The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign

Yves Lepage and Adrien Lardilleux

GREYC, University of Caen Basse-Normandie, France
Yves.Lepage@info.unicaen.fr, Adrien.Lardilleux@etu.info.unicaen.fr

Abstract

The GREYC machine translation (MT) system is a slight evolution of the ALEPH machine translation system that participated in the IWLST 2005 campaign. It is a pure example-based MT system that exploits proportional analogies. The training data used for this campaign were limited on purpose to the sole data provided by the organizers. However, the training data were expanded with the results of sub-sentential alignments. The system participated in the two classical tasks of translation of manually transcribed texts from Japanese to English and Arabic to English.

1. Introduction

This paper gives a sketch of the GREYC machine translation system that participated in the IWSLT 2007 evaluation campaign. This system is an evolution of the ALEPH machine translation system that participated in the IWSLT 2005 evaluation campaign [1]. In particular, the core engine has been redesigned. It is still basically the same example-based machine translation system, but it now adds an alignment preprocessing phase.

This system participated in the two classical tasks of IWLST 2007: the translation of read speech from Japanese into English and from Arabic into English in the so-called clean conditions, *i.e.*, manually transcribed texts. Our main goal is to assess how much can be achieved by example-based techniques using the minimum amount of data and to show progress year after year. In this respect, we did not use any data outside of the training data sets provided by the organizers: 40,000 aligned utterances in Japanese-English and 20,000 in Arabic-English. No use was made of the development sets and no use was made of any extra resources or tools like the ones listed on the IWSLT 2007 resources page.[1]

It was shown in previous experiments and during a previous participation to IWSLT that the system was unable to translate in the absence of sufficient data because the core engine did not compile the data in any respect. 20,000 aligned sentences were shown to be too small to get any translation, and consequently, for the previous participation in the Japanese-English IWSLT 2005 track, 140,000 extra aligned sentences from the BTEC corpus were used in addition to the training data provided by the organizers. [2] reports a fall in BLEU from 0.53 with a training corpus of 160,000 aligned sentences to 0.42 when using only a quarter of the same training corpus. The challenge we addressed this year was precisely to cope with the provided data alone. Our answer was to inflate the training data by adding sub-sentential alignment data obtained from the training data using example-based techniques. An example in Section 4.3 will explain why sub-sentential alignments may increase the coverage of sentences translated by analogy.

[1] http://iwslt07.itc.it/menu/resources.html

2. Preprocessing of the data

We applied two basic preprocessing methods on the training data as well as on the test data. They are directly related to the way the core engine works. Thirdly, we also preprocessed punctuation for alignment purposes, but not for the translation itself.

2.1. Constant-length encoding

First, our translation engine uses proportional analogy on strings of characters as a fundamental operation, whatever the character set is [3]. The only requirement is that the encoding for the same language be constant-length in byte number. This is why we had to convert all data in each of the languages to the same constant-length encoding. The data were provided in UTF-8, an encoding that possibly mixes different byte-lengths. For instance, an English word in the middle of a Japanese sentence is coded in UTF-8 with one-byte long characters, while the surrounding Japanese characters are encoded in two-bytes. This is a problem for the core engine. For reasons of visualisation, so as to be able to verify our data at each further processing step, we decided to convert the UTF-8 data into different constant-length encodings that we are easily able to display. The encodings we adopted were: EUC-JP for Japanese (a 2-byte encoding), ISO-8859-6 for Arabic (thus a 1-byte encoding for that language), and ASCII for English (1-byte).

2.2. Sentence splitting

Second, previous experiments have confirmed the intuition that analogies are relatively much more important in number with short utterances [4]. Consequently, in order to increase

Table 1: Japanese word and chunk counts.

	Training set	Test set	In test set but not in training set
Words	12,534	895	35
Chunks	42,721	982	209

the possibilities of the system, we split those utterances composed of several sentences into their individual sentences. To do so consistently, both the source and target utterances have to comprise the same number of sentences. In addition it is better to verify which sentence in the source language corresponds to which sentence in the target language. This check was performed using in-house alignment methods.

2.3. Punctuation normalization

A third preprocessing we applied on the training data only was to separate all punctuations from the words they are adjoined to, in order to make them similar to words. This is a preprocessing for the alignment method, not for the translation as will be explained below. In this way, we were able to consistently align punctuations in the different languages (see the example of the question mark in Table 3). As a detail, because there were no commas in the test set, we deleted any comma in the training set before any other preprocessing.

Reciprocally, in Arabic, the data presented inconsistencies concerning punctuation. For instance, question marks are either separated from the previous word by a space (preferred convention in the Maghreb) or by nothing (more frequent in the Mashriq). The provided data contained both notations and we normalized the data according to the Maghreb convention.

3. Expanding the training data with sub-sentential alignments

3.1. Chunking

As a preprocessing step for sub-sentential alignment, we segmented the Japanese and Arabic texts into chunks. The writing systems in the two languages and also the type of the data provided led to different preprocessing methods.

For Japanese, the data provided a segmentation into words for each utterance. Such a segmentation seems to have been performed by ChaSen.[2] Standard Japanese usually does not insert any space between words. We could thus compute a dictionary of words for both the training set and the test set. The counts are to be found in Table 1. We also performed an experiment in specifically translating unknown words from Japanese into English with the set of words from the training set. This kind of experiment has been reported in [5] and [6]. As a result, we could correctly generate half of them (18 words out of 35).

As for the chunking method, it is based on markers. Indeed, Japanese is a language with case markers, a closed set of words (or morphemes) appearing at the end of chunks. Our list of markers consisted of eight nominal case markers, one verbal ending and one punctuation mark.[3] Some data concerning the number of chunks obtained as well as sequences of chunks for the Japanese training data are given in Table 2.

The case of Arabic is still different from English and Japanese. The Arabic writing system composes the root of a word with a set of proclitic and enclitic affixes, like conjunctions, the article or pronouns, into an agglutinate form that is separated from other similar agglutinate forms by spaces. Consequently, a form separated by two spaces in Arabic corresponds to some extent to the notion of a chunk in English, but with some differences. For instance, the English chunk *the black cat* will spell as two separate agglutinate forms in Arabic: *the-cat the-black*. For simplicity, we kept agglutinate forms as the basic unit (which we abusively call a `chunk') for Arabic in the alignment process. Some statistics about the number of such agglutinate forms in the Arabic training set are given in Table 3. Relative to the specificities of the Arabic writing system, we should point out that we did not call on any morphological analysis to disambiguate agglutinate forms. This is in accordance with our decision not to use any extra resources. To summarize, we used the agglutinate forms as they are in the training set and in the test set, and considered such forms as a unit of processing for sub-sentential alignment (recall that the unit of processing for the translation engine is the character).

3.2. Sub-sentential alignment using hapaxes

In order to increase the possibilities of the translation engine, we inflated the training data with sub-sentential alignments obtained from the training data themselves.

Previous research showed that hapaxes (words that appear only once in a corpus) were responsible for the majority of the best alignments using the cosine method [7]. Indeed, given a pair of aligned sentences, if a source word from the source sentence and a target word from the target sentence are both hapaxes in their respective corpora, and if they are the *only* hapaxes on the aligned sentences at hand, then the resulting word alignment is almost guaranteed to be correct.

Such results allowed us to design a new alignment method, where the string to be aligned is artificially made a hapax in a new subset of the corpus. Given a pair of aligned sentences from which two words in the source and the target languages are to be aligned, the basic problem is to determine if there exists a subset of lines from the bilingual corpus such that both words are hapaxes in their respective language, and

[2] `http://chasen.naist.jp/hiki/ChaSen/`.

[3] The eight nominal case markers are の /*no*/ (genitive), で /*de*/ (instrumental or location), へ /*e*/ (direction), に /*ni*/ (dative or location), を /*wo*/ (accusative, *i.e.*, object), は /*wa*/ (topic), が /*ga*/ (subject), から /*kara*/ (origin). The verbal ending is the past ending ました /*-masita*/. The punctuation is the symbol corresponding to fullstop.

Table 2: Some statistics on the Japanese-to-English sub-sentential alignments.

	Total number	Average number of words in English (Mean ± Std.dev.)
Words	4,097	3.93 ± 1.84
Chunks	12,582	4.51 ± 2.39
Sequences of 2 chunks	31,695	5.80 ± 2.44
Sequences of 3 chunks	39,758	6.85 ± 2.56
Total	88,132	

Table 3: Some statistics on the Arabic-to-English sub-sentential alignments.

	Total number	Average number of words in English (Mean ± Std.dev.)
Chunks	2,843	2.47 ± 1.33
Sequences of 2 chunks	11,558	3.35 ± 1.39
Sequences of 3 chunks	35,268	7.84 ± 2.74
Total	49,669	

that they be the only hapaxes in the considered pair of aligned sentences. If such a subset does not exist, the pair cannot be aligned and thus will get the worst possible score. If it does, then the two words are potentially good translations.

The method was applied to chunks and sequences of chunks in Japanese and Arabic. We obtained 84,035 Japanese-English alignments and 49,669 Arabic-English alignments, thus tripling the size of the training data available, as the sizes of the original training data were 40,000 aligned utterances in Japanese-English and 20,000 in Arabic-English. In addition we also aligned Japanese words with English words (see above). Tables 2 and 3 summarize the number of alignments obtained relatively to the length of the sequences of chunks. Some examples of alignments are shown in Tables 5, 6 and 7.

4. The translation engine

4.1. The principle

The translation engine used in this experiment is similar to the one described in [2]. It is an example-based engine that also adopts a translation memory approach. For a new sentence to translate, the engine first looks for the sentence in the training data. If the sentence is found in the training set, the translation engine just outputs its translation without any further computation. This is the most felicitous case. This was the case for 188 Japanese sentences and 34 Arabic sentences in the test data.

If the sentence does not already exist in the training data, then computation is carried out using the principle of corresponding proportional analogies between two language domains. Refer to [3] for a thorough presentation of proportional analogies between strings of symbols and some properties in formal language theory. To translate sentence A, the engine basically solves all possible analogical equations of the type:

$$A : x :: C : D$$

where C and D are two source sentences from the training data. If the solution of the equation $x = B$ belongs to the training data, then its translation $\hat{B}$ is known and the analogical equation:

$$y : \hat{B} :: \hat{C} : \hat{D}$$

can be built and possibly solved in the target language. Any solution $y = \hat{A}$ to this equation is a possible translation of A. Figure 1 illustrates the principle with actual data.

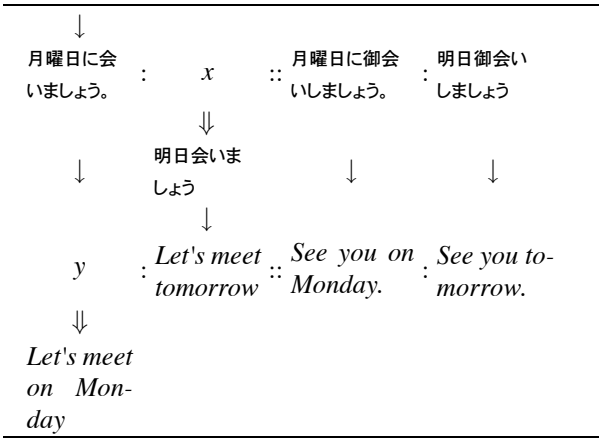

Figure 1: A sketch of the principle of corresponding proportional analogies on actual data. The Japanese sentence on the left is the input. The second Japanese sentence on the same line is from the training data. The Japanese chunk on the right is from the chunk alignments. Solving the Japanese equation yields the Japanese chunk in the middle, which, incidentally, belongs to the list of aligned chunks. From the corresponding translations, a candidate translation for the input is obtained in place of y. Notice the missing fullstop.

When no solution at all can be found by analogy, then, the engine backs off to the basic behavior of a translation memory, *i.e.*, it outputs the translation of the source sentence closest to the input sentence.

4.2. Two improvements

A first difference with the procedure used in the IWSLT 2005 campaign is in the type of data used. All the data are no longer entire sentences. As seen above in the Chunking and Sub-sentential alignments sections, we expanded the training data with sub-sentential alignments. Obviously, this drastically increases the number of cases where D may be a subsequence in C. In this way, it is expected that more commutations between B and D are revealed in A and C. To this

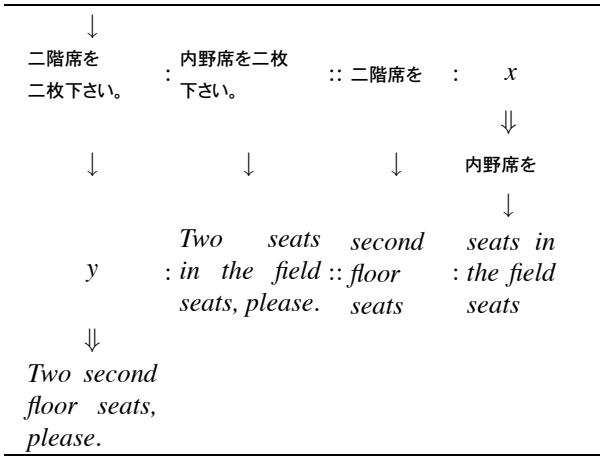

Figure 2: An actual example of the application of the extra heuristic: notice the difference with Figure 1 for the place of the unknown x in the analogical equation in Japanese.

end and because trying all possible pairs of (C,D) is just impossible, we selected the C's and the D's using the following heuristic. The corpus was sorted relatively to A in two ways: by distance (edit distance) to A for the selection of C's and by inclusion score (length of D minus its similarity to A) for the selection of D's. All the notions involved, like distance, length and similarity, refer to classical notions on character strings and can be efficiently implemented.

A second difference with the engine used in the IWSLT 2005 campaign is the use of a heuristic to increase the number of successfully solved analogical equations. In addition to the previous analogical equation in the source language, the following equation is also tried:

$$A : B :: C : x$$

where B is close to A, and C is well included in D (we use the same sorted lists as mentioned just above). Figure 2 gives an actual example. In case the solution $x = D$ of the analogical equation belongs to the training data, the following equation can be built in the target language:

$$y : \hat{B} :: \hat{C} : \hat{D}$$

This heuristic has proved to be productive thanks to the expansion of the training data with sub-sentential alignments.

4.3. The benefit of sub-sentential alignments

We now illustrate the benefit of sub-sentential alignments in the previous framework with a simple example. Suppose we want to translate the input sentence *I'd like two coffees, please.* with a corpus composed of the two following aligned sentences only:

I'd like a cup of coffee, please. コーヒーを下さい。
Two coffees, please. コーヒーを二つ下さい。

As analogies require three sentences in the source domain, the method is simply unable to compute any translation.

Suppose now that we are able to extract the following sub-sentential alignments from the previous data:

a cup of coffee コーヒーを
two coffees コーヒーを二つ
, please 下さい

Then, it becomes possible to form the following analogical equation in the source domain: *I'd like two coffees, please.* : *I'd like a cup of coffee, please.* :: *two coffees* : *a cup of coffee*, which corresponds to a proportional analogy in the target domain: y : コーヒーを下さい。 :: コーヒーを二つ : コーヒーを , whose resolution delivers a correct translation for the input sentence:

I'd like two coffees, please. コーヒーを二つ下さい。

This example shows that increasing the data with sub-sentential alignments makes it possible for the system to translate sentences that it was unable to translate before.

4.4. Time-out in number of analogies tried

For the translation of each input sentence a limited number of one million analogies is allowed so as to limit computational time. This number is passed with the `-time-out` parameter to the core engine. When this number is exceeded, the translation process stops and a translation memory back-off solution is output: the translation of the sentence closest to the input sentence in the training data.

4.5. Filtering of incorrect character sequences

Because the core engine works with character strings, we have to limit the production of incorrect sequences of characters in English during the production of new character strings. For that, we use a filtering technique which consists in rejecting any new character string that contains a sequence of n characters absent from the training data. In this campaign, we set n to 4. This is parameter `-target-ngram` for the actual program.

4.6. Translation parameters

To summarize, the parameters passed to the core engine were the following for the two translation tasks. For Japanese-to-English:

```
-bicorpus=training-data
          +word-alignments
          +chunk-alignments
          +seq-of-2-chunks-alignments
-source-B=2
-time-out=1000000
-target-ngram=4
```

The parameter `-source-B` sets the number of bytes to 2 for the source language, here Japanese. It should be stressed

Table 4: A comparison of the BLEU scores obtained during the 2005 and 2007 IWSLT evaluation campaigns.

Year	Task	Training data size	Number of references	BLEU scores
2005	ja-en	160,000	16	0.593
2007	ja-en	40,000	6	0.396
2005	ar-en	20,000	16	0.382
2007	ar-en	20,000	6	0.329

that we deleted any space in the Japanese data as well in the training data as in the sub-sentential alignments.

For Arabic-to-English the parameters were similar, the main difference being that we did not have word alignments:

```
  -bicorpus=training-data
            +chunk-alignments
            +seq-of-2-chunks-alignments
-time-out=1000000
-target-ngram=4
```

5. Postprocessing

As this year translation outputs were required to be in conformity with standard English writing conventions (case sensitive and punctuation) we restored a proper punctuation style. For instance, we introduced a punctuation sign at the end of a sentence when there was none, and deleted any sentence-final markers in the middle of a sentence (the case of 35 sentences translated from Japanese and also 35 sentences from Arabic). We also capitalized the first letter of a sentence when it did not appeared so (the case of 19 sentences translated from Japanese and 30 from Arabic).

As a final postprocessing, we put back multiple sentences in the same utterance to conform to the format of the test data (this was the case of 30 sentences in the Japanese test set, and 35 in the Arabic test set).

6. Results

The BLEU scores obtained are 0.396 in the Japanese to English task and 0.329 in Arabic to English task. According to the organizers, the BLEU scores were obtained using 6 references.

In IWSLT 2005 [8], the scores obtained by a similar system were 0.593 in Japanese-English, and 0.382 in Arabic-English with a number of references of 16. As the number of references decreases, the values of the scores somehow mechanically decrease. But the amount in decrease depends on the references themselves and nothing can really be said when the test sets are different. If the amount of training data was the same for Arabic-English in both years, 20,000 aligned utterances, the size of the training set was quite different for Japanese-English: 160,000 aligned utterances in 2005 versus 40,000 in 2007.

From all these facts, it is difficult to compare the results obtained. Only in the case of Arabic to English can it be said that the results obtained seem comparable in the two campaigns. In the case of Japanese to English we feel uneasy in drawing any clear interpretations.

7. Conclusion

This paper has given a sketch of the GREYC machine translation system that participated in the IWSLT 2007 evaluation campaign in the two proposed classical tasks: the translation of manually transcribed sentences from Japanese into English and from Arabic into English.

The system did not use any data outside of the training data provided by the organizers: 40,000 aligned utterances in Japanese-English and 20,000 in Arabic-English. However, sub-sentential alignments were obtained from the training data themselves and allowed to triple the size of the aligned example data.

As a whole, the system is an example-based system that relies on the correspondence of analogies between two languages. In this sense, it is an evolution of the ALEPH machine translation system that participated in the IWSLT 2005 evaluation campaign. It is nevertheless difficult to compare the results obtained this year with those of 2005 for several reasons: size of the training data, number of references used for the computation of the BLEU scores, etc. Despite these differences, the results seem comparable for Arabic to English while no clear conclusion can be drawn for Japanese to English.

8. References

[1] Y. Lepage and E. Denoual, ``Aleph: an EBMT system based on the preservation of proportional analogies between sentences across languages," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, Pittsburgh, PA., Oct. 2005, pp. 47--54. [Online]. Available: http://www.slt.atr.co.jp-/∼lepage/pdf/iwslt2005.pdf.gz

[2] ------, ``Purest ever example-based machine translation: detailed presentation and assessment," *Machine Translation*, vol. 19, pp. 251--282, 2005.

[3] Y. Lepage, ``Analogy and formal languages," *Electronic notes in theoretical computer science*, vol. 47, pp. 180--191, Apr. 2004.

[4] Y. Lepage, J. Migeot, and E. Guillerm, ``Analogies of form between chunks in Japanese are massive and far from being misleading," in *Proceedings of the 4th language and technology conference*, Poznan, 2007 (to appear).

[5] P. Langlais and A. Patry, ``Translating unknown words by ananlogical learning," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2006, pp. 877--886. [Online]. Available: http://www.aclweb.org/anthology/D/D07/D07-1092

[6] E. Denoual, ``Analogical translation of unknown words in a statistical machine translation framework," in *Proceedings of Machine Translation Summit XI*, Copenhagen, September 2007.

[7] A. Lardilleux and Y. Lepage, ``The contribution of the notion of hapax legomena to word alignement," in *Proceedings of the 4th language and technology conference*. Poznan, 2007 (to appear).

[8] T. Eck and C. Hori, ``Overview of the IWSLT 2005 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*. Pittsburgh, PA., 2005, pp. 1--22.

Table 5: Examples of sub-sentential alignments for Arabic chunks, sequences of 2 Arabic chunks and sequences of 3 Arabic chunks.

؟	?
كلا	No
سياتل	Seattle
سأرحل	I 'm getting off
البحيرة ؟	is the lake
اثنين عشرة	two ten
ألو .	Hello .
يفتح المصرف	does the bank open
رحلة إلى لندن	a flight to London
إنني مسرور لحضورك	I 'm glad you could come
للأمام ، وهو	Go straight ,
لي بالتحدث إليك	Can I talk to

Table 6: Examples of alignments for Japanese words.

万年筆	a fountain pen
平土間	the stalls
開き	do you open
隠れろ	Keep hidden .

Table 7: Examples of sub-sentential alignments for Japanese chunks, sequences of two Japanese chunks and sequences of three Japanese chunks.

もちろんごゆっくりどうぞ	Sure , take your time .
ピンは	sell pins
一時間ほどで	about an hour
英国航空です	, this is British Airways .
今空いていますか	the toilet vacant now
すぐに御持ち致します	I 'll bring one right away .
ダウンタウンの地図は	maps of the downtown area
では明日迄に	Well then , we can get it done tomorrow .
野ウサギが欲しいのです	like hare .
このバスは動物園迄	the zoo ?