The 'purest' EBMT system ever built: no variables, no templates, no training, examples, just examples, only examples

Yves Lepage
yves.lepage@atr.jp

Etienne Denoual
etienne.denoual@atr.jp

ATR – Spoken language communication labs
Keihanna gakken tosi, 619-0288 Kyoto, Japan

Abstract

We designed, implemented and assessed an EBMT system that can be dubbed the "purest ever built": it strictly does not make any use of variables, templates or training, does not have any explicit transfer component, and does not require any preprocessing of the aligned examples. It uses a specific operation, namely proportional analogy, that implicitly neutralises divergences between languages and captures lexical and syntactical variations along the paradigmatic and syntagmatic axes without explicitly decomposing sentences into fragments. In an experiment with a test set of 510 input sentences and an unprocessed corpus of almost $160,000$ aligned sentences in Japanese and English, we obtained BLEU, NIST and mWER scores of 0.53, 8.53 and 0.39 respectively, well above a baseline simulating a translation memory.

1 Introduction

In contrast to some "least effort" approaches to machine translation, which do not view linguistic data as specific data, we claim that natural language tasks are specific because their data are specific. The goal of this paper is to show that the use of a specific operation, namely proportional analogy in our present proposal, is profitable in terms of trading off preprocessing time of the data and quality of the results. Our proposed technique does not require any preprocessing of the data whatsoever, a definite advantage over techniques that require intensive preprocessing.

1.1 Dealing with the specificity of linguistic data

Trivially, any linguistic datum belongs to one specific natural language that constitutes a "system" in the Saussurian sense of the term. A consistent consequence is to process linguistic data using operations that specifically capture this systematicity. This systematicity appears at best in commutations exhibited by proportional analogies like in the following example.

$$\frac{\textit{I'd like to open these windows.}}{\textit{Could you open a window?}} :: \frac{\textit{I'd like to cash these traveler's checks.}}{\textit{Could you cash a traveler's check?}}$$

Such commutations make paradigmatic and syntagmatic variations explicit and allow for lexical and syntactical variations that ought to be exploited by machine translation system to express different meanings. Indeed, each sentence in any language can be cast into a wide number of such proportional analogies that form a kind of meshwork around it. In (LEPAGE and PERALTA, 2004) we have shown how to automatically extract tables (or matrices) from a linguistic resource so as to visualize these meshworks: each cell in a table contains a sentence, and rectangles formed with four cells in the tables are proportional analogies.

1.2 Dealing with divergences across languages

Machine translation has specific problems to address: one of them, at the core of translation, is to tackle divergences across languages.

A classical and simple example of divergence is the exchange of the arguments of a predicate in Vauquois's famous example between English and French:

$$Elle_1\ lui_2\ plaît.\qquad \leftrightarrow\qquad He_2\ likes\ her_1.$$

To confirm the importance of the phenomenon, (HABASH, 2002) quotes a study on a sample of $19,000$ sentences between English and Spanish that shows that one sentence in three presents divergences that can be classified into five different types. An example of type 4 is the classical translation of a Spanish verb into an English preposition.

1: $Atraves\acute{o}_{\text{V}}$		0: It
2: $el\ r\acute{i}o_{\text{N}}$	$\leftrightarrow$	3: $floated_{\text{V}}$
3: $flotando_{\text{particip.}}$		1: $across_{\text{prep.}}$
		2: $the\ river_{\text{N}}$

Approaches that rely on the word as the unit of processing forget the fact that corresponding pieces of information in different languages are indeed distributed over the entire strings and do not necessarily correspond to complete words. For this reason, the correspondence between words given in the example above is in fact not detailed enough. Actually, the ending *-ó* of the first Spanish word accounts for 3rd person singular past tense. So, not only does *atravesó* correspond to the English preposition *across* for its meaning, but, in addition, it also corresponds to another complete word in English (the pronoun *it*), plus a portion of yet a third English word (the final ending *-ed* of *floated*).

1.3 Dealing with structures (meshworks of proportional analogies)

Following the previous idea that a sentence belongs to a meshwork of proportional analogies, any particular translation correspondence between two sentences belonging to two different languages should be viewed as a part of the global correspondence between the two languages at hand. The technique that we thus propose for automatic translation exploits the translation links that incidentally exist between sentences as part of the meshwork of proportional analogies found around them.

I'd like to open these windows.	*Could you open a window?*	*I'd like to cash these traveler's checks.*	*Could you cash a traveler's check?*
$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
Est-ce que ces fenêtres, là, je peux les ouvrir?	*Est-ce que vous pouvez m'ouvrir une fenêtre?*	*Ces chèques de voyage, là, je peux les échanger?*	*Vous pouvez m'échanger un chèque de voyage?*

Figure 1: Two proportional analogies in two different languages that correspond.

Figure 1 gives the example of the two following sentences taken as part of particular proportional analogies that correspond.

$$Could\ you\ cash\ a\ traveler's\ check? \leftrightarrow Vous\ pouvez\ m'\acute{e}changer\ un\ ch\grave{e}que\ de\ voyage?$$

The correspondence can only be established because each sentence in the lower part of the figure is a possible translation of the sentence above it in the upper part of the figure.

A consequence of this view is that the difficulty which is usually seen in translating between some particular pairs of languages simply vanishes. The claim that it is costly to translate between some specific language pairs like, *e.g.*, Japanese and English, relies indeed on the idea that translating would basically consist of rearranging, transforming, or decoding. However, to make a comparison with clothes, to localise what corresponds to the left shoulder of a shirt on, say, a jacket, one does not take material from the left shoulder of the shirt, unweave it, weave it back again in a different way, and then patch it somewhere on the jacket. Although this sounds strange, this is precisely what second generation MT systems do when they use lexical and structural transfer rules; and SMT systems (BROWN et al., 1993) when they use lexicon models with distortion models.

Rather, it is reasonable to point at the left shoulder of the jacket by looking at the gen-

eral constitution of the jacket, and by following the different wooves and threads *on* the jacket to localise some point more precisely if needed, as the jacket is made of a different material from the shirt. Transposing to machine translation, the translation of a source sentence should be looked for by relying on the paradigmatic and syntagmatic meshworks, *i.e.*, by using the proportional analogies in the target language which correspond to the proportional analogies of the source language that involve the source sentence, until a corresponding sentence is obtained.

2 Example-based machine translation (EBMT) by proportional analogy

2.1 The algorithm

Suppose we have a corpus of aligned sentences in two languages (a bicorpus) at our disposal. The following gives the basic outline of our method to perform the translation of an input sentence:

Form all analogical equations with the input **sentence** D and with all relevant pairs of **sentences** (A_i, B_i) from the source part of the bicorpus[1];

$$A_i : B_i :: x : D$$

For those sentences that are solutions of the previous analogical equations which do not belong to the bicorpus, translate them using the present method recursively. Add them with their newly generated translations to the bicorpus;

For those sentences $x = C_{i,j}$ that are solutions of the previous analogical equations[2] which belong to the bicorpus, do the following;

---

[1]Relevant pairs of sentences are selected on-the-fly according to a similarity criterion. A_i,B_i and D are **sentences**; they are **not fragments** of sentences. Sentences are **not cut into pieces**. Also, **pairs** of sentences are retrieved to form an analogical equation with D; consequently, there is no such thing as **analogous examples**, as such an expression does not make any sense in this framework; indeed, A_i's and B_i's may be quite "far away" from D.

[2]One analogical equation may yield several solutions.

Form all analogical equations with all possible target language sentences corresponding to the source language sentences[3];

$$\widehat{A_i}^k : \widehat{B_i}^k :: \widehat{C_{i,j}}^k : y$$

Output the solutions $y = \widehat{D_{i,j}}^k$ of the analogical equations as a translation of D, sorted by frequencies[4].

2.2 An example

Suppose that we wanted to translate the following Japanese input sentence:

濃いコーヒーが飲みたい。[5]

Among all possible pairs of sentences from the bicorpus, we may find the following two Japanese sentences:

| 紅茶をください。 | $\leftrightarrow$ | *May I have some tea, please?* |
| コーヒーをください。 | $\leftrightarrow$ | *May I have a cup of coffee?* |

that will allow us to form the following analogical equation:

$$\text{紅茶をください。} : \text{コーヒーをください。} :: x : \text{濃いコーヒーが飲みたい。}$$

This equation yields $x = $ 濃い紅茶が飲みたい。[6] as a solution. If this sentence already belongs to the bicorpus, *i.e.*, if the following translation pair is found in the data

| 濃い紅茶が飲みたい。 | $\leftrightarrow$ | *I'd like some strong tea, please.* |

the following analogical equation is formed with the corresponding English translations:

May I have some tea, please? : *May I have a cup of coffee?* :: *I'd like some strong tea, please.* : x

By construction, the solution: $x = $ *I'd like a cup of strong coffee.* is a candidate translation of the input sentence: 濃いコーヒーが飲みたい。

---

[3]Several target sentences may correspond to the same source sentence.

[4]Different analogical equations may yield identical solutions.

[5]Gloss: strong coffee NOMINATIVE-PARTICLE drink-VOLITIVE. Literally: *I want to drink strong coffee.*

[6]Lit.: *I want to drink strong tea.*

May I have some tea, please? ·········· ↔ ·········· 紅茶をください。

I'd like some strong tea, please. ·········· ↔ ·········· 濃い紅茶が飲みたい。

May I have a cup of coffee? ·········· ↔ ·········· コーヒーをください。

I'd like a cup of strong coffee. ·········· ↔ ·········· 濃いコーヒーが飲みたい。
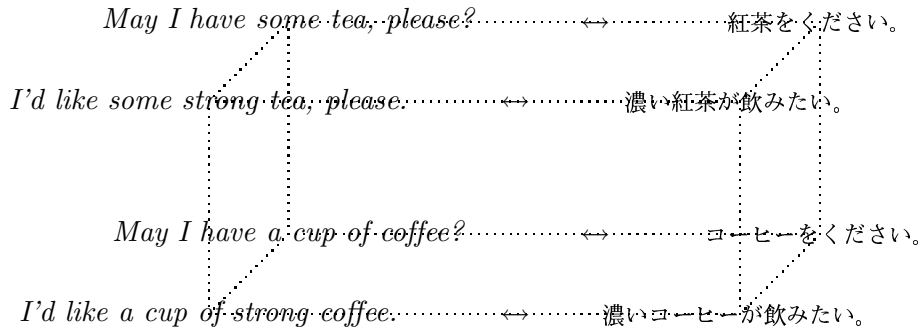
Figure 2: The parallelopiped: in each language, four sentences form a proportional analogy. There exist four translation relations between the sentences.

2.3 A geometric view of the principle

The processing of the previous example, which is reminiscent of distributionalism (HARRIS, 1954), can be viewed in the shape of a parallelopiped as shown in Figure 2. The left plane of this parallelopiped is the plane of the English analogy. The right plane is the Japanese one. Because each of these planes resides in one and only one language, the terms of the proportional analogy involve monolingual data only so that they can be processed by algorithms like the one proposed in (LEPAGE, 1998).

3 Features of the method

3.1 No transfer

To stress that the choice of a correct translation is really left to an implicit use of the structure of the target language, and does not imply any explicit transfer processing, consider the Spanish example of Section 1.2 again. The correspondences between the source and the target language in a proportional analogy will be entirely responsible not only for the selection of the correct lemmas wit their lexical POS, but also for the correct word order[7].

This could be compared to some extent to the translation of the adnominal particle N_1 *no* N_2 from Japanese into English in (SUMITA

and IIDA, 1991) where the choice of the correct preposition (or word order) is left to the list of examples.

$$
\begin{array}{ccccc}
\textit{They} & \textit{They} & & \textit{It floated} & \textit{It floated} \\
\textit{swam in} : & \textit{swam} & :: & \textit{in the} : & \textit{across} \\
\textit{the sea.} & \textit{across} & & \textit{sea.} & \textit{the river.} \\
& \textit{the river.} & & & \\
\updownarrow & \updownarrow & & \updownarrow & \updownarrow \\
\textit{Nadaron} & \textit{Atravesa-} & & & \\
\textit{en el} : & \textit{ron el rio} :: & \textit{Flotó en} : & x \\
\textit{mar.} & \textit{nadando} & & \textit{el mar.} &
\end{array}
$$

However, it should be stressed that in proportional analogies like the two above, nowhere is it said which word corresponds to which word, or which syntactic structure corresponds to which syntactic structure. The sole action of proportional analogy with (necessarily) **the character as the only unit of processing**, is sufficient to produce the exact translation of *It floated across the river*, that is, the correct Spanish sentence: $x = Atravesó\ el\ rio\ flotando$, provided that the three sentence pairs on the left are valid translation pairs.

3.1.1 No extraction of symbolic knowledge

In a second generation MT system, one makes the knowledge relevant to such divergences explicit in the form of lexical and structural transfer rules. In the EBMT approach too, one makes this knowledge explicit by automatically acquiring templates that capture these divergences. In both cases, the knowledge about these divergences has to be made explicit. In

---

[7]As for reordering of words, with its translation knowledge reduced to the sole two translation pairs: $abc \leftrightarrow abc$, $abcabc \leftrightarrow aabbcc$, the system needs only to solve $2 \times (n-2)$ proportional analogies recursively to translate members of the regular language $\{ (abc)^n \mid n \in \mathbb{N}^* \}$ into the corresponding members of the context-sensitive language $\{ a^n b^n c^n \mid n \in \mathbb{N}^* \}$, and reciprocally: $(abc)^n \leftrightarrow a^n b^n c^n$.

our view the choice of the correct expression ought to be left implicit as it pertains to the structure of the target language. Indeed, paradigmatic and syntagmatic commutations neutralise these divergences as they are the implicit constitutive material of proportional analogies.

Our system definitely positions itself in the EBMT stream, however it departs from it in one important aspect: it does not make any use of explicit symbolic knowledge such as templates with variables. Direct use of bicorpus data in their raw form is made, without any preprocessing.

The reason for doing so is that we consider that templates may well be insufficient in representing all of the implicit knowledge contained in examples. Indeed, variables in templates allow for paradigmatic variations at some predefined positions only[8]. For instance, extracting the template *X salts Y* from the example sentence *the butcher salts the slice* where X may be replaced by *the butcher*, etc. and Y by *the slice*, etc. [9] does not make the most of the potential of the example. Firstly, it prevents *the butcher* from being changed into a plural: *the butchers*. Moreover, it misses the fact that *salts* may also commute with its past and future forms, etc.: *salted*, *will salt*, etc., or with *cuts*, *smokes*, etc.; and so forth. To summarise, there is a risk of loss of information when replacing examples with templates.

The situation is in no way better with translation patterns. They make explicit which variables in the source have to be replaced by which variables in the target[10]. But it is well known that a single variable at one single position in a source template often needs to be linked to several positions distributed over a target template, and may even imply different levels of description (morphological, syntactical, etc.) For instance, negation is expressed at one single position in Japanese, whereas it may also imply a change in the form of the main verb in English: *he eats → he does not eat.*

Our view is that *every* position in a language datum is subject to paradigmatic variation[11]. The consequence being that a lot more exploitable information is to be found in unprocessed examples than in templates. And it may well be the case that the number of templates necessary to encode the same amount of information contained in a set of examples is much larger in size than the actual size of these unprocessed examples themselves. Thus, extracting templates from examples may well entail a loss in generative power as well as in space. It must however be stressed that the generative power of the unprocessed examples does not actually reside in their bare listing but in their capacity for getting involved in proportional analogies.

3.2 No training, no preprocessing

As a consequence of the abovementioned features, there is no such thing as a training phase or a preprocessing phase in our system: the bicorpus is just loaded into memory at program startup. No language model is computed; no other alignment than the one given by the bicorpus is extracted; no segmentation or tagging whatsoever is performed. Needless to say, the possibility of adding new information to the bicorpus is left open. For instance, adding dictionaries or paraphrases to the corpus is a possibility that may improve results but leaves the structure of the system absolutely unchanged (see Sections 4.4.2 and 4.4.3).

4 Evaluation and comparison with other systems

4.1 Resources used in the evaluation

To assess the performance of the proposed method, we used the C-STAR Basic Traveler's Expressions Corpus[12]. It is a multilingual resource of expressions from the travel and tourism domain that contains almost 160,000 aligned translations in English and Japanese. In this resource, the sentences are quite short as the figures in the following table show. As the same sentence may appear several times with different translations, the number of different

---

[8] In (SATO, 1991), so as to acquire a grammar, sentences are fed into a system, which differ by one word only.

[9] Examples from (CARL, 1998).

[10] (SASAYAMA et al., 2003) for the use of arrays describing these kinds of associations.

[11] Putting it to the extreme, even phonetic variations have to be considered: *wolf : wolves :: leaf : leaves.* So that one definitely has to go below words. For this reason, our system processes **strings of characters,** not strings of words.

[12] http://www.c-star.org/.

	コーヒーのおかわりをいただけますか。
2318	*I'd like another cup of coffee.*
2296	*May I have another cup of coffee?*
1993	*Another coffee, please.*
1982	*May I trouble you for another cup of coffee?*
1982	*Can I get some more coffee?*
530	*Another cup of coffee, please.*
516	*Another cup of coffee.*
466	*Can I have another cup of coffee?*
337	*May I get some more coffee?*
205	*May I trouble you for another cup of coffee, please?*

	小銭をまぜてください。
924	*Can you include some small change?*
922	*Can you include some small change, please?*
899	*Would you include some small change?*
896	*Include some small change, please.*
895	*I'd like to have smaller bills mixed in.*
895	*Please change this into small money.*
895	*Will you include some small change?*
885	*Could you include some small change, please?*
880	*May I have some small change, too?*

Figure 3: Two examples of translations. The figures on the left are the frequencies with which each translation candidate has been output.

sentences in each language is indicated in the following table.

	Number of ≠ sentences	Size in characters avg. ± std. dev.
English	97,395	35.17 ± 18.83
Japanese	103,051	16.22 ± 7.84

The method relies on the assumption that analogies of form are almost always analogies of meaning. Thus, prior to its application, we (LEPAGE, 2004) estimated the relative number of analogies of form which are not analogies of meaning in the resource used: less than 4% (p-value = 0.1% on a sample of 666 analogies). This proportion is too small to seriously endanger the quality of the results obtained during translation.

4.2 Gold Stantard and baseline

In order to evaluate the performance of our system, we use a test set of 510 input sentences. These sentences are from the same domain as the bicorpus. For each of them, we also have a set of 16 translation references in the target language at our disposal.

This allows us to perform an evaluation using several standard objective measures, like BLEU, NIST or mWER.

Firstly, we determined a Gold Standard in the following way. For each sentence of the test set, we evaluated the first reference translation as if it were given by an MT system. In this way, we obtained the "best" values for each of the measures considered (see Table 1).

Then, we determined a baseline by simulating a translation memory. For each sentence of the test set, we took the closest sentence in the corpus according to edit distance and output its translation that we evaluated with each of the objective measures. This gives baseline scores for each of the measures considered.

4.3 Results with the resource only

Our system was then evaluated on the translations it output for the sentences of the test set, with the sole source of examples being the resource data (see Table 1, line: resource only). Some examples of translations are shown in Figure 3, with the frequencies for each candidate[13]. As we assumed that the most frequent candidate should be the most reliable one, the evaluation was performed on the first candidates only.

4.4 Choice and influence of linguistic resources

4.4.1 Influence of the amount of examples

In an EBMT system, one would trivially expect the amount and nature of examples to strongly influence translation quality. The figures in Table 1 on the lines marked 1/2 resource and 1/4 resource, which were obtained by sampling the original resource confirm this fact. In this case, the more data, the better the results.

---

[13]Different analogical equations may yield the same solutions (see Section 2.1).

Table 1: Scores for the Gold Standard, the baseline, and the system with various data. We also compare with two other EBMT systems that require heavy preprocessing of the bicorpus to extract patterns either automatically (system A) or by hand (system B).

System:	Number of translation pairs	BLEU	NIST	mWER	PER	GTM
Gold Standard	n.r.	1.00	14.95	0.00	0.00	0.91
System A	unknown	0.66	10.36			
+ Src + tgt paraphrases	438,817	0.50	**8.98**	0.46	0.42	0.67
+ Tgt paraphrases	158,409	0.49	8.91	0.47	0.43	0.67
+ Src paraphrases	158,409	0.53	8.53	**0.38**	**0.35**	**0.68**
+ Dictionary	206,382	**0.54**	8.54	0.39	0.36	**0.68**
Resource only	158,409	0.53	8.53	0.39	0.36	**0.68**
1/2 resource	81,058	0.45	7.78	0.50	0.45	0.63
1/4 resource	40,580	0.42	7.18	0.53	0.49	0.60
System B	unknown	0.41	9.00			
Baseline: transl. memory	n.r.	0.38	7.54	0.58	0.53	0.61

4.4.2 Dictionaries as lists of particular examples

Whole sentences contained in the resource (as opposed to isolated words or idioms) may not allow the translation of particular expressions if commutations cannot be found between them. This case is particularly plausible when translating sentences that contain multi-word expressions or numbers, for instance.

A possible remedy is to add dictionary entries to the original resource to be used as additional examples. As a matter of fact, this system does not make any difference between a bicorpus or a dictionary as long as both are aligned strings of data, be they sentences or words. The following examples illustrate that the data format for a bicorpus or a dictionary does not differ in any way.

フィルムを買いたいのですが。 ↔ *I'd like a film, please.*

三十六枚撮りを二本ください。 ↔ *Two rolls of thirty-six exposure film, please.*

このカメラの電池がほしいのです。 ↔ *I'd like a battery for this camera, please.*

フィルム ↔ *film*
映画 ↔ *film*
電池 ↔ *battery*
砲台 ↔ *battery*

The scores obtained by adding a dictionary to our resource are not different from those with the resource only, except for a slight improvement in BLEU.

4.4.3 Paraphrases generated from the resource as additional examples

Previous research has shown that the introduction of paraphrases may improve the quality of machine translation output. Paraphrases may be added in the source language (YAMAMOTO, 2004) or in the target language (HABASH, 2002).

In order to increase the chances of a sentence entering into proportional analogies, we grouped sentences in the source language data by paraphrases. To do so, we grouped sentences that share at least one common translation because, in this case, they share the same meaning, (*i.e.*, they are paraphrases). In our bicor-

pus, an average of 3.03 paraphrases per source sentence was obtained[14]. This new information allows the translation process to test a larger number of proportional analogies. When a pair of sentences (A, B) is proposed for an input sentence D, not only the equation $A : B :: x : D$ will be tried, but also all possible equations of the form $A' : B' :: x : D$, where A' and B' are paraphrases of A and B.

The evaluation of translation quality when adding paraphrases in the source language are shown in Table 1 on the line marked: + Src paraph. They show a slight improvement in word error rate.

The same thing can be done on the target language side with a similar effect of increasing the number of proportional analogies tried, this time in the target language. As for scores, they decrease in BLEU but show a real improvement in NIST.

The scores obtained when adding paraphrases in the source and in the target language are shown on the line marked: + Src + tgt paraph. They are not better than those with the resource only, except for NIST, as paraphrases are expected to have introduced lexical and syntactical variation in expressing identical meanings. An explanation for the loss in quality according to all other measures may be that the increase in computation to perform may have overloaded the system (all experiments are done with the same time-out).

5   Discussion and future work

5.1   Translation time

It could have been feared that the complexity of the algorithm, which is basically square in the amount of data, would have enormously impaired the method. However, using a simple heuristics to select only relevant pairs entering in analogical equations allowed us to keep translation times reasonable. Within a time-out of 1 CPU second, the average translation time per sentence was 0.73 second on a 2.8 GHz processor machine with 4 Gb memory.

---

[14]However, the distribution is not uniform: 71,192 sentences (out of 103,274) don't get any new paraphrase, while 54 sentences get more than 100 paraphrases, with a maximum of 410 paraphrases for one sentence.

5.2   Proportion of successful analogies

As the fundamental operation in the system is analogy, we measured the proportion of analogical equations sucessfully solved over the total number of analogies formed in the source language. Between half a million and one million analogical equations $(687, 641)$ are formed on average to translate one sentence from the test set. The proportion of analogical equations sucessfully solved is 28%. In other words, the heuristics used to select sentence pairs from the corpus in order to form analogical equations is successful only a quarter of the time. Future work should include finding a heuristics that would increase this proportion so as to reduce the number of unnecessary trials.

5.3   Recursion level needed

As was explained in Section 2.1, recursive applications are expected to be made in order to reach translations of a single input sentence. Over all input sentences of the test set, one recursive call is needed on average, and a maximum of two is necessary on some sentences. This shows that the sentences in the test set were in fact quite "close" to the resource used: the number of recursive calls is a measure of how "far" a sentence is to a corpus.

5.4   Relevance / suitability of the examples

The translation of an input sentence depends crucially on the two following points. Firstly, whether the input sentence belongs to the domain (and the style) of the corpus of examples. Secondly, whether the corpus covers the linguistic phenomena present in the input sentence. A positive point of our system is that the absence of any training phase reduces the development cycle to the problem of choosing / coining suitable examples that cover a given domain and the linguistic phenomena of the language. To address these two issues, we see two possible directions of research.

Firstly, as was mentioned in Sections 4.4.3 and 4.4.2, we are studying various ways to add paraphrases or dictionaries and how to improve their efficiency in terms of lexical and syntactical variation, so as to further densify the bicorpus in terms of coverage

Secondly, we are investigating the possibility of designing a core grammar by examples, *i.e.*,

a collection of examples that would cover the basic linguistic phenomena in a given language. In the same way as school grammars illustrate rules by examples, our methodology will be to choose a formal grammar known to have a large coverage, and to illustrate its rules with examples. Distributionalist grammars (HARRIS, 1982) seem to be better candidates for this purpose as they rely on the notion of the expansion and embedding of strings, a notion that is precisely captured by proportional analogy. In particular, *string grammars* (SAGER, 1981) or (SALKOFF, 1973) are well known for having a large coverage.

6   Conclusion

In this paper, we have shown that the use of a specific operation, namely proportional analogy, leads to reasonable results in machine translation without any preprocessing of the data whatsoever, an advantage over techniques requiring intensive preprocessing. In an experiment with a test set of 510 input sentences and an unprocessed corpus of almost $160,000$ aligned sentences in Japanese and English, we obtained BLEU, NIST and mWER scores of 0.53, 8.53 and 0.39, respectively, well above a baseline simulating a translation memory. Slight improvements could be obtained by adding paraphrases.

The use of an operation that suits by essence the specific nature of linguistic data, *i.e.*, their capacity of commutation on the paradigmatic and syntagmatic axes, allowed us to dispense with any preprocessing of the data whatsoever. In addition, this operation has the advantage of tackling the issue of divergences between languages in an elegant way: it neutralises them implicitly. As a consequence, the system implemented does not include any transfer component (either lexical or structural).

To summarise, we designed, implemented and assessed an EBMT system that, we think, can be dubbed the "purest ever built" as it strictly does not make any use of variables, templates or training, does not have any explicit transfer component, and does not require any preprocessing of the aligned examples, a knowledge that is, of course, indispensable.

As an extra feature, the system is learning as it keeps translating. Recursive calls add trans-

lation knowledge to the bicorpus, so that, in standard use, the history of translations will influence the results of coming translations. In the reported experiment we had to disallow this feature to be placed in conditions comparable with, say, SMT systems. However, such a use denatures our system.

7   Acknowledgements

References

Peter E. BROWN, Vincent J. DELLA PIETRA, Stephen A. DELLA PIETRA, and Robert L. MERCER. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, Special Issue on Using Large Corpora: II*, 19(2):263–311, March.

Michael CARL. 1998. A constructivist approach to machine translation. In *Proceedings of NeMLaP'98*, Sydney.

Nizar HABASH. 2002. Generation-heavy hybrid machine translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02)*, New York.

Zellig HARRIS. 1954. Distributional structure. *Word*, 10:146–162.

Zellig HARRIS. 1982. *A grammar of English on mathematical principles*. John Wiley & Sons, New York.

Yves LEPAGE and Guilhem PERALTA. 2004. Using paradigm tables to generate new utterances similar to those existing in linguistic resources. In *Proceedings of LREC-2004*, volume 1, pages 243–246, Lisbonne, May.

Yves LEPAGE. 1998. Solving analogies on words: an algorithm. In *Proceedings of COLING-ACL'98*, volume I, pages 728–735, Montréal, August.

Yves LEPAGE. 2004. Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus. In *Proceedings of COLING-2004*, volume 1, pages 736–742, Genève, August.

Naomi SAGER. 1981. *Natural language information processing: a computer grammar of English and its applications.* Adelson-Wesley, Reading, Massachusetts.

Morris SALKOFF. 1973. *Une grammaire en chaîne du français.* Dunod, Paris.

Manabu SASAYAMA, Fuji REN, and Shigo KUROIWA. 2003. Super-function based Japanese-English machine translation system. In *Proceedings of Natural Language Processing and Knowledge Engineering*, volume 1, pages 555–560, Beijing, October.

Satoshi SATO. 1991. *Example-based Machine Translation.* Ph.d. thesis, Kyoto University, September.

Eiichiro SUMITA and Hitoshi IIDA. 1991. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th Conference on Association for Computational Linguistics*, pages 185–192, Morristown, NJ, USA. Association for Computational Linguistics.

Kazuhide YAMAMOTO. 2004. Interaction between paraphraser and transfer for spoken language translation. *Journal of Natural Language Processing*, 11(5):63–86, October.