Objective evaluation of the analogy-based machine translation system ALEPH

Yves Lepage & Etienne Denoual
ATR 音声言語コミュニケーション研究所
619-0288「けいはんな学研都市」光台 2-2-2
{yves.lepage,etienne.denoual}@atr.jp

1 Introduction

The ALEPH machine translation system (Lepage and Denoual, 2005) is an example-based machine translation system that 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. Its characteristic is that it relies solely on the use of a linguistically justified operation, proportional analogy.

2 Comparison with other systems

We assessed the ALEPH system using the IWSLT 2004 tasks in both Japanese-English and Chinese-English directions. As we used a bicorpus of 160,000 examples (the C-STAR Basic Traveler's Expressions Corpus or C-STAR BTEC (Takezawa et al., 2002)) 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, the ALEPH system did not make any use of any NLP tool such as a tagger or the like 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 delete segmentation in the provided test sets! This clearly demonstrates that segmentation is not a necessity to perform a translation task from Japanese or Chinese. As for data, no dictionary was used. The C-STAR corpus of around 160K aligned sentences was used for both language pairs. We refer to this as our 'training data', although there is absolutely no training phase within the ALEPH framework.

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 the ALEPH 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 are given below. The numbers on the left of a translation candidate are the frequencies with which it has been output. We assumed that the most frequent candidate should be the most reliable one, so that the evaluation was performed on the first candidates only.

	席はありますか。
27634	Can I have a table?
27634	Can we have a table?
27628	Do you have any seats?
	ライター用の詰めかえをください。
2	Give me a refill for a lighter, please.
	ここに行きたいんですけど。
11	Here's where I would like to go.
2	This is where I want to go to.
2	This is where I want to go.

Tables 1 and 2 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 ALEPH system achieves second place in Chinese-English, and third place in Japanese-English. A standout point is the achievement in BLEU: 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 the ALEPH system. It must be stressed again that the above results were obtained without any training performed in advance on the data, and that no tuning whatsoever of the system towards the 'training data' was performed.

3 Comparison for different language pairs

The 2006 IWSLT 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 the evaluation procedure was also the same. Our goal was to learn some lessons on the difficulty of translating some language pairs relatively to others with our proposed method. As one configuration only was allowed, we chose to use the *open* configuration of the ALEPH system because it seemed the most honest attitude to inspect the potentialities 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. On 506 sentences to translate, 90 did in fact belong to the training set (and even to the supplied data of 20,000 sentences)! In an example-based system, by essence, such expressions should be translated by a mere memory access.

Again as far as data are concerned, we limited 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.

The results obtained are shown in Table 3. 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 the ALEPH system should be considered as a sort of baseline for all these language pairs in the C-STAR tracks.

We have already said that because we used only 20,000 translation pairs in Arabic-English, we are not able to compare with other language pairs. 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. This prevents us from comparing the results.

To summarise, we are only able to conduct a comparison 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 performance of the ALEPH system is lower for Korean-English whereas the best performance is achieved in Japanese-English, Chinese-English being in the middle.

In both the IWSLT2004 and the IWSLT2005 tasks, the ALEPH system's scores in BLEU and NIST are lower in the Chinese-English track than in the Japanese-English track, an observation which seem to hold true for the other competing systems. One could possibly infer that the Chinese data allow for less 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 whithin such a sequence. Relevant commutations should logically be sought at a scale lower than that of characters whereas we had the ALEPH 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 the ALEPH 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 statistics that gives the number of formal analogies present in each language part of the C-STAR BTEC. According to these statistics, Chinese exhibits less 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 in average. There may be several reasons for this. Firstly, the Korean data may not be so homogenous and consistent as for the other languages as they seem to have been produced by different people using quite different levels of language for similar situations. Secondly, as we said above, our method may miss commutations in Korean by relying on the character unit. Thirdly, 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.

4 N-Best list

The ALEPH system delivers a list of translation candidates that are ordered according to the number of times each candidate has been produced, *i.e.*, their frequency of output. As we said above, in the previous evaluation settings, we used only the candidate with the highest frequency of output for each sentence to translate. In a new experiment, we inspected the gain obtained when choosing among different translation candidates. The data we used in this experiment were, for each sentence to be translated, the N-best list of candidates with $N = 3$ for the IWSLT 2005 JE C-STAR track.

Let us consider one of the objective measures, say mWER. For each test sentence, we evaluated the mWER score of each of its 3 possible translations and selected the one which delivers the best score. We then gathered these sentences to form a new set of 506 test translations that we evaluated according to all possible objective measures.

We performed the same for the two other measures BLEU and NIST, that are usually considered to reflect fluency and adequacy respectively. The results are shown in Table 4.

These results show that an amelioration of 7% in BLEU, 9% in NIST and 10% in mWER is obtained when we choose the best candidate in the N-best list of translation candidates.

5 Conclusion

We have shown that the use of a specific operation, namely proportional analogy, leads to reasonable results in machine translation without any preprocessing of the data whatsoever, an advantage over techniques requiring intensive preprocessing.

It is the use of an operation that suits by essence the specific nature of linguistic data, *i.e.*, their capacity of commutation on the paradigmatic and syntagmatic axes, that allowed us to dispense with any preprocessing of the data whatsoever.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Yves Lepage and Etienne Denoual. 2005. The purest EBMT system ever built: no variables, no templates, no training, examples, just examples, only examples. In *Proceedings of the Workshop on Example-based machine translation, hosted by MT Summit X*, pages 81–90, Phuket, Thailand, September.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002*, pages 147–152, Las Palmas, May.

Table 1: Scores for the IWSLT 2004 Chinese-to-English *Unrestricted Data* track: no restriction on linguistic resources.

	mWER	mPER	BLEU	NIST	GTM
[s]ISL-S	**0.379**	**0.319**	**0.524**	**9.56**	**0.748**
[e]ALEPH *standard*	0.434	0.400	0.522	8.42	0.687
[e]ALEPH *open*	0.437	0.404	0.512	8.24	0.682
[s]IRST	0.457	0.393	0.440	7.24	0.671
[s]IBM	0.525	0.442	0.350	7.36	0.684
[h]ISL-E	0.531	0.427	0.275	7.50	0.666
[s]ISI	0.573	0.499	0.243	5.42	0.602
[h]NLPR	0.578	0.531	0.311	5.92	0.563
[e]HIT	0.594	0.487	0.243	6.13	0.611
[r]CLIPS	0.658	0.542	0.162	6.00	0.584
[e]ICT	0.846	0.765	0.079	3.64	0.386

Table 2: Scores for the IWSLT 2004 Japanese-to-English *Unrestricted Data* track: no restriction on linguistic resources.

	mWER	mPER	BLEU	NIST	GTM
[h]ATR-H	**0.263**	**0.233**	0.630	10.72	0.796
[s]RWTH	0.305	0.249	0.619	**11.25**	**0.824**
[e]ALEPH *standard*	0.324	0.300	**0.634**	9.19	0.731
[e]ALEPH *open*	0.437	0.403	0.534	8.97	0.697
[e]UTokyo	0.485	0.420	0.397	7.88	0.672
[r]CLIPS	0.730	0.597	0.132	5.64	0.568

Table 3: Scores for all IWSLT 2005 C-STAR tracks.

	mWER	mPER	BLEU	NIST	GTM	*Remarks*
English-Chinese	0.798	0.746	0.098	3.029	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	

Table 4: Scores for the IWSLT 2005 JE C-STAR track with an oracle on 3-best candidates list.

	mWER	mPER	BLEU	NIST	GTM	METEOR
IWSLT 2005 results	0.361	0.323	0.593	9.82	0.607	0.720
best mWER	**0.325**	0.300	0.634	9.86	0.620	0.734
best BLEU	0.343	0.311	**0.638**	10.34	0.617	0.732
best NIST	0.349	0.315	0.627	**10.50**	0.610	0.727