Mach Translat (2005) 19:251–282
DOI 10.1007/s10590-006-9010-x

ORIGINAL PAPER

Purest ever example-based machine translation: Detailed presentation and assessment

Yves Lepage · Etienne Denoual

Received: 16 December 2005 / Accepted: 25 August 2006 /
Published online: 19 December 2006

Abstract We have 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 patterns, does not have any explicit transfer component, and does not require any preprocessing or training of the aligned examples. It uses only a specific operation, proportional analogy, that implicitly neutralizes divergences between languages and captures lexical and syntactic variations along the paradigmatic and syntagmatic axes without explicitly decomposing sentences into fragments. Exactly the same genuine implementation of such a core engine was evaluated on different tasks and language pairs. To begin with, we compared our system on two tasks of a previous MT evaluation campaign to rank it among other current state-of-the-art systems. Then, we illustrated the "universality" of our system by participating in a recent MT evaluation campaign, with exactly the same core engine, for a wide variety of language pairs. Finally, we studied the influence of extra data like dictionaries and paraphrases on the system performance.

Keywords Example-based machine translation · Proportional analogies · Divergences across languages

Y. Lepage (✉) · E. Denoual
ATR Spoken Language Communication Research Labs, 619-0288 Keihanna, Kyōto, Japan

Present Address:
Y. Lepage
Université de Caen, campus 2,
Bd Maréchal Juin, BP 5186, F-14032 Caen Cedex, France
e-mail: Yves.Lepage@info.unicaen.fr

E. Denoual
GETA-CLIPS-IMAG, Université Joseph Fourier, 38041 Grenoble Cedex 9, France
e-mail: etienne.denoual@atr.jp

1 Introduction

In contrast to some other approaches to machine translation (MT), namely statistical MT (SMT), which do not view linguistic data as specific data, we believe that natural language tasks are specific because their data are specific. The object 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 the preprocessing time of the data and the quality of the results.

We present a novel example-based MT system which relies entirely on proportional analogy. An appealing feature of our system is that it requires no training whatsoever: the data are simply loaded into memory at start-up and they are not preprocessed in any way. This is a definite advantage over techniques that require intensive preprocessing. Another consequence is that it can be applied directly to any language pair for which there are sufficient available data.

We evaluated our system on the tasks of the IWSLT 2004 evaluation campaign (Akiba et al. 2004) in the Japanese–English and Chinese–English "unrestricted data" tracks. This evaluation showed that, at that time, our system would have positioned itself among the top systems for these tracks. We also demonstrated the ubiquity of the system by having exactly the same translation engine participating in all possible language pairs of the IWSLT 2005 evaluation campaign. As the data loaded in memory at start-up constitute the only translation knowledge of our system, we also inspected the influence of these data on translation results.

After this introduction, we detail in Sect. 2 the claims on which our proposal relies. In Sect. 3, we sketch the core process of the translation engine, make some remarks on its features and illustrate it with a number of examples. Section 4 gives the theoretical foundations of the proposal. Section 5 details a number of experiments carried out, their results and some comparison with other state-of-the-art systems. The final section provides a conclusion of the features of the system, discusses the main contributions of the paper, and an evaluation of the results obtained.

2 The claims

This section details the claims on which our proposal relies, dealing in particular with the specificity of linguistic data, divergences across languages, and handling structured representations of proportional analogies.

2.1 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 (cf. also Carl 2006). A consistent consequence is to process linguistic data using operations that specifically capture this systematicity. This systematicity appears at best in commutations exhibited in proportional analogies like those in examples (1).

(1) a. It walks across the street. : It walked across the street. :: It floats across the river. : It floated across the river.

b. 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?

A proportional analogy is noted $A : B :: C : D$ in its general form and reads "A is to B as C is to D". It is a logical predicate that necessarily takes *four* arguments, no less. We stress that four objects are required, because analogy is, alas, too often confused with mere similarity between two objects.[1] Because it is a logical predicate, when we take four sentences in a language to form a proportional analogy, the result is either true or false. For instance, the proportional analogies in example (1) are true, while the one in example (2) is not.

(2) Good morning. : Can I exchange these traveler's checks? $\not::$ It walks across the street. : It floated across the river.

This is exactly the same thing as with proportions on numbers, the difference being that here we use sentences instead of numbers. Hence, for instance, $\frac{5}{15} = \frac{10}{30}$ also written as $5 : 15 = 10 : 30$ is a valid proportion, whilst $\frac{5}{15} \neq \frac{4}{8}$ is not.

The human interpretation of proportional analogies between sentences is that some pieces of the sentences commute with other pieces, so that human beings perceive it as a kind of parallel replacement. In example (3), we put such pieces into boxes. It is clear that such pieces are not necessarily words in general. In example (3), although an English speaker would perceive an exchange of the word *street* with *river*, a shorter explanation in terms of total length of substrings exchanged is possible: *str* commutes with *riv* while *et* commutes with *r*. This may offend one's linguistic sensitivities, but the same speaker would easily admit that the *s* in *walks* in the first sentence is exchanged with *ed* in the second sentence because *s* and *ed* are inflectional morphemes affixed to verbal roots like *walk*.

(3) It walk[s] across the str[e]et. : It walk[ed] across the str[e]et. $\not::$ It float[s] across the riv[e]r. : It float[ed] across the riv[e]r.

From a global point of view, such commutations make paradigmatic and syntagmatic variations explicit and allow for lexical and syntactic variations that ought to be exploited in an MT system to express different meanings. Indeed, any sentence in any language may be cast into a wide number of such proportional analogies that form a kind of meshwork around it. For instance, the sentence in (4a) can be seen in the analogies (4b) to (4d).

(4) a. It floated across the river.

b. It walks across the street. : It walked across the street. :: It floats across the river. : It floated across the river.

c. It swam. : It swam across the river. :: It floated. : It floated across the river.

d. They swam in the sea. : It swam across the river. :: They floated in the sea. : It floated across the river.

The analogy (4b) can be interpreted as showing an opposition between the present and the past tenses in two different sentences. The analogy (4c) can be interpreted as

[1] The confusion of analogy with mere similarity has its root in the scholastic elaboration of the notion. Seemingly, it originates in the writings of St Thomas of Aquinas and their interpretation by St Cajetan (who, nonetheless, duly acknowledged that the only rigorous acception of analogy is when one can say that A is to B as C is to D). Boethius introduced a distinction between "proportions" for ratios and "proportionality" for an equality of ratios, that is, an analogy. The recent work by Gentner (1983) and her colleagues on what they call "analogy" should rigorously be characterized as dealing with the fourth species of metaphors in Aristotle's definitions in the *Poetics*, namely, those metaphors that are based on an analogy: "an atom is like a solar system" because (and only because) "an electron is to the nucleus as a planet is to the sun".

showing that a prepositional phrase may expand some verbs. Consequently, these two analogies can easily be labeled ("present/past" or "with/without PP") and thus located according to some linguistically accepted categorization. But this is not the case in the analogy (4d), where it is difficult to find a label that would adequately characterise all oppositions involved (change in pronouns, singular/plural, different verbs, different circumstantial complements). The purpose of giving such examples of analogies is to show that, if it is usually understood that analogies exemplify well documented and described linguistic phenomena, actual occurrences in corpora are not always ideal examples of definite, and well classified phenomena. In the discussion (see Sect. 7.5), we shall go back to the fact that what our method lacks in order to be more efficient is precisely consistent examples for well described phenomena, such as simple tense oppositions, singular/plural, affirmative/negative/interrogative sentences, and so on, because they do not appear consistently in actual corpora.

In Lepage and Peralta (2004), we have shown how to extract tables (or matrices) automatically 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. An example of such a table is given in Table **??**. It was obtained starting with the sentence (5a). The line with *seafood* and the other lines with *Chinese*, *Italian*, … *food* show that word boundaries do not count as a specific place for commutations. From the table, it is also clear that new sentences may be added into the table into some of the cells that were left blank. For instance, the cell marked (x) can be filled with the sentence (5b). Such a sentence is obtained by solving an analogical equation in the same way as equations in proportions are solved. (5b) can fill the cell marked (x) in example (6a) in the same way as the implication in (6b).

(5) a. I like Japanese food.
b. I'd prefer Italian food.

(6) a. I like Japanese food. : I'd prefer Japanese food. :: I like Italian food. : $x \Rightarrow x =$ I'd prefer Italian food.

b. $5 : 15 = 10 : x \quad \Rightarrow \quad x = 30$

2.2 Dealing with divergences across languages

MT has specific problems to address: one of them, at the core of translation, is to tackle divergences across languages. Back in the early times of MT, the problem was

Table 1 An extract of a table that visualizes several analogical relation between (simple) sentences extracted from our corpus

I like Japanese food.	I prefer Japanese	I'd prefer Japanese food.	I feel like Japanese food.
Do you like Italian food?			Do you feel like Italian food?
I'd like Western food.	I'd prefer Western food.		
I like Chinese food.	I prefer Chinese food.		
I like Italian food.	I prefer Italian food.	(x)	
I like Mexican food.			I feel like Mexican food.
I like seafood.	I prefer seafood.		I feel like seafood.
I like Western food.		I'd prefer Western food.	

pointed out by Vauquois and exemplified with the exchange of predicate arguments between French and English in the famous example in (7) in which the French phrase (a) corresponds to the English (b).

(7) a. *Elle* $_1$ *lui* $_2$ *plaît.* lit. 'She $_1$ to-him $_2$ pleases.'
b. He $_2$ likes her $_1$.

Recent studies (Dorr et al. 2002) confirm the importance of the phenomenon: on a sample of 19,000 sentences between English and Spanish it was estimated that one sentence in three presents divergences. Dorr's (1994) classification into five different types was reused in Habash (2002):

1. categorial divergences: *tener celos* (N) (lit. 'to have jealousy') ↔ *to be jealous* (A)
2. conflation: *ir flotando* (lit. 'to go floating') ↔ *to float*
3. structural divergence: *entrar en* N (lit. 'to enter in N') ↔ *to enter* N
4. head switching: *entrar corriendo* (lit. 'to enter running') ↔ *to run in*
5. thematic divergence: *me gustan uvas* (lit. 'to-me they-please grapes') ↔ *I like grapes*

Let us examine an example of type 4 in further detail, that is, the classical translation of a Spanish verb into an English preposition (Amores and Mora 1998).[2] We can express the word-to-word correspondences by indexing words. The same index shows words in translation correspondence. The correspondence of *atravesó* with *across* through index 1 and that of *flotando* and *floated* through index 3 exemplify a translation divergence of type 4, head switching (8).

(8) a. *Atravesó el río flotando.* IT-CROSSED THE RIVER FLOATING
b. It floated across the river.

c.
1: *Atravesó* $_{V}$		0: It
2: *el río* $_{N}$	↔	3: floated $_{V}$
3: *flotando* $_{\text{particip.}}$		1: across $_{\text{prep.}}$
		2: the river $_{N}$

To show that the complexity of divergences is often underestimated, let us stick with the above interpretation that Spanish *atravesó* would correspond to English *across*. This kind of divergence easily gives rise to a configuration which is excluded by construction from Inversion Transduction Grammars. It is the 14th configuration in Wu (1997, p. 386) , one of the transpositions called "inside-out matchings" by the author, who further claims that he has "been unable to find real examples in [his Chinese–English] data of constituent arguments undergoing 'inside-out' transposition" (ibid., p. 385). However, the introduction of an adverb, which appears after the finite verb in Spanish, and before in English, leads to this very configuration as shown in example (9).

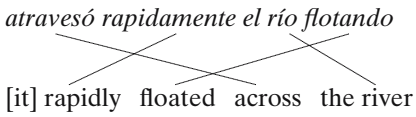
[2] One often generalizes divergences across language families, by saying that motion verbs in Romance languages are usually translated into prepositions or verbal particles in Germanic and Slavic languages. Hence, one would oppose the series (i) to their Germanic or Slavic counterparts in (ii).

(i) a. Fr. *Il traversa la rivière à la nage.* 'he crossed the river at the swim'
b. It. *Ha attraversato nuotando il fiume.* 'he-has crossed swimming the river'
c. Sp. *Atravesó el río nadando.* 'he-crossed the river swimming'

(ii) a. En. He swam across the river.
b. Ger. *Er durchschwamm den Fluss.* 'he cross-swam the river'
c. Pol. *Przepływał przez rzekę.* '[he] cross-swam across river'

A remark on this last example. Although attested on the Web (*argumentum ad Gogulum*!), the verb *przepłynąć/przepływać* does not appear in Polański's (1984) dictionary. In our opinion, this "latency" is a testimony of the productivity of such morphological constructs.

(9)



Let us now show that the view of word-to-word correspondence is all but partial and incomplete. Approaches that adopt the word as the unit of processing neglect 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 example (10) is in fact not sufficiently detailed. Actually, the ending *-ó* of the first Spanish word accounts for the third 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*). Consequently the representation in (10), where boxes show the correspondence, is more correct.

(10) [*Atravesó*] *el río flotando.* ↔ [It] float[ed] [across] the river.

Unfortunately, again, this representation is partial, as it should be repeated for any word in the source language, or any word in the target language, or, even, taken to the extreme, any sequence of characters in both the source and the target language.

If we wanted to drop the view where words correspond to words, we would logically have to deal with more fine-grained units than words, and go to the level of characters. This would mean that, in order to express correspondences, we should compute the correspondences between characters or character strings. This approach is obviously risky as it would imply a combinatorial explosion in the number of correspondences to explore.

The following section will show that indeed, making such correspondences explicit can be avoided. The solution is achieved by stating only the necessary correspondence, the one that exists between two entire sentences in two different languages, as in (11), and relying on the structure of the languages to perform monolingual commutations instead of computing finer bilingual correspondences.

(11) [*Atravesó el río flotando.*] ↔ [It floated across the river.]

2.3 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.

The translation in (12a) is extracted from corresponding sentences of the proportional analogies in (12b) and (12c).[3]

[3] Literal glosses of French examples are shown in (iii) and (vi).

(iii) *Vous pouvez m'échanger un chèque de voyage?*
YOU CAN TO-ME-CHANGE A CHECK OF JOURNEY

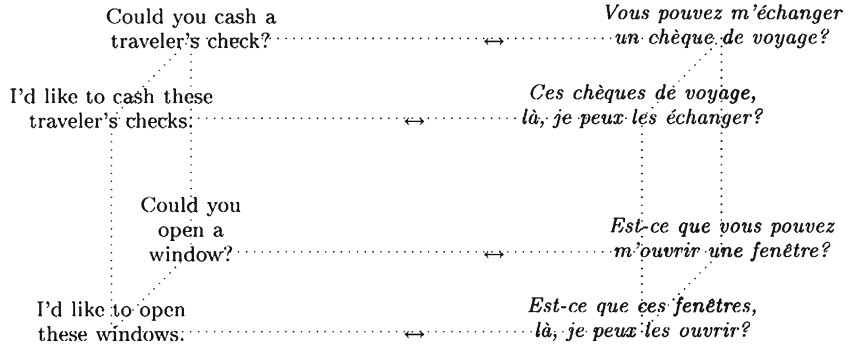


Fig. 1 The parallelopiped: in each language, four sentences form a proportional analogy. There exist four translation relations between the sentences. This is just a geometric representation of example (12)

(12) a. Could you cash a traveler's check? ↔ *Vous pouvez m'échanger un chèque de voyage?*

b.	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?
	↕		↕		↕		↕	
c.	*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?*

Here, we have chosen an example which is much more complex (at least in the French part) than the previous one in (9) with *across the river*, so as to convince the reader of the difficulty of the puzzles that can be solved by our approach. The correspondence can only be established because each sentence in (12c) is a possible translation of the sentence in the corresponding part of (12b) as indicated by a vertical arrow.

Another view of the scene is given by the parallelopiped of Fig. 1. Each of the vertical planes of the parallelopiped resides in one and only one language. The one on the left is the English part, and the one on the right is the French part.

A consequence of this view is that the difficulty which is usually faced in translating between some particular languages partly vanishes (at least in theory!). The claim that it is costly to translate between some specific languages such as Japanese and English relies indeed on the idea that translating would basically consist of rearranging, transforming, or decoding. For instance, Sumita explicitly says: "Because we have succeeded in J[apanese]-to-E[nglish], one of the most difficult translation pairs, we have little concern about other pairs" (2003, p. 205).

Footnote 3 continued

(iv) *Est-ce que ces fenêtres-là, je peux les ouvrir?*
IS-IT THAT THOSE WINDOWS-THERE, I CAN THEM OPEN

(v) *Est-ce que vous pouvez m'ouvrir une fenêtre?*
IS-IT THAT YOU CAN TO-ME-OPEN A WINDOW

(vi) *Ces chèques de voyage-là, je peux les échanger?*
THOSE CHECKS OF JOURNEY-THERE, I CAN THEM CHANGE

However, to make a comparison with clothes, to localize what corresponds to the left shoulder of a shirt on, say, a jacket, one does not rearrange or transform the material of the shirt, that is, 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 may sound strange, this is precisely what second-generation MT systems actually do when they use lexical and structural transfer rules; and what SMT systems (Brown et al. 1993) do when they use lexicon models with distortion models (IBM models 4 and 5).

Rather, it is reasonable to point at the left shoulder of the jacket by looking at the general constitution of the jacket, and by following the different weaves and threads *on* the jacket to localize some point more precisely if needed, as the jacket is made of a different material from the shirt. Transposing to MT, the translation of a source sentence should be sought by relying on the paradigmatic and syntagmatic meshworks, that is, 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.

Basically, the method that we propose for translation is reminiscent of distributionalism (Harris 1954): a sentence can be generated in the target language as long as there is a place for it in the meshwork of the target language. And a sentence of the source language can be translated only to the extent that it occupies some place in the meshwork of the source language. Consequently, expressions that are proper to a particular language, an example of which is the famous English idiom *to kick the bucket*, shall be translated in our proposed framework with no added difficulty than for any other usual sentence, as exemplified in (13).

(13) a.

He swam across the river.	:	She swam across the river.	::	He kicked the bucket.	:	She kicked the bucket.
↕		↕		↕		↕
Il traversa la rivière à la nage.	:	*Elle traversa la rivière à la nage.*	::	*Il mourut.*	:	*Elle mourut.*

b.

He swam across the river.	:	She swam across the river.	::	He died.	:	She died.
↕		↕		↕		↕
Il traversa la rivière à la nage.	:	*Elle traversa la rivière à la nage.*	::	*Il mourut.*	:	*Elle mourut.*

3 EBMT by proportional analogy

3.1 The algorithm

The following gives the basic outline of the method we propose to perform the translation of an input sentence. Let us suppose that we have a corpus of aligned sentences in two languages at our disposal. Let D be an input sentence to be translated into one or more target sentences $\widehat{D}$.

— Form all analogical equations with the input **sentence** D and with all pairs of **sentences** (A_i, B_i) from the source part of the parallel aligned corpus (14).

(14) $A_i : B_i :: \mathrm{x} : D$

— For those sentences that are solutions of the previous analogical equations, but that do not belong to the parallel aligned corpus, translate them using the present method recursively. Add them with their newly generated translations to the parallel aligned corpus.

— For those sentences $x = C_{i,j}$ that are solutions of the previous analogical equations (one analogical equation may yield several solutions) and which do belong to the parallel aligned corpus, do the following:

— Form all analogical equations with all possible target-language sentences corresponding to the source-language sentences (several target sentences may correspond to the same source sentence) (15).

(15) $\widehat{A_i}^k : \widehat{B_i}^k :: \widehat{C_{i,j}}^k : \mathrm{y}$

— Output the solutions $y = \widehat{D_{i,j}}^{k,l}$ of the analogical equations as a translation of D, sorted by frequencies (different analogical equations may yield identical solutions).

3.2 Some remarks

As the above algorithm may be misunderstood in various ways, it is necessary for us to make some remarks and clarify some points. This section serves this purpose.

First, in order to avoid any misinterpretation where the method would be considered a method by decomposition where some breaking operation is involved, let us stress that A_i, B_i and D are *sentences*; they are *not fragments* of sentences. Sentences are *not cut into pieces* by the proposed method.

Second, *pairs* of sentences are retrieved to form an analogical equation with D. Consequently, speaking about "analogous examples" does not make any sense in this framework. Again proportional analogy should not be confused with mere similarity, and it should be stressed that, indeed, A_is and B_is may be "far away" from D in terms of edit distance (Levenshtein 1965).

Third, according to the previous description, the complexity of the translation method is basically quadratic in the size of the examples. Of course to reduce this complexity in our actual implementation, relevant pairs of sentences are selected on the fly according to some criterion. It suffices to say that in our current implementation, we do not inspect pairs (A, B) where the length of B is less than half that of D or more than twice that of D (and the same for A relative to B). We shall not elaborate on this criterion as it is obvious that it is not optimal, neither theoretically, nor in terms of efficiency. It still remains to be determined what kind of criterion would be most efficient.

Four, it follows from the algorithm above and the properties of proportional analogies that the method is *nondeterministic*. A plurality of translations may be obtained for one input sentence. This was made explicit in the algorithm above by the use of indices for As, Bs, Cs, and their counterparts in the target language, and also in the remarks in parentheses. Let us make it clear that the nondeterminism of the method has four reasons:

(a) many pairs (A_i, B_i) can lead to a translation for D (hence the use of i to index these pairs);
(b) analogical equations may have a plurality of solutions, so that any analogical equation (14) in the source language may yield several solutions (hence the introduction of index j to denote such solutions: $C_{i,j}$);
(c) each of A_i, B_i, and $C_{i,j}$ may have different translations, so that for any such source triple, there may be several corresponding analogical equations to solve in the target language (hence index k);
(d) again, as analogical equations may have a plurality of solutions, the analogical equations (15) in the target language may yield different solutions (hence index l for $\widehat{D_{i,j}}^{k,l}$).

Finally, it is necessary to mention that the same translation for the same input sentence may be output through different paths, that is, different pairs (A, B); different analogical equations $A : B :: x : D$; different solutions to such analogical equations; different translations $\widehat{A}$ for A, $\widehat{B}$ for B, and $\widehat{C}$ for C; different analogical equations $\widehat{A} : \widehat{B} :: \widehat{C} : y$ and different solutions to such analogical equations. To sum up, each different candidate translation $\widehat{D}$ output for D may be assigned a number which is the number of times this particular $\widehat{D}$ was output. In Figs. 3 and 4 below, where actual examples of translations are shown, these numbers are given at the left of each particular translation. It must be noted that these numbers are not small: in our experiments with a corpus of 160,000 aligned sentences, the same translation for the same input sentence may be output thousands of times. Currently, we exploit these numbers poorly, as we consider only that the most frequent translation should be the best one. This elementary criterion is used to select which translation candidate we use in evaluation with mWER, BLEU, NIST, etc.

3.3 A simple example without recursion

To illustrate the method, suppose that we wanted to translate the Japanese input sentence in (16).[4]

(16) D = 濃いコーヒーが飲みたい。
koi kōhī ga nomitai. STRONG COFFEE-NOM DRINK-VOLITIVE
'I want to drink strong coffee.'

At some point in the exploration of all possible pairs of sentences from the parallel aligned corpus, we will find the two Japanese sentences in (17) that literally translate as shown, but are actually translated as (18) in our database of examples.

(17) a. (A) 紅茶をください。*kōcha wo kudasai*. TEA-obj PLEASE 'Tea, please.'
b. (B) コーヒーをください。*kōhī wo kudasai*. COFFEE-obj PLEASE 'Coffee, please.'

(18) a. ($\widehat{A}$) May I have some tea, please?
b. ($\widehat{B}$) May I have a cup of coffee?

Sentences A and B will allow us to form the analogical equation (19a). This equation yields (19b).

[4] Transliterations, literal glosses and translations are shown for the reader's information, and are of course not part of the example.

(19) a. $(A) : (B) :: (C) : (D)$

b. C = 濃い紅茶が飲みたい。
koi kōcha ganomitai. STRONG TEA-nom DRINK-volitive
'I want to drink strong tea.'

If sentence C is already part of the parallel aligned corpus, that is, if the translation pair (20a) is found in the data, the analogical equation (20b) is formed with the corresponding English translations. By construction, the solution (20c) is a candidate translation of the input sentence (16).

(20) a. (C) 濃い紅茶が飲みたい。 $\leftrightarrow (\widehat{C})$ I'd like some strong tea please.
b. $(\widehat{A}) : (\widehat{B}) :: (\widehat{C}) : (\widehat{D})$
c. $\widehat{D}$ = I'd like a cup of strong coffee.

The processing of the previous example can be viewed in the shape of a parallelopiped similar to the one of Fig. 2. The left plane of this parallelopiped is the plane of the English analogy. The right plane is the Japanese one. The translation that was established is the one along the bottom front horizontal line.

3.4 Actual examples of translation

Figures 3 and 4 show some actual examples of translation. As our motivation was to deal with translation divergences, Fig. 3 shows the fact that some parts of speech in Japanese are rendered by different English parts of speech in actual translation results. The numbers on the left are the frequencies with which each translation candidate has been output. Unacceptable translation candidates have been struck out.

Some parts of speech were changed during translation. In the first example in Fig. 3, the Japanese noun 料金 *ryōkin* 'fee/price' appears to have been translated by verbs (*cost*, *charge*) in addition to being translated by nouns (*price*, *fee*) in other sentences. In the second example, the Japanese adjective 痛い *itai* 'painful' has been translated by a verb (*hurts*) or different nouns (*ache*, *pain*). This example is the linguistic realisation

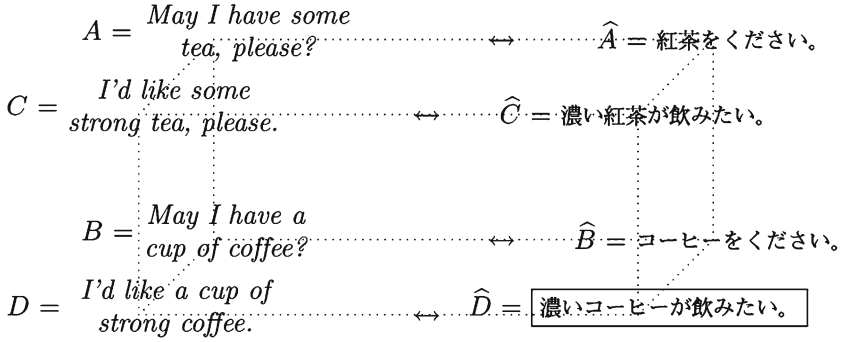


Fig. 2 The parallelopiped for a translation from English into Japanese. D is the input, $\widehat{D}$ is the output

このツアーの 料金 はいくらですか。

kono tsuā no *ryōkin* *wa ikura desu ka?*

THIS TOUR gen FEE/PRICE topic HOW-MUCH TO-BE interrog

271	How much does this tour cost?
160	How much do you charge for this tour?
141	What's the price of this tour?
94	What does this tour cost?
43	What's the price of the tour?
6	What is the price of the tour?
6	~~How much is the green fee~~?

胃が 痛い んです。

i ga *itai* *n desu.*

STOMACH nom PAINFUL insist TO-BE

1744	I have a stomach ache.
552	My stomach hurts
124	I've got a stomach ache.
56	~~Do you have a stomach ache~~.
51	~~Do you have a stomach ache~~?
50	~~I have a stomach ache~~?
2	My stomach hurts me.
1	I have an abdominal pain in my stomach.
1	I have a pain in my stomach.
1	~~I have a soare throat~~.

Fig. 3 Examples of translations

of a statement about pain. In various languages, this kind of statement is realised in various ways, some of which can be schematised as in (21) (not an exhaustive list).

(21) a. <body-part> <hurts> [<somebody>]
b. [<somebody>-DAT | <somebody>-GEN] <body-part> <hurts>
c. [<there-is> | <somebody> has] <pain> in <body-part>

The first form is used in the second example in Fig. 3 repeated here has (22a), (where 'hurts' is expressed by an adjectival predicate). The second form is typical of Spanish (22b), while the last one is usual to French (22c).

コーヒーのおかわりをいただけますか。

kōhī no o-kawari o itadakemasu ka.
COFFEE gen polite CHANGE/AGAIN acc CAN-RECEIVE interrog

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。

kozeni o mazete kudasai.
COINS/SMALL-CHANGE acc MIX PLEASE

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?
877 Please give me some small change as well.

Fig. 4 Further examples of translations

(22) a. 胃が痛いんです。
i ga itai n desu.
STOMACH nom PAINFUL insist TO-BE
b. *Me duele la cabeza, la mano, . . .*
TO-ME HURTS THE HEAD, THE HAND, . . .
c. *J'ai mal* à *la tête,* à *la main, . . .*
I-HAVE PAIN IN THE HEAD, IN THE HAND, . . .

The translation example of Fig. 3 shows that our system could synthesize some of these forms of expression for the sentence to be translated. Of course this was only possible because, the only knowledge of the system being the parallel corpus, realisations of such patterns were actually present in the corpus. What the example demonstrates is that the system could properly synthesize various forms of

expression for the sentence to be translated because it could exploit actual realisations of such forms of expression and perform proper commutations. This demonstrates that actual unprocessed examples in conjunction with proportional analogy are at least as powerful as predefined patterns or templates with variable positions marked as in (23).

(23) I have a <body part> ache

Moreover, such patterns are often too restricted for actual language usage, and they prevent the elastic use of idioms. Some years ago, Marcel Bigeard, a former general in the French army, wrote a book entitled *J'ai mal à la France*, meaning something like 'I have a France ache', or 'France hurts me', with the implied meaning that France was no external object for him, in contrast to a situation like *a needle hurts me*. A system where examples remain unprocessed has a greater chance of translating such an expression than a system where the previous pattern would be associated with a constraint restricting the choice to body parts.

4 Theoretical foundations of the method

This section gives some deeper insights into the theoretical aspects of the method proposed above. It should be mentioned that the method was in fact derived from the theoretical work that will now be described, not the contrary.

4.1 Proportional analogies between strings of characters

Our notion of analogies between sentences, or to be more precise between strings of characters, reaches back as far as Euclid and Aristotle: "*A* is to *B* as *C* is to *D*", postulating the identity of types for *A*, *B*, *C*, and *D*. The notion was put forward in morphology by Apollonius Dyscolus and Varro in antiquity. In modern linguistics, de Saussure (1955) (Part III, Chap. 5) considers analogical equations as a typically synchronic operation by which, given two forms of a given word, and only one form of a second word, the fourth missing form is coined: "*honor* is to *honōrem* as *ōrātor* is to *ōrātōrem*",[5] as shown in (24).

(24) $\bar{o}r\bar{a}t\bar{o}rem : \bar{o}r\bar{a}tor :: hon\bar{o}rem : x \quad \Rightarrow \quad x = honor$

According to de Saussure, this explains the fact that in the second century BC, the form *honor* actually competed with the etymologically correct form *honos*. While analogy has been largely mentioned and used in linguistics, only recently can we see applications of the notion in natural language processing to pronounciation, morphology or terminology (Skousen 1989, Damper and Eastman 1996, Hathout 2001, Claveau and L'Homme 2005, Stroppa and Yvon 2005, among others).

That analogy applies also to syntax, which is the foundation of our framework, has been advocated by Paul (1920, p. 110) and Bloomfield (1933, p. 275). More recently, Itkonen (1999) showed how to deliver grammatical sentences by application of proportional analogies to structural representations.

[5] Latin *ōrātor* 'orator/speaker' and *honor* 'honor' are nominative singular, *ōrātōrem* and *honōrem* accusative singular.

Algorithmic ways to solve proportional analogies between strings of characters have never been proposed, perhaps because the operation seems so misleadingly "intuitive". An exception is Copycat (Hofstadter et al. 1994, pp. 205–265), which adopts an Artificial Intelligence point of view, unfortunately of little use for linguistic applications that require very fast computation. To our knowledge, we were the first to give an efficient algorithm for the resolution of analogical equations in Lepage (1998). Our proposal is based on the following formalisation of proportional analogies (Lepage 2003) in terms of edit distances, or equivalently, in terms of similarity (refer to Chap. 3 of Stephen (1994) for these notions and see Delhay and Miclet (2004) for an extension of this formalisation to alphabets equipped with an algebraic structure).

We denote $\sigma(A, B, \ldots, N)$ as the length of the longest common subsequence in the strings A, B, …, N, that is, their similarity. We also denote $|A|_a$ as the number of occurrences of character a in string A and $|A|$ as the length of A. The formula in (25) consistently puts the unknown D on the left of all equal signs, so as better to suit the resolution of analogical equations.

$$A : B :: C : D \;\Rightarrow\; \begin{cases} \sigma(B,D) = -|A| + |B| + \sigma(A,C) \\ \sigma(C,D) = -|A| + |C| + \sigma(A,B) \\ \sigma(A,B,C,D) = -|A| + \sigma(A,B) + \sigma(A,C) \\ |D|_a = -|A|_a + |B|_a + |C|_a, \quad \forall a \end{cases} \tag{25}$$

As a remarkable property of proportional analogies, it is worth mentioning that the last relation in (25), which expresses the fact that the number of occurrences of any character a in A and D is equal to the number of its occurrences in B and C, trivially implies that the lengths of the extremes (A and D on one hand and B and C on the other hand) are equal as in example (4d) above, reproduced here as (26).

$$\begin{array}{ccccccc} \text{They swam in the sea.} & : & \text{It swam across the river.} & :: & \text{They floated in the sea.} & : & \text{It floated across the river.} \end{array} \tag{26}$$

For instance, applied to (26), by counting characters we have the outcome as shown in (27a) and (27b) (a space counts as one character).

(27) a. $|A| + |D| = |B| + |C|$
b. $21 + 28 = 25 + 24$

The step-by-step mechanism we adopt during resolution is inspired by Itkonen (1999, p. 149), where they take sentence A as the axis against which sentences B and C are compared, and by opposition to which output sentence D is built.

Rather than explaining once again the algorithm given in Lepage (1998), we sketch its application in an actualised way on a particular analogical equation in example (28). The Arabic words ارسل *arsala* 'he sent' and أسـلم *aslama* 'he converted [to Islam]' are third person singular past verbs; مرسل *mursil* 'a sender' and مسـلم *muslim* 'a convert', (that is, 'a muslim') are agent nouns.

(28) *aslama* : *muslim* :: *arsala* : $x \quad \Rightarrow \quad x =$ *mursil*

For this sketch, we use words rather than sentences for reasons of space; the same algorithm applies to analogical equations between sentences considered as strings of characters; and it also applies to languages like Japanese, Chinese or Korean where a character is encoded by two bytes instead of just one byte for English.

We use (transliterated) Arabic words to show that solving analogical equations is not reduced to a trivial matter of exchanging prefixes or suffixes. In the morphology

of Semitic languages, proportional analogies ought to capture parallel infixing (something that may help in the Arabic–English C-STAR track of IWSLT 2005, see Sect. 6.3). But more generally, parallel infixing is indispensable in our framework because proportional analogies between sentences involve parallel infixing in almost all of the cases.

As for the algorithm, pseudodistance matrices between strings A and B, and A and C, are first computed. The pseudodistance used here counts only insertions, and overlooks substitutions and deletions. A cell in a pseudodistance matrix contains the value of the pseudodistance computed from the beginning of the strings up to the current positions. For instance, the cell with a value of 2 in boldface in the example shown in Fig. 5 gives the value of the pseudodistance between the two prefixes of C and A, *ars*, and *asla*. To pass from *ars*, to *asla*, it is necessary and sufficient to insert two symbols, *l* and *a*, and to delete one: *r*. Deletions do not count, only insertions count, hence a value of 2. Based on the fast algorithm by Allison and Dix (1986) such a computation is performed in an efficient way. Also, a result by Ukkonen (1985) allows us to compute only minimal diagonal bands in those matrices.

The algorithm which computes the fourth term of the analogy follows all possible paths in parallel in the pseudodistance matrices, in a way similar to that taken in Wagner and Fischer (1974) for the output of edit-distance traces. In the example in Fig. 5, a particular path is made explicit by circles, the indices of which indicate the move number. Paths start from the bottom of the matrices. Some constraints apply: circles with the same index must appear on the same line in both matrices; an arithmetic formula on the symbols in A, B, and C must yield a result. Because of the constraints, paths may deviate from a trace in the sense of Wagner and Fischer (1974). For instance, in the example, move number 5 lands on 1, not on the 0 to its right.

The succession of moves (see Table 2) is read in parallel in both matrices from the bottom to the top. Each move triggers the copies of characters into the solution D (thus, in reverse order) according to "rules" that tell which character to choose from which string B or C according to the different combinations of moves (diagonal, horizontal or vertical in each matrix). For instance, vertical moves forbid writing into D, two vertical moves are forbidden, and so on. As a result, the solution D = *mursil* is output as a possible solution for the analogical equation at hand.

In some cases, no path exists, which means that there is no solution to the analogical equation, and in general, several paths may exist so that there may be several solutions to an analogical equation. (Unfortunately, our formalization of proportional analogies is yet to be completed, so that, with our implementation, we sometimes obtain more

B =	m	i	l	s	u	m		a	r	s	a	l	a	= C
	1	1	1	1	①$_{6,7}$	①$_{8}$	a	⓪$_{6,7,8}$	0	0	0	0	**0**	
	1	1	1	①$_{5}$	2	2	s	1	①$_{5}$	0	0	0	0	
	1	1	①$_{4}$	2	3	3	l	2	2	①$_{4}$	1	0	0	
	2	②$_{3}$	2	3	4	4	a	3	3	**2**	①$_{3}$	1	0	
	②$_{2}$	3	3	4	4	4	m	4	4	3	2	②$_{2}$	1	
	②$_{1}$	3	3	4	4	4	a	5	5	4	3	3	②$_{1}$	
							‖							
							A							

Fig. 5 Ilustration of the pseudodistance algorithm

Table 2 Trace of moves in Fig. 4

Move	Dir_{AB}	Dir_{AC}	Copy onto D	from string
1	vertical	diagonal	ϵ	none
2	diagonal	diagonal	$-m+m-l=l$	C
3	diagonal	diagonal	$-a+i+a=i$	B
4	diagonal	diagonal	$-l+l+s=s$	C
5	diagonal	diagonal	$-s+s+r=r$	C
6	horizontal	no move	$-a+u+a=u$	B
7	no move	no move	$-a+m+a=m$	B
8				

solutions than desired, but never fewer, at least as far as our experiments show on hundreds of linguistic examples in morphology and examples from formal languages.)

Analogical equations are thus a ternary operation, that is, a mapping $\alpha: \mathcal{L}\times\mathcal{L}\times\mathcal{L} \mapsto \wp(\mathcal{L})$ where $\wp(\mathcal{L})$ is the power set of $\mathcal{L}$, the set of strings considered. The set of the solutions of an analogical equation is given in (29).

(29) $\alpha(A,B,C) = \{\ D \in \mathcal{L} \ \mid \ A : B :: C : D \ \}$

4.2 Languages of analogical strings

Based on proportional analogies, we have shown (Lepage 2001) how to define a family of formal languages, called "languages of analogical strings". It is important to note that their construction does not make any use of nonterminals as is the case with simple contextual grammars (Ilie 1998).[6] In fact, our proposal shares some aspects and concerns of contextual grammars.

Languages of analogical strings are built by transitive closure starting from a corpus of given sentences (strings of characters) Λ_0. We denote $\alpha(\Lambda,\Lambda,\Lambda)$ as the set of sentences produced as in (30) by solving all possible analogical equations formed with three sentences in Λ.

(30) $\alpha(\Lambda,\Lambda,\Lambda) = \{D \mid \exists(A,B,C) \in \Lambda^3,\ A : B :: C : D \ \}$

Then, the language $\mathcal{L}(\Lambda_0)$ of analogical strings built from a corpus Λ_0 is defined as shown in (31).

(31) $\mathcal{L}(\Lambda_0) = \bigcup_{n=0}^{+\infty} \Lambda_n$ where $\Lambda_{n+1} = \alpha(\Lambda_n,\Lambda_n,\Lambda_n)$

In fact, there is a chain of set inclusions $\Lambda_{n+1} \supset \Lambda_n$ because $A : A :: A : x \Rightarrow A = x$.

As for the position of such languages in Chomsky–Schützenberger hierarchy, it is easy to show that the classical regular language $\{a^n | n \geq 1\}$, the context-free language $\{a^n b^n | n \geq 1\}$, and the context-sensitive language $\{a^n b^n c^n | n \geq 1\}$ are all languages of analogical strings. Moreover, we have shown (Lepage 2001) that the famous context-sensitive language $\{a^n b^m c^n d^m \mid m,n \geq 1\}$ used by Shieber (1985) to refute the context-freeness hypothesis of natural language, is a language of analogical strings. More importantly, every language of analogical strings meets the "constant growth

[6] Note that *contextual* grammars, are not to be confused with *context-sensitive* grammars.

property", a property that intervenes partially in the definition of mild context-sensitivity, a notion introduced in Joshi et al. (1991) to cope with the apparent power of human languages.

4.3 Homomorphisms between languages of analogical strings

The framework for translation by proportional analogies that we propose sees both the source and the target languages as languages of analogical strings that are defined from the set of sentences given in the parallel corpus. If we denote Λ_0 as the source-language part of the parallel corpus and $\widehat{\Lambda_0}$ as its target-language counterpart, we idealize the entirety of both source and target languages as being $\mathcal{L}(\Lambda_0)$ and $\mathcal{L}(\widehat{\Lambda_0})$ according to the above notations.

Let us denote $\widehat{A}$ as the (set of) translations of a sentence A. The principle of translation is based on the intuitive formula (32a) that is a transcription of the parallelopiped of Fig. 2. Using the α operation that structures the source and target languages of analogical strings, an equivalent form of this formula is shown in (32b)

(32) a. $A : B :: C : D \;\Leftrightarrow\; \widehat{A} : \widehat{B} :: \widehat{C} : \widehat{D}$

b. $\widehat{D} = \widehat{(\alpha(A, B, C))} = \alpha(\widehat{A}, \widehat{B}, \widehat{C})$

This shows that this translation principle "distributes" translation on the arguments of the structuring internal operation α. Thus, it is a homomorphism between two languages of analogical strings that preserves the structuring operation, proportional analogy.

As we are concerned with translation divergences, and because divergences often imply different reorderings in different languages, let us add a word on this question and mention that the previous formalisation is indeed able to solve "difficult" reordering problems. With its translation knowledge reduced to the two translation pairs $abc \leftrightarrow abc$, $abcabc \leftrightarrow aabbcc$, the system translates members of the regular language $\{ (abc)^n \mid n \in N^* \}$ into the corresponding members of the context-sensitive language $\{ a^n b^n c^n \mid n \in N^* \}$, and reciprocally by solving $2 \times (n - 2)$ proportional analogies recursively (33).

(33) $(abc)^n \leftrightarrow a^n b^n c^n$

5 Features of the method

5.1 No explicit 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, let us consider the Spanish example (8) of Sect. 2.2 again (34). The correspondences between the source and target language in a proportional analogy will be entirely responsible not only for the selection of the correct lemmas with their lexical part of speech, but also for the correct word order (see above for reordering).

The technique is also more general than the translation of the adnominal particle N_1 *no* N_2 from Japanese into English in Sumita and Iida (1991) (cf. also Hutchins 2006) where the choice of the correct preposition (or word order) is left to the list of examples.

(34) a.	They swam in the sea.	:	They swam across the river.	::	It floated in the sea.	:	It floated across the river.
	↕		↕		↕		↕
b.	*Nadaron en el mar.*	:	*Atravesaron el río nadando.*	::	*Flotó en el mar.*	:	$\widehat{D}$

However, it should be stressed that in proportional analogies like the two in example (34), nowhere is it said which word corresponds to which word, or which syntactic structure corresponds to which syntactic structure. To go back to the previous Spanish–English example (8) on which we made explicit word-to-word correspondences, we stress again that in our method, the system does not see any subcorrespondences below that of the global correspondence between sentences. Hence, if we keep the same convention as before and put corresponding parts of the sentences into boxes, all that the system sees appears in fact as in (35), that is, the system sees only the entire correspondence between two sentences: a sentence in the source language corresponds to a sentence in the target language.

(35) [*Atravesó el río flotando.*] ↔ [It floated across the river.]

The sole action of proportional analogy with (necessarily) *the character as the only unit of processing* is sufficient to produce the exact translation of (36a), which is the correct Spanish sentence (36b), provided that the three sentence pairs to the left of (34) are valid translation pairs.

(36) a. It floated across the river.
b. $\widehat{D}$ = *Atravesó el río flotando.*

5.2 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 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 neutralize these divergences as they are the implicit constitutive material of proportional analogies.

Our system definitely positions itself in the EBMT stream, but it departs from it in one important aspect: it does not make any use of explicit symbolic knowledge such as templates with variables, nor does it produce any such templates. Direct use of parallel aligned corpus data in their raw form is made without any preprocessing.

The reason for doing so is that templates may well be insufficient in representing all of the implicit knowledge contained in examples. Indeed, variables in templates allow for paradigmatic variations only at some predefined positions. In Sato (1991), so as to acquire a grammar, sentences which differ by one word only are fed into a system. However, only regular languages can be learned by this technique.

For instance, extracting the template *X salts Y* from the example sentence (37), where *X* may be replaced by *the butcher* and so on, and *Y* by *the slice* and so on, (example taken from Carl 1998 p. 251), does not make the most of the potential of the example.

(37) The butcher salts the slice.

First, it prevents *the butcher* from being changed into a plural: *the butchers*. Moreover, it overlooks the fact that *salts* may also commute with its past and future forms: *salted*, *will salt*, and so on, or with *cuts*, *smokes*, and so on. To summarize, there is a risk of loss of information when replacing examples with templates.

The situation is in no way better with translation patterns. On the one hand, it is true that such translation patterns can be very efficiently indexed so that their retrieval is very fast. Superfunctions introduced by Sasayama et al. (2003) are such a means to extract and retrieve these kinds of translation associations using arrays. They make it explicit which variables in the source have to be replaced by which variables in the target. 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, syntactic, and so on). 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 (38).

(38) He eats. → He does not eat.

Our view is that *every* position in a language datum is subject to paradigmatic variation. Taking it to the extreme, even phonetic variations have to be considered: *wolf* : *wolves* :: *leaf* : *leaves*, so that one definitely has to go below the level of words. For this reason, our system processes *strings of characters,* not strings of words. The consequence is that a lot more exploitable information should be found in unprocessed examples than in templates. And it may well be the case that the templates necessary to encode the information contained in a set of examples are 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 rather in their capacity to get involved in proportional analogies.

5.3 No training, no preprocessing

As a consequence of the above-mentioned features, there is no such thing as a training phase or a preprocessing phase in our system: the parallel aligned corpus is just loaded into memory at program start-up. No language model is computed; no alignment other than the one given by the parallel aligned corpus is extracted; no segmentation or tagging whatsoever is performed. Needless to say, the possibility of adding new information to the parallel aligned corpus 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 Sects. 6.4.2 and 6.4.3).

6 Experiments, evaluation, comparison

6.1 Resources used in evaluation

To assess the performance of the proposed method, we used the C-STAR Basic Traveler's Expressions Corpus (BTEC) (Takezawa et al. 2002).[7] This is a multilingual resource of expressions from the travel and tourism domain that contains almost 160,000 aligned translations in Chinese, Korean, English, and Japanese. In this resource, the sentences are quite short as the figures in Table 3 show. As the same sentence may appear several times with different translations, the number of different sentences in each language is indicated in Table 3.

The method relies on the assumption that analogies of form are almost always analogies of meaning. Thus, prior to its application, in Lepage (2004) we 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 endanger seriously the quality of the results obtained during translation. An example of a false analogy, that is, an analogy of form extracted from the sampling set that is not valid, is given in (39).

(39) Could you tell me how to fill this from. : Could you tell me how to fill this form. $\not::$ Where is the conference centre? : Where is the conference center?

In example (39), the commutation between *center* and *centre* is just a matter of dialects of English, whereas the commutation between *from* and *form* is not acceptable, the first one being a spelling error.

6.2 Comparison with other systems

We assessed our system using the IWSLT 2004 tasks in both Japanese–English and Chinese–English directions. As we used a parallel aligned corpus of 160,000 examples our results should be compared with those of the "unrestricted data" track reported in the proceedings of the evaluation workshop (Akiba et al. 2004).

In this track, no restrictions were imposed on linguistic resources. As for tools, our system did not make any use of any NLP tool such as a tagger to preprocess the data. In particular, we chose to place ourselves in the condition of standard natural Japanese and Chinese texts (in which no segmentation appears), so that we had to undo segmentation in the test sets provided. This clearly demonstrates that segmentation is

Table 3 The C-STAR BTEC: number of sentences and sentence lengths

	Number of different sentences	Size in characters	
		Average	Std. deviation
Chinese	96,224	9.40	5.17
English	97,395	35.17	18.83
Japanese	103,051	16.22	7.84
Korean	92,626	11.67	5.99

[7] The corpus is not in the public domain, but it is of easy access for research institutes wishing to become a C-STAR partner, and parts of it are delivered every year during the IWSLT evaluation campaign as supplied data (training corpus).

not a necessity to perform a translation task from Japanese or Chinese. We consider that translation ought to be performed as much as possible on unmodified real texts without preprocessing as we want to evaluate MT systems, not preprocessing tools. As for data, no dictionary was used. The C-STAR corpus of around 160,000 aligned sentences described above was used for both language pairs. We refer to this as our "training data", although there is absolutely no training phase within our framework. All these conditions are summarised in Table 4.

In addition to the previous conditions, and in order to avoid the fact that some sentences in the test data may be included in the "training data", we assessed our system in two configurations: "standard" and "open". The difference between the two is that, in the latter, any sentence from the test set was removed from the "training data", if found there.

Some examples of Japanese–English translations have already been given in Figs. 3 and 4. Let us recall that the numbers to the left of a translation candidate are the frequencies with which it has been output (see Sects. 3.1 and 3.2). As we assumed that the most frequent candidate should be the most reliable one, the evaluation was performed on the first candidates only.

Tables 5 and 6 summarize the evaluation results obtained with the objective criteria used in this evaluation campaign. The results for other systems were copied from Akiba et al. (2004, p. 11). The evaluation measures used are multiple-translation word-error rate (mWER), multiple reference position-independent word-error rate (mPER), BLEU score (Papineni et al. 2002), NIST score (Doddington 2002), and General Text Matcher (GTM) (Turian et al. 2003). Higher scores are better, except for mWER and PER, where lower scores indicate better results. The top score is shown in bold in each case. In its "open" configuration our system tries to translate an input sentence again if it already belongs to the data, whereas in its "standard" configuration, it outputs the translation found in the data.

The results obtained for our system are very promising as our system achieves second place in Chinese–English, and third place in Japanese–English. A standout point is the achievement with the BLEU score: a close second for Chinese–English (0.522, first at 0.524), and the best one for Japanese—English (0.634). Unfortunately, we are not in a position to reproduce the subjective evaluation for the translation results output by our system. It must be stressed again that these results were obtained with-

Table 4 Among the permitted resources, our system used only the C-STAR 160,000 aligned sentences

Resources	Data track	
	Unrestricted	Our configuration
IWSLT 2004 corpus	yes	yes
LDC resources	yes	no
Tagger	yes	no
Chunker	yes	no
Parser	yes	no
External bilingual dictionaries	yes	no
Other resources	yes	140,000 additional aligned sentences

The IWSLT 2004 supplied corpus of 20,000 sentences is a subset of the C-STAR corpus, so that the other resources that our system used are just the remaining 140,000 sentences

Table 5 Scores for the IWSLT 2004 Chinese-to-English "unrestricted data" track: no restriction on linguistic resources

System	Type	mWER	mPER	BLEU	NIST	GTM
ISL-S (Vogel et al. 2004)	SMT	**0.379**	**0.319**	**0.524**	**9.56**	**0.748**
ours (standard)	EBMT	0.434	0.400	0.522	8.42	0.687
ours (open)	EBMT	0.437	0.404	0.512	8.24	0.682
IRST (Bertoldi et al. 2004)	SMT	0.457	0.393	0.440	7.24	0.671
IBM (Lee and Roukos 2004)	SMT	0.525	0.442	0.350	7.36	0.684
ISL-EDTRL (Reichert and Waibel 2004)	hybrid	0.531	0.427	0.275	7.50	0.666
ISI (Thayer et al. 2004)	SMT	0.573	0.499	0.243	5.42	0.602
NLPR (Zuo et al. 2004)	hybrid	0.578	0.531	0.311	5.92	0.563
HIT (Yang et al. 2004)	EBMT	0.594	0.487	0.243	6.13	0.611
CLIPS (Blanchon et al. 2004)	rule-based	0.658	0.542	0.162	6.00	0.584
ICT (Hou et al. 2004)	EBMT	0.846	0.765	0.079	3.64	0.386

Table 6 Scores for the IWSLT 2004 Japanese-to-English "unrestricted data" track: no restriction on linguistic resources

	Type	mWER	mPER	BLEU	NIST	GTM
ATR-Hybrid (Sumita et al. 2004)	hybrid	**0.263**	**0.233**	0.630	10.72	0.796
RWTH (Bender et al. 2004)	SMT	0.305	0.249	0.619	**11.25**	**0.824**
ours (standard)	EBMT	0.324	0.300	**0.634**	9.19	0.731
ours (open)	EBMT	0.437	0.403	0.534	8.97	0.697
UTokyo (Aramaki and Kurohashi 2004)	EBMT	0.485	0.420	0.397	7.88	0.672
CLIPS (Blanchon et al. 2004)	rule-based	0.730	0.597	0.132	5.64	0.568

out any training performed in advance on the data, and that no tuning whatsoever of the system toward the "training data" was performed.

6.3 Comparison for different language pairs

Our system does not require any training phase, so that data are merely loaded into memory before the system is made ready to translate. The IWSLT 2005 campaign offered a number of language pairs, with the possibility of using a multilingual corpus, where the amount and meaning of sentences are identical. We chose to participate in all C-STAR data tracks with exactly the same core engines in order to be able to compare the results obtained on different language pairs, provided that the evaluation procedure was also the same. Our goal was to learn some lessons on the difficulty of translating some language pairs relative to others with our proposed method. As only one configuration was allowed, we chose to use the "open" configuration of our system because it seemed the most honest attitude to inspect the potential of our method: whenever an input sentence was recognised as belonging to the training data, we excluded it from the database of translation pairs and tried to translate it anew. To do so seriously handicapped us, because such cases did actually occur. Of 506 sentences to translate, 90 did in fact belong to the training set (and even to the supplied data of

Table 7 Scores for all IWSLT 2005 C-STAR tracks

Language pair	mWER	mPER	BLEU	NIST	GTM	Remarks
English–Chinese	0.798	0.746	0.098	3.03	0.363	1 reference
Arabic–English	0.527	0.497	0.382	6.22	0.481	20,000 pairs
Korean–English	0.530	0.486	0.412	7.12	0.446	
Chinese–English	0.454	0.418	0.477	7.85	0.553	
Japanese–English	0.361	0.323	0.593	9.82	0.607	

Unless otherwise mentioned in Remarks, the system (open configuration) used the roughly 160,000 translation pairs of the C-STAR multilingual corpus in each language pair, and the evaluation was performed with 16 references

20,000 sentences). In an example-based system, by essence, such expressions should be translated by a simple memory access.[8]

Again as far as data are concerned, we intended to limit ourselves to the use of the core 160,000 C-STAR translation pairs. However, this was not possible for the Arabic–English track where only 20,000 translation pairs were supplied. Consequently, a comparison of the Arabic–English results with other language pairs is not possible. We face another problem with the English–Chinese language pair: although the amount of data was 160,000 translation pairs as for other language pairs, evaluation was performed with only one reference whereas 16 references were used in all other pairs. It is well known that the number of references used enormously influences the scores in objective evaluation measures.[9] This prevents us from directly comparing the results.

The results obtained are shown in Table 7. Again for all language pairs, no tool of any sort was used, which means that prior to translation, no segmentation or tagging whatsoever was performed. No dictionary was added to the corpus of example sentences. In fact, the results of our system should be considered as a sort of baseline for all these language pairs in the C-STAR tracks.

To summarize, we are able to conduct a comparison only between the following language pairs: Korean–English, Chinese–English, and Japanese–English. The scores obtained in these three language pairs may be compared because the amount of linguistic data used as examples does not change. Only the source language changes while the target language remains English in all cases with the very same examples. The results in all three main evaluation scores (mWER, BLEU and NIST) show that the

[8] In participating in IWSLT 2005, we wanted to demonstrate the potential of our approach, rather than obtain the best possible numerical results. Our real goal was to compare results obtained in different language pairs. It would have been trivially possible for us to get excellent results with our system using a standard configuration and suitable data. Indeed the ultimate essence of an example-based system is to comprise a translation memory and this is the case with our system. The track we participated in was the so-called "C-STAR data track" for which it was formally specified that "[t]here are no limitations on the linguistic resources used to train the MT systems. Full BTEC corpus and proprietary data can be used" (`www.is.cs.cmu.edu/iwslt2005/eval_restrictions.html`). Almost all the test sentences could be found in our proprietary data, so that, with the standard configuration of our system, such data, and a minimum of computation, we actually got the following scores in a "false" run: mWER = 0.07, BLEU = 0.93, and NIST = 14.13.

[9] An epistemological remark: mWER, BLEU, NIST, and the like are often said to be measures for translation. Strictly speaking, this is not true. They are just families of measures. Only the given formulae of the metrics *plus* a set of reference sentences constitute a measure, not the formulae alone.

Table 8 Number of analogies in the BTEC multilingual corpus

Language	Number of analogies	Number of sentences involved
English	2,384,202	53,250
Japanese	1,910,065	53,572
Chinese	1,639,068	49,675
Korean	266,504	25,088

performance of our system is lower for Korean–English whereas the best performance is achieved in Japanese–English, with Chinese–English being somewhere in between.

In both the IWSLT 2004 and IWSLT 2005 tasks, our system's scores are lower in the Chinese–English track than in the Japanese–English track, an observation which also holds true for the other competing systems. One could possibly infer that the Chinese data allow for fewer commutations than the Japanese data.

In the case of the Korean language, an issue is that of encoding. The Hangul writing system uses one character to represent a syllable of the type CVC. Morphological commutations may take place within such a sequence. Relevant commutations should logically be sought at a scale lower than that of characters whereas we had our system working on the character level.

A more general interpretation of the results is that, in the view of our approach, the scores obtained by our system may well be interpreted as a measure of the "systematicity" of the data contained in the linguistic resources used. In this view, our scores are consistent with the fact that the C-STAR BTEC is usually believed to be internally more homogeneous in Japanese than in Chinese, which is in turn usually believed to be more homogeneous than in Korean. This impression is confirmed by the statistics of Table 8, which gives the number of formal analogies present in each language part of the C-STAR BTEC. According to these statistics, Chinese exhibits fewer analogies than Japanese. In Korean, the number of sentences involved in at least one analogy is nearly half the number of sentences involved in other languages, which implies a much lower number of analogies in comparison with the other languages: roughly one eighth on average. There may be several reasons for this. First, the Korean data may not be as homogenous and consistent as the other languages as they seem to have been produced by different people using quite different levels of language for similar situations. Second, as we said above, our method may miss commutations in Korean by relying on the character unit. Third, and in accordance with the previous point, Korean is known to be much richer morphologically than Japanese or English (not to mention Chinese) so that much more textual data should be logically needed to reflect the same amount of commutations in meaning.

6.4 Choice and influence of the data

In a third experiment, we evaluated the influence of adding or substracting data on the performance of our system. The test set used consists of 510 input sentences from the same domain as the parallel aligned corpus. Sixteen translation references in the target language were used for evaluation. As the data are all known to us in the experiment, we were able to determine a baseline and the upper bound for them.

The "gold standard" was determined 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

Table 9 Scores for the Gold Standard, the baseline, and the system with various data

System	Transl. pairs	mWER	mPER	BLEU	NIST	GTM
Gold standard	(N/A)	0.00	0.00	1.00	14.95	0.91
+ Source + target paraphrases	438,817	0.46	0.42	0.50	**8.98**	0.67
+ Target paraphrases	318,668	0.47	0.43	0.49	8.91	0.67
+ Source paraphrases	369,822	**0.38**	**0.35**	0.53	8.53	**0.68**
+ Dictionary	206,382	0.39	0.36	**0.54**	8.54	**0.68**
Resource only	158,409	0.39	0.36	0.53	8.53	**0.68**
1/2 resource	81,058	0.50	0.45	0.45	7.78	0.63
1/4 resource	40,580	0.53	0.49	0.42	7.18	0.60
Baseline: translation memory	158,409	0.58	0.53	0.38	7.54	0.61

system. In this way, we obtained the "best" values for each of the measures considered (see Table 9).

The baseline was determined 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, which we evaluated with each of the objective measures. This gives baseline scores for each of the measures considered.

Our system was then evaluated on its outputs for the sentences of the test set, with the sole resource of our 158,409 translation pairs (see Table 9, line: Resource only). Again, the evaluation was performed using the first candidates only, that is, those with the highest output frequencies.

6.4.1 Influence of the amount of examples

In an EBMT system, one would trivially expect the amount and nature of examples to influence translation quality strongly. The figures in Table 9 on the lines marked "1/2 resource" and "1/4 resource" confirm this fact. They were obtained by sampling the original resource. In this case, the more data, the better the results.

6.4.2 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 multiword 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, in this system, there is no difference between a parallel aligned corpus or a dictionary as long as both are aligned strings of data, be they sentences or words. The examples in (40) illustrate that the data format for a parallel aligned corpus or a dictionary does not differ in any way.[10]

(40) a. フィルムを買いたいのですが。*firumu wo kaitai no desu ga.*
FILM obj BUY-WANT polite
↔ I'd like a roll of film, please.

[10] Once again, the transcription and literal gloss are provided for the reader's benefit, and are not part of the stored data.

b. 三十六枚撮りを二本ください。 *sanju-roku-mai-dori wo nihon kudasai.*
36 TAKE TWO-ROLLS PLEASE
↔ Two rolls of thirty-six exposure film, please.
c. このカメラの電池がほしいのです。
kono kamera no densi ga hoshi n desu.
THIS CAMERA gen BATTERY subj WANT affirm
↔ I'd like a battery for this camera, please.
d. 電池 *firumu* ↔ film
映画 *eiga* ↔ film ('movie')
電池 *densi* ↔ battery
砲台 *hōdai* ↔ battery ('gun 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 score.

6.4.3 Paraphrases generated from the resource as additional examples

Previous research has shown that the introduction of paraphrases may improve the quality of MT 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 (that is, they are paraphrases). In our parallel aligned corpus, an average of 3.03 paraphrases per source sentence was obtained. However, the distribution is not uniform: 71,192 sentences (out of 103,274) do not receive any new paraphrase, while 54 sentences receive more than 100 paraphrases, with a maximum of 410 paraphrases for one sentence.

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 will the equation $A : B :: x : D$ 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 is shown in Table 9 on the line marked "+ Source paraphrases". They show a slight improvement in mWER.

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 target languages are shown on the line marked "+ Source + target paraphrases". They are not better than those with the resource only, except for NIST, as paraphrases are expected to have introduced lexical and syntactic variation in expressing identical meanings. An explanation for the loss in quality according to all other measures may be that the increase in computation may have overloaded the system (all experiments are performed with the same time-out).

7 Discussion and future work

7.1 Learning and lazy processing

In opposition to MT methods that "eagerly compile input samples and use only the compilations to make decisions" (Aha 1998, p. 13) our method "perform[s] less precompilation and use[s] the input samples to guide decision making" (*idem.*). In this sense, the system presented here may be seen as a "lazy learning" system (Aha 1997).

There is indeed an extra feature in our system: it learns as it keeps translating. As it appears from the description in Sect. 3.1, the system increases its knowledge by recursive calls because it adds new translation pairs to the parallel aligned corpus, so that, in a normal setting, the history of translations influences the results of subsequent translations. However, in all experiments reported above, we had to disable this feature so as to be placed in conditions comparable with (say) SMT systems. Of course, such a use disadvantages our system.

7.2 Translation time

It could have been feared that the complexity of the algorithm, which is basically quadratic in the amount of data, would have enormously impaired the method. However, using a simple heuristic (see Sect. 3.2) 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 seconds on a 2.8 GHz processor machine with 4 Gb memory.

7.3 Proportion of successful analogies

As the fundamental operation in the system is analogy, we measured the proportion of analogical equations successfully solved over the total number of analogies formed in the source language. On average, 687,641 analogical equations are formed in translating one sentence from the test set. The proportion of analogical equations successfully solved is 28%. In other words, in comparison with an ideal heuristic that would select only those pairs that lead to a solution, the current heuristic used to select sentence pairs from the corpus in order to form analogical equations is successful only a quarter of the time. Reaching 100% may be unattainable in practice but future work should include finding a heuristic that would increase this proportion so as to reduce the number of unnecessary trials.

7.4 Recursion level needed

As was explained in Sect. 3.1, 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 from a corpus.

7.5 Relevance/suitability of the examples

The translation of an input sentence depends crucially on the two following points: first, whether the input sentence belongs to the domain (and the style) of the corpus of examples, and second, 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.

First, as was mentioned in Sects. 6.4.2 and 6.4.3, we are studying various ways to add paraphrases or dictionaries and how to improve their efficiency in terms of lexical and syntactic variation, so as to further improve the coverage of the parallel aligned corpus.

Second, we are investigating the possibility of designing a core grammar by examples, that is, 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" (Salkoff 1973, Sager 1981) are well known for having a wide coverage.

8 Conclusion

In this paper, we have shown that the use of a specific operation, namely proportional analogy, leads to reasonable results in MT 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 were obtained by adding paraphrases.

The use of an operation that suits by essence the specific nature of linguistic data, namely their capacity for 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 neutralizes them implicitly. As a consequence, the implemented system does not include any transfer component (either lexical or structural).

To summarize, 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 patterns, does not have any explicit transfer component, and does not require any training or preprocessing of the aligned examples, a knowledge that is, of course, indispensable.

Acknowledgements The research reported here was supported in part by a contract with the Japanese National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus". Both authors were at the time of writing with the Japanese National Institute of Information and Communications Technology (NiCT). We are particularly indebted to Prof. C. Boitet for his many comments on an earlier version of the draft that considerably helped to improve clarity. Thanks also to the reviewers who pointed out some errors in the draft.

References

Aha DW (1997) Editorial. Artif Intell Rev 11:7–10

Aha DW (1998) Feature weighting for lazy learning algorithms. In: Liu H, Motoda H (eds) Feature extraction, construction and selection: A data mining perspective, Kluwer, Dordrecht, The Netherlands, pp. 13–32

Akiba Y, Federico M, Kando N, Nakaiwa H, Paul M, Tsujii J (2004) Overview of the IWSLT04 evaluation campaign. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 1–12

Allison L, Dix TI (1986) A bit string longest common subsequence algorithm. Inform Proc Lett 23:305–310

Amores JG, Mora JP (1998) Machine translation of motion verbs from English to Spanish. In: Martín-Vide C (1998) pp 191–206

Aramaki E, Kurohashi S (2004) Example-based machine translation using structual translation examples. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 91–94

Bender O, Zens R, Matusov E, Ney H (2004) Alignment templates: The RWTH SMT system. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp. 79–84

Bertoldi N, Cattoni R, Cettolo M, Federico M (2004) The ITC-irst statistical machine translation system for IWSLT-2004. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, 51–58

Blanchon H, Boitet C, Brunet-Manquat F, Tomokiyo M, Hamon A, Hung VT, Bey Y (2004) Towards fairer evaluations of commercial MT systems on basic travel expressions corpora. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, 21–26

Bloomfield L (1933) Language. Holt. New York, NY

Brown PE, Della Pietra VJ, Della Pietra SA, Mercer RL (1993) The mathematics of statistical machine translation: Parameter estimation. Comput Ling 19:263–311

Carl M (1998) A constructivist approach to machine translation. In: Proceedings of the joint conference on new methods in language processing and computational natural language learning, NeMLaP3/CoNLL98. Macquarie University, [Sydney, Australia], pp 247–256

Carl M (2006) A system-theoretic view of EBMT. Mach Translat 19:147–167

Carl M, Way A (eds) (2003) Recent advances in example-based machine translation. Kluwer Academic Publishers, Dordrecht, The Netherlands

Claveau V, L'Homme M-C (2005) Apprentissage par analogie pour la structuration de terminologie—Utilisation comparée de ressources endogènes et exogènes [Analogical learning of terminological structure—Comparison of the use of endogenous and exogenous resources]. In: TIA 2005: 6èmes rencontres terminologie et intelligence artificielle. Rouen, France, p 10

Damper RI, Eastman JEG (1996) Pronouncing text by analogy. In: COLING-96: The 16th international conference on computational linguistics. Copenhagen, Denmark, pp 268–269

Delhay A, Miclet L (2004) Analogical equations in sequences: Definition and resolution. In: Paliouras G, Sakakibara Y (eds) Grammatical inference: Algorithms and applications, 7th international colloquium, ICGI 2004. Springer, Berlin, Germany, pp 127–138

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: ARPA workshop on human language technology notebook proceedings. San Diego, CA, pp 139–145

Dorr BJ (1994) Machine translation divergences. Comput Ling 20:597–633

Dorr BJ, Pearl L, Hwa R, Habash N (2002) DUSTer: A method for unraveling cross-language divergences for statistical word-level alignment. In: Richardson S (ed) Machine translation: From research to real users (Fifth conference of the Association for Machine Translation in the Americas AMTA-2002. Tiburon, CA, USA, …), Springer, Berlin, pp 31–43

Gentner D (1983) Structure mapping: A theoretical model for analogy. Cognitive Sci 7:155–170

Habash N (2002) Generation-heavy hybrid machine translation. In: Proceedings of the international natural language generation conference (INLG'02). New York, NY, pp 185–191

Harris ZS (1954) Distributional structure. Word 10:146–162

Harris Z (1982) A grammar of English on mathematical principles. J Wiley, New York, NY

Hathout N (2001) Analogies morpho-synonymiques: Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes [Morpho-synonymic analogies: A method of automatically acquiring morphological links starting from a synonym dictionary]. In: TALN-Récital 2001: 8ème conférence annuelle sur le traitement automatique des langues

naturelles et 5ème rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues. Tours, France, pp 223–232
Hofstadter D, Fluid Analogies Research Group (1994) Fluid concepts and creative analogies. Basic Books, New York, NY
Hou H, Deng D, Zou G, Yu H, Liu Y, Xiong D, Liu Q (2004) An EBMT system based on word alignment. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 47–49
Hutchins J (2006) Example-based machine translation: A review and commentary. Mach Translat 19:116–130
Ilie L (1998) On ambiguity in internal contextual languages. In: Martín-Vide C (1998) pp 29–46
Itkonen E (1999) Grammaticalization: Abduction, analogy, and rational explanation. In: Shapiro M, Haley M (eds) The Peirce seminar papers: Essays in semiotic analysis vol IV, Berghahn Books, Oxford, England, pp 159–175
Joshi A, Vijay-Shanker K, Weir D (1991) The convergence of mildly context-sensitive grammar formalisms. In: Sells P, Shieber SM, Wasow T (eds) Foundational issues in natural language processing, MIT Press, Cambridge, MA, pp 31–81
Lee Y-S, Roukos S (2004) IBM spoken language translation system evaluation. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 39–46
Lepage Y (1998) Solving analogies on words: An algorithm. In: COLING-ACL '98: 36th annual meeting of the Association for Computational Linguistics and 17th international conference on computational linguistics. Montreal, Quebec, Canada, pp 728–735
Lepage Y (2001) Analogy and formal languages. In: Proceedings of the joint meeting of the sixth conference on formal grammar and the seventh conference on the mathematics of language (FG/MOL 2001). Helsinki, Finland, pp 1–12
Lepage Y (2003) De l'analogie rendant compte de la commutation en linguistique [On analogy considering commutation in linguistics]. Mémoire d'habilitation à diriger les recherches. Université de Grenoble, Grenoble, France
Lepage Y (2004) Lower and higher estimates of the number of "true analogies" between sentences contained in a large multilingual corpus. In: Coling: 20th international conference on computational linguistics. Geneva, Switzerland, pp 736–742
Lepage Y, Peralta G (2004) Using paradigm tables to generate new utterances similar to those existing in linguistic resources. In: Proceedings of the fourth international conference on language resources and evaluation (LREC-2004). Lisbon, Portugal, pp 243–246
Levenshtein VI [Левенштейн, ВИ] (1965) Двоичные коды с исправлением выпадений, вставок и замещений симболов. Докл Акад Наук СССР 163:845–848; appeared (1966) as Binary codes capable of correcting deletions, insertions and reversals, Sov Phys Dokl 10:707–710
Martín-Vide C (ed) (1998) Mathematical and computational analysis of natural language. John Benjamins, Amsterdam, The Netherlands/Philadelphia, PA
Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: A method for automatic evaluation of machine translation. In: 40th annual meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, pp 311–318
Paul H (1920) Prinzipien der Sprachgeschichte [Principles of the history of language]. Niemayer, Tübingen, Germany
Polański K (1984) Słownik syntaktyczno-generatywny czasowników polskich [Syntactic-generative dictionary of Polish verbs]. Wydawnictwo im. Ossolińskich, Warszawa
Reichert J, Waibel A (2004) The ISL EDTRL system. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, 61–64
Sager N (1981) Natural language information processing: A computer grammar of English and its applications. Addison-Wesley, Reading, MA
Salkoff M (1973) Une grammaire en chaîne du français [A string grammar of French]. Dunod, Paris, France
Sasayama M, Ren F, Kuroiwa S (2003) Super-function based Japanese-English machine translation system. In: NLPK-KE 2003: International conference on natural language processing and knowledge engineering. Beijing, China, pp 555–560
Sato S (1991) Example-based machine translation. PhD thesis, Kyoto University, Kyoto, Japan
de Saussure F (1955) Cours de linguistique générale [A course in general linguistics]. Payot, Lausanne, Switzerland
Shieber SM (1985) Evidence against the context-freeness of natural language. Ling Philos 8:333–343
Skousen R (1989) Analogical modeling of language. Kluwer, Dordrecht, The Netherlands
Stephen GA (1994) String searching algorithms. World Scientific, Singapore

Stroppa N, Yvon F (2005) An analogical learner for morphological analysis. In: CoNLL-2005: Ninth conference on computational natural language learning. Ann Arbor, MI, pp 120–127
Sumita E (2003) EBMT using DP-matching between word sequences. In: Carl M, Way (2003) pp 189–209
Sumita E, Akiba Y, Doi T, Finch A, Imamura K, Okuma H, Paul M, Shimohata M, Watanabe T (2004) EBMT, SMT, hybrid and more: ATR spoken language translation system. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 13–20
Sumita E, Iida H (1991) Experiments and prospects of example-based machine translation. In: 29th annual meeting of the Association for Computational Linguistics. Berkeley, CA, pp 185–192
Takezawa T, Sumita E, Sugaya F, Yamamoto H, Yamamoto S (2002) Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In: LREC 2002: Third international conference on language resources and evaluation. Las Palmas de Gran Canaria, Spain, pp 147–152
Thayer I, Ettelaie E, Knight K, Marcu D, Munteanu DS, Och FJ, Tipu Q (2004) The ISI/USC system. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 59–60
Turian JP, Shen L, Melamed ID (2003) Evaluation of machine translation and its evaluation. In: MT Summit IX: Proceedings of the ninth machine translation summit. New Orleans, USA, pp 386–393
Ukkonen E (1985) Algorithms for approximate string matching. Inform Control 64: 100–118
Vogel S, Hewavitharna S, Kolss M, Waibel A (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
Wagner RA, Fischer MJ (1974) The string-to-string correction problem. J Assoc Comput Mach 21:168–173
Wu D (1997) Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Comput Ling 23:377–403
Yamamoto K (2004) Interaction between paraphraser and transfer for spoken language translation. 自然言語処理 J Nat Lang Proc 11.5:63–86
Yang M, Zhao T, Liu H, SHi X, Jiang H (2004) Auto word alignment based Chinese-English EBMT. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 27–29
Zuo Y, Zhou Y, Zong C (2004) Multi-engine based Chinese-to-English translation system. In: Proceedings of the international workshop on spoken language translation. Kyoto, Japan, pp 73–77