the script which we downloaded at **www.nist.gov/speech/tests/mt/resources/scoring.htm**

[5] We used the alignments produced by IBM model 2.

[6] We experimented with all the heuristics without noticing any significant impact on performance.

[7] This meant extending the number of different symbols that could be aligned by ClustalW and modifying the cost matrix.

[8] The sentences being considered are without meta-tokens.

References

Akiba, Y., M. Federico, N. Kando, H. Nakaiwa, M. Paul and J. Tsujii: 2004, 'Overview of the IWSLT04 Evaluation Campaign'. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, pp. 1–12.

Bangalore, S., G. Bordel and G. Riccardi: 2001, 'Computing consensus translation from multiple Machine Translation systems'. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Italy, pp. 351–354.

Bangalore, S., V. Murdock and G. Riccardi: 2002, 'Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system'. In *COLING 2002: Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 50–56.

Bender, O. R. Zens, E. Matusov and H. Ney: 2004, 'Alignment templates: the RWTH SMT system'. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2004)*, Kyoto, Japan, pp. 79–84.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis and N. Ueffing: 2004, 'Confidence estimation for Machine Translation'. In *The 20th International Conference on Computational Linguistics, COLING 2004*, Geneva, Switzerland, pp. 315–321.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer: 1993, 'The mathematics of statistical Machine Translation: Parameter estimation'. *Computational Linguistics* **19**, 263–311.

Carl, M. and A. Way: 2003, *Recent Advances in Example-Based Machine Translation*, Dordrecht, The Netherlands: Kluwer.

Chandioux, J.: 1988, 'Meteo: An operational translation system'. In *Proceedings of the 2nd Conference on RIAO*, Cambridge, Massachusetts, pp. 829–839.

Coch, J.: 1998, 'Interactive generation and knowledge administration in MultiMeteo'. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, Niagara-on-the-lake, Ontario, Canada, pp. 300–303.

Collobert, R., S. Bengio and J. Mariéthoz: 2002, *Torch: A Modular Machine Learning Software Library*, Technical report IDIAP-RR 02-46, IDIAP, Martigny, Switzerland.

Doddington, G.: 2002, 'Automatic evaluation of Machine Translation quality using n-gram co-occurrence statistics'. In *Proceedings of the Second International Conference on Human Language Technology Research*, San Diego, CA, pp. 138–145.

Frederking, R. and S. Nirenburg" 1994, 'Three heads are better than one'. In *Fourth Conference on Natural Language Processing*, Stuttgart, Germany, pp. 95–100.

Frederking, R. E. and K. B. Taylor: 2004, (eds) *Machine Translation: From Real Users to Research. 6th Conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, ...* . Berlin: Springer Verlag.

Gandrabur, S. and G. Foster: 2003, 'Confidence estimation for text prediction'. In *Conference on Computational Natural Language Learning (CoNLL-2003)*, Edmonton, Canada, pp. 315–321.

Goldberg, E., N. Driedger and R. Kittredge: 1994, 'Using natural language processing to produce weather forecasts'. *IEEE Expert* **9**, 2:45–53.

Grimaila, A. and J. Chandioux: 1992, 'Made to measure solutions'. In J. Newton (ed.), *Computers in Translation: A Practical Appraisal*, London: Routledge, London, pp. 33–45.

Hutchins, W. J.: 1986, *Machine Translation: Past, Present, Future*. Chichester, Sussex: Ellis Horwood.

Hutchins, W. J. and H. L. Somers: 1992, *An Introduction to Machine Translation*, London: Academic Press.

Isabelle, P.: 1987, 'Machine Translation at the TAUM Group'. In M. King (ed.) *Machine Translation Today: The State of the Art*, Edinburgh: Edinburgh University Press, pp. 247–277.

Jayaraman, S. and A. Lavie: 2005, 'Multi-engine Machine Translation guided by explicit word matching'. In *European Association for Machine Translation (EAMT) 10th Annual Conference*, Budapest, Hungary, pp. 143–152.

Kittredge, R., A. Polguère and E. Goldberg: 1986, 'Synthesizing weather reports from formatted data'. In *Coling '86: Proceedings of the 11th International Conference on Computational Linguistics*, Bonn, West Germany, pp. 563–565.

Knight. K. and Y. Al-Onaizan: 1999, 'A Primer on Finite-State Software for Natural Language Processing', unpublished report, August 1999, available at `www.isi.edu/licensed-sw/carmel/carmel-tutorial2.pdf`. Last accessed 11 August 2005.

Koehn, P.: 2004, 'Pharaoh: a beam search decoder for phrase-based SMT'. In Frederking and Taylor (2004), pp. 115–124.

Koehn, P., F.-J. Och and D. Marcu: 2003, 'Statistical phrase-based translation'. In *HLT-NAACL 2003, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, pp. 127–133.

Langlais, P.: 1997, 'Alignement de corpus bilingues: intérêts, algorithmes et évaluations'. [Bilingual corpus alignment: points of interest, algorithms and evaluation], in *Actes du colloque international FRACTAL 1997, Linguistique et Informatique : Théories et Outils pour le Traitement Automatique des Langues*, Besançon, France, pp. 245–254.

Leplus, T.: 2004, *Étude de la traduction automatique des bulletins météorologiques* [A study on machine translation of weather reports], Master's thesis, Université de Montréal.

Leplus, T., P. Langlais and G. Lapalme: 2004, 'Weather report translation using a Translation Memory'. In Taylor (2004), pp. 154–163.

Leusch, G., N. Ueffing and H. Ney: 2003, 'A novel string-to-string measure with applications to Machine Translation'. In *MT Summit IX: Proceedings of the Ninth Machine Translation Summit*, New Orleans, LA, pp. 240–247.

Macklovitch, E.: 1985, *A Linguistic Performance Evaluation of METEO 2*, Technical report, Canadian Translation Bureau, Montreal, Canada.

Och, F.-J., D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin and D. Radev: 2004, 'A smorgasbord of features for Statistical Machine Translation'. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, MA, pp. 161–168.

Och, F.-J. and H. Ney: 2000, 'Improved statistical alignment models'. In *38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, pp. 440–447.

Papineni, K., S. Roukos, T. Ward and W-J. Zhu: 2002, 'BLEU: A method for automatic evaluation of Machine Translation'. In *40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia PA, pp. 311–318.

Reiter, E. and R. Dale: 2000, *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

Rosenfeld, R.: 2000, 'Two decades of statistical language modeling: Where do we go from here?' *Proceedings of the IEEE* **88**, 1270–1278.

Stolcke, A.: 'SRILM – An extensible language modeling toolkit'. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP-2002*, Denver, Colorado, pp. 901–904.

Thompson, J. D., D. G. Higgins and T. J. Gibson: 1994, 'ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice'. *Nucleic Acids Research* **22**, 4673–4680.

van Zaanen, M. and H. Somers: 2005, 'DEMOCRAT: Deciding between Multiple Outputs Created by Automatic Translation'. In *MT Summit X: The Tenth Machine Translation Summit*, Phuket, Thailand, pp. 173–180.

Vogel, S., S. Hewavitharana, M. Kolss and A. Waibel: 2004, 'The ISL statistical translation system for spoken language translation'. In *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, pp. 65–72.

Wahlster, W. (ed.): 2000, *Verbmobil: Foundations of Speech-to-speech Translation*. Berlin: Springer Verlag.

Zens, R. and H. Ney: 2004, 'Improvements in phrase-based Statistical Machine Translation'. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Boston, MA, pp. 257–264.

Zhao, B., S. Vogel and A. Waibel: 2004, 'Phrase pair rescoring with term weightings for Statistical Machine Translation'. In *2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain, pp. 206–213.