Robust Large-Scale EBMT with Marker-Based Segmentation

Nano Gough and Andy Way
National Centre for Language Technology
School of Computing
Dublin City University
Dublin 9, Ireland
{ngough,away}@computing.dcu.ie

Abstract

Previous work on marker-based EBMT [Gough & Way, 2003, Way & Gough, 2004] suffered from problems such as data-sparseness and disparity between the training and test data. We have developed a large-scale robust EBMT system. In a comparison with the systems listed in [Somers, 2003], ours is the third largest EBMT system and certainly the largest English-French EBMT system. Previous work used the on-line MT system *Logomedia* to translate source language material as a means of populating the system's database where bitexts were unavailable. We derive our sententially aligned strings from a *Sun* Translation Memory (TM) and limit the integration of *Logomedia* to the derivation of our word-level lexicon. We also use *Logomedia* to provide a baseline comparison for our system and observe that we outperform *Logomedia* and previous marker-based EBMT systems in a number of tests.

1 Introduction

It is widely acknowledged that MT systems are more suitable for domain-specific applications. When the training data is restricted to a particular sublanguage and the testset is also tuned to this domain, there is less margin for error.

The size of the example base is dependent on the system in question and the objectives of its developers. A large example base reduces the problems associated with data-sparseness and has been proven in some instances to improve translation performance [Sumita & Iida, 1991, Sato, 1993, Mima *et al*., 1998].

[Way & Gough, 2004] extend the research of [Gough & Way, 2003] on integrating controlled language data in an EBMT system. They translate a set of controlled English documents derived from *Sun* computer manuals using the on-line system *Logomedia*.[1] They acknowledge that the use of *Logomedia* in constructing an example base is not ideal, but justify the use of the on-line system given the current absence of controlled bitexts. *Logomedia* was selected as it was deemed to be the better of the three on-line systems tested in [Way & Gough, 2003]. The testset is extracted from a *Sun* TM, which while not written according to controlled language (CL) specifications, addresses the same sublanguage area. The Marker Hypothesis [Green, 1979] is applied to produce additional lexical resources which are then used to train the EBMT system.

Following a number of subsequent improvements to the system, [Way & Gough, 2004] report an overall improvement in the average Bleu score over [Gough & Way, 2003] and show that their EBMT system

[1] http://www.logomedia.net

outperforms the on-line MT system *Logomedia*. They find that although the training data and the testset are of a similar domain, the disparity between the data may be extensive enough to reduce the overall Bleu score. Data sparseness was also a major factor in reducing translation performance. Although the system obtained 100% coverage, many translations were generated word for word, as few chunk matches were located. [Way & Gough, 2004] suggest that applying the Marker Hypothesis to an extended example base should reduce data-sparseness. They also suggest that a training set with less dependence on *Logomedia* would yield improved translation results.

With this in mind, we have developed a robust EBMT system which uses the Marker Hypothesis to derive linguistic resources from a large-scale *Sun* TM. We restrict the use of *Logomedia* to the generation of our word-level lexicon. The memories of our system are populated with automatically extracted resources derived from over 200,000 sentence pairs. In a number of experiments using a testset of 3,939 sentences, randomly extracted from the *Sun* TM, we show that our system considerably outperforms *Logomedia* and previous marker-based EBMT systems according to a number of automatic and human evaluation metrics.

The rest of this paper is structured as follows. In section 2, we describe relevant previous research in the area of marker-based EBMT and refer to other scalable EBMT systems. We also present our large-scale EBMT system. In section 3, we report on a number of experiments carried out to test the system. We use automatic evaluation metrics to assess the quality of the translations produced and also provide a manual evaluation on a smaller dataset. Finally, we conclude and outline some potential areas for future research.

2 Example-based Machine Translation (EBMT)

EBMT systems translate new input with recourse to a set of <source, target> sub-sentential resources extracted from a bitext. There are many different methods of sub-sentential alignment in EBMT including, deriving transfer rules from examples [Furuse & Iida, 1992], generalisation by syntactic category [Kaji *et al*., 1992] and generalisation by semantic features [Matsumoto & Kitamura, 1995].

Another method which has met with some success in recent research [Gough *et al*., 2002, Way & Gough, 2003] uses sets of marker-words to segment the source and target sentences to derive additional sub-sentential resources.

2.1 Marker-Based EBMT

The 'Marker Hypothesis' [Green, 1979] is a universal psycholinguistic constraint, which states that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes. A number of EBMT systems have used the Marker Hypothesis as a basis for translation including METLA [Juola, 1994], *Gaijin* [Veale & Way, 1997], and the *wEBMT* system [Gough *et al*., 2002, Way & Gough, 2003]. This 'linguistics-lite' approach was also used as a basis for work on controlled EBMT [Gough & Way, 2003, Way & Gough, 2004].

We define six sets of marker words for English and French and assign these to categories <DET>, <PREP> etc. These are then used to segment the <source, target> aligned sentences. As an example, consider the strings in (1) appearing in the *Sun* TM:

(1) you click apply to view the effect of the selection ⇒ vous cliquez sur appliquer pour visualiser l'effet de la sélection

Marker-based segmentation is applied in a pre-processing stage. The <source, target> strings in the sentential database are traversed word by word and automatically tagged with their marker categories, as in (2):

(2) <PRON> you click apply <PREP> to view <DET> the effect <PREP> of <DET> the selection ⇒ <PRON> vous cliquez <PREP> sur appliquer <PREP> pour visualiser <DET> l' effet <PREP> de <DET> la sélection

A new fragment begins where a marker word is encountered and ends at the occurrence of the next marker word. In addition, we impose a further constraint that each chunk must contain at least one non-marker word. This restriction is implemented in (2) when generating the last chunk in both the English and French strings. For example, in the English string, a new chunk begins at *of*. Although the next word *the* is also defined as a marker word, it does not indicate the beginning of a new chunk as this would leave the chunk beginning with *of* without any content word. From the tagged strings in (2), the marker chunks in (3) are generated:

(3) a. <PRON> you click apply : vous cliquez sur appliquer
b. <PREP> to view : pour visualiser
c. <DET> the effect : l'effet
d. <PREP> of the selection : de la sélection

In [Gough *et al*., 2002, Gough & Way, 2003, Way & Gough, 2003] these marker lexicons are predicated on the naïve yet effective assumption that marker-headed chunks in the source map sequentially to their target equivalents, subject to the source and target strings having the same number of marker tags and their marker categories matching. We generate a marker-lexicon using the improved sub-sentential alignment algorithm of [Way & Gough, 2004], which enables much more data to be retained. This method checks that chunks are marked with similar tags but also uses a base-dictionary created via *Logomedia* to check for word-equivalences between chunks. Along with cognate matches, these word-equivalences are used to predict chunk alignment. The more lexical equivalences which can be established between chunks, the more likely these chunks are to produce a correct alignment. The position of chunks can also be used to predict alignments – the more distance between two chunks, the less likely they are to align.

Previously it was only possible to produce 1:1 alignments, whereas this algorithm facilitates the merging of chunks to produce 2:1 and 3:1 alignments. As an example, consider the tagged strings in (2). The Marker Hypothesis segments the English sentence into four chunks and the corresponding French sentence into five chunks. Given the differing number of chunks, this sentence pair would not be considered for sub-sentential alignment under the method of [Gough *et al*., 2002, Gough & Way, 2003, Way & Gough, 2003]. Using the method of [Way & Gough, 2004], the base-dictionary generated via *Logomedia* can be used to establish lexical equivalences between <*click, cliquez*>, <*apply, appliquer*>, <*view, visualiser*>, <*effect, effet*> and <*selection, sélection*>. The final three chunks in the source and target can also be linked with their marker tags: <PREP>, <DET>, <PREP>. The first two chunks in the French sentence, <PRON> *vous cliquez* and <PREP> *sur appliquer* can be merged as they both share lexical equivalences with the first chunk in the English sentence <PRON> *you click apply*. In this way the marker chunks in (3) are derived.

Further lexical information can be extracted from the marker chunks in (3). We take advantage of the fact that chunks containing just one non-marker word in both source and target are assumed to be translations of each other. In this way we can extract the 'word-level' translations in (4)

(4) <PREP> to : pour <LEX> view : visualiser <LEX> effect : effet <PRON> you : vous <DET> the : l' <PREP> of : de

More general examples can add flexibility to the matching process and improve coverage. In a final pre-processing stage, we produce a set of marker templates by replacing marker words with their associated tags. For example, from the entries in (3), the templates in (5) can be generated:

(5) a. <PRON> click apply : <PRON> cliquez sur appliquer
b. <PREP> view : <PREP> visualiser
c. <DET> effect : <DET> effet
d. <PREP> selection : <PREP> sélection

Using the templates in (5) it is now possible to insert any marker word after the relevant tag if it appears with its translation in the lexicon. As an example, consider the translation of *an effect*. Assuming this string cannot be located in the marker lexicon, it is generalised to <DET> *effect* in this process. The generalised lexicon is then searched and assuming this string is now located, its translation <DET> *effet* can be retrieved. The final translation can be produced by inserting the translation of *the* in the place of <DET>. Of course, it is likely that the word-level lexicon will contain multiple translations for *the* and several erroneous strings will be produced alongside the correct translation. Weightings are calculated for each translation according to the formula in (6) and the same *n*-gram search method as in [Gough & Way, 2003, Way & Gough, 2004] is used to derive translations. As the 'best' translation is ranked first in 92% of cases, the task of identifying the most accurate translation from a set of candidates is much simplified.

(6)

$$weight = \frac{\textit{no. occurrences of the proposed translation}}{\textit{total no. of translations produced for source language chunk}} \quad (1)$$

The weight for the complete translation is calculated by multiplying the weights of each chunk making up the translation. The lexical resources deduced using the Marker Hypothesis are considerably larger than those used to train the system in [Gough & Way, 2003, Way & Gough, 2004]. In the next section, we discuss the potential advantages of increasing the example base.

2.2 Scalability

In his overview of EBMT, [Somers, 2003] lists a number of systems in terms of the size of the example base used. A number of experiments report an improvement in translation quality based on augmentation of the example database. The work of [Sumita & Iida, 1991] and [Sato, 1993] promotes a positive link between increasing the example base and improving translation quality. [Mima *et al*., 1998] report an improvement of 35% on translation accuracy following a continuous incrementation of the example base in measures of 100 examples. However, they also suggest that there may be a ceiling to this pattern where adding further examples will not improve translation quality.

Indeed, the adverse effects of an increased example base cannot be overlooked. Where a number of similar examples occur these can mutually reinforce each other and this can be manipulated by the system. Some systems [Somers *et al*., 1994, Öz & Cicekli, 1998, Murata *et al*., 1999] use a similarity metric in their matching algorithm so that a higher score can be applied to more frequently occurring examples. However, if such a metric is not used this can create ambiguity and/or result in overgeneration of certain linguistic phenomena.

We output our translations with an associated weight calculated according to the formula in (**??**). As shown in [Way & Gough, 2003], our system consistently ranks good translations more highly than bad ones, and the 'best' translation by human standards is always to be found in the top 1% of candidate translations. Accordingly, we do not consider the presence of other similar training examples to be detrimental to the quality of translations output by our system.

2.2.1 Scalability in Marker-based EBMT

[Gough & Way, 2003, Way & Gough, 2004] train their EBMT system on a set of *Sun* documents translated from English into French via the on-line system *Logomedia*. They extract their testdata from a much

larger *Sun* TM. The authors acknowledge that this is an unusual approach but aim to filter the data using CL specifications. This led to problems with data-sparseness and somewhat reduced the quality of the translations produced. In this paper, we do not use controlled language data or apply *Logomedia* to produce our sententially aligned strings. Instead, we train our system on a portion of the uncontrolled *Sun* TM and test our system on another portion of the same resource. In this way, we not only increase the amount of training data but also heighten the similarity between the training data and the testset. We use *Logomedia* only to produce our word-level lexicon and as a baseline comparison.

In a comparison with the systems listed in [Somers, 2003], this is by far the largest English/French EBMT system and certainly the largest marker-based EBMT system. Although, the data used in the *wEBMT* system [Gough *et al*., 2002, Way & Gough, 2003] contained over 200,000 English-French phrases, no sententially aligned pairs existed in the example base. We use the improved sub-sentential alignment algorithm of [Way & Gough, 2004] to align chunks derived via the Marker Hypothesis. We weight these alignments favouring frequently occurring source/target chunks. Using automatic evaluation metrics we assess the impact of a scalable EBMT system on translation quality. We provide a comparison with existing marker-based systems and also with the on-line MT system *Logomedia*.

3 Translation Experiments and Evaluation

In this section we report on a number of experiments carried out to test the system. The *Sun* TM contains 207,468 sententially aligned English-French pairs. For both languages, we randomly extracted 3,939 sentences (ave. sentence length for English 13.2 words, min. 1 word, max. 87 words; for French, ave. sentence length 15.7 words, min. 1 word, max. 91 words) from the TM as a testset, and as in [Gough & Way, 2003, Way & Gough, 2004] ensured that each unique word in the testset was contained somewhere within the training data. We used the remaining 203,529 sentences as training data.

We segmented these English-French aligned pairs using the Marker Hypothesis (cf. section 2.1), and added the sub-sententially aligned fragments to the marker lexicon. We then generalised the sub-sentential alignments and extracted entries for our word-level lexicon. As in [Gough & Way, 2003, Way & Gough, 2004], we used the on-line MT system *Logomedia* to translate any words that could not be assigned to the word-level lexicon via this method.

Initially we translated the 3,939 sentence testset in both language directions. In order to quantify the effect of exact matches on the Bleu score, we then extracted the non-exact matches from the English testset and translated these. Finally, we performed a novel filtering of the marker-lexicon to remove incorrect alignments and assessed the impact of this process on the quality of English translations. In the following sections, we present both an automatic and a human evaluation of the translations produced by the system and compare these to previous figures obtained using marker-based EBMT. We provide results for *Logomedia* as a baseline comparison.

3.1 Automatic Evaluation

[Gough & Way, 2003] and [Way & Gough, 2004] calculated IBM Bleu scores for the translations produced by their systems using the NIST MT Evaluation Toolkit.[2] They also calculated Bleu scores for *Logomedia* on the same testset. [Gough & Way, 2003] report that when Bleu was utilised, *Logomedia* appears to considerably outperform their EBMT system. However, there was far less disparity between the two systems in a manual evaluation, indicating that the Bleu metric was quite harsh. [Way & Gough, 2004] incorporate some novel refinements to the sub-sentential alignment algorithm of [Gough & Way, 2003] (cf. section 2.1)

[2] http://www.nist.gov/speech/tests/mt/mt2001/index.htm

and apply two simple improvements to the EBMT system. They report a 104% improvement in the average Bleu score over [Gough & Way, 2003] and obtain a Bleu score 0.66% higher than *Logomedia* when evaluating the same data. [Way & Gough, 2004] suggest that increasing the example base may improve the Bleu score as the testset is considerably smaller than the training data used in the experiment.

In this paper, our training data contains 203,529 sentences. This is a significant increase over [Gough & Way, 2003] and [Way & Gough, 2004], where the number of sententially aligned pairs in the example base was just over 1,600. We apply the Marker Hypothesis to these aligned pairs. In [Way & Gough, 2004], 85% of sentence pairs threw up candidates for sub-sentential alignment. For the larger training set this figure is 69.7%. Nevertheless, the total number of unique sub-sentential alignments derived is 275,822, a considerable increase over [Way & Gough, 2004] where even with the refined alignment algorithm just over 3,000 sub-sentential alignments were generated.

3.1.1 Evaluating Translations (En-Fr / Fr-En)

Using our augmented training set and testing on more similar data, the average Bleu score for our system on French translations is 0.3435. This is an improvement of 154% over [Way & Gough, 2004]. An improvement of 42% is also noted in the Bleu score obtained for English translations. These figures are presented in Table 1:

System *(En-Fr)*	**Bleu**	**System *(Fr-En)***	**Bleu**
Our System	0.3435	Our System	0.2419
Logomedia	0.1292	Logomedia	0.1677
Way/Gough 04	0.1352	Way/Gough 04	0.1703

Table 1: Comparing our large-scale EBMT system with *Logomedia* and [Way & Gough, 2004] using the IBM Bleu automatic evaluation metric on a 3939 sentence testset

Based on the average Bleu score, our system outperforms the on-line MT system *Logomedia* by 73.4% from English to French and 44.2% from French to English. Figures for Precision and Recall and word and sentence error rates are also calculated using the tools reported in [Turian *et al*., 2003].[3] The scores obtained for our system and for *Logomedia* are presented in Table 2.

System	**Precision**	**Recall**	**WER**	**SER**
Our System *(En-Fr)*	0.5318	0.6648	66.7	88.3
Logomedia	0.4500	0.4766	79.7	98
Way/Gough 04	0.3891	0.5293	64.8	84
Our System *(Fr-En)*	0.4730	0.6879	83.2	93.8
Logomedia	0.4420	0.5838	96.5	95.4
Way/Gough 04	0.3005	0.3646	80.1	88

Table 2: Comparing our large-scale EBMT system with *Logomedia* and [Way & Gough, 2004] using automatic evaluation metrics

The Bleu scores in Table 1 and the Precision and Recall figures in Table 2 suggest that our enhanced system presented in this paper outperforms *Logomedia* in both directions (French-English and English-French). The translations produced by our enhanced system presented in this paper also obtain better word and sentence error rates. Similarity between the training data and the testset is likely to account for this.

[3] http://nip.cs.nyu.edu/GTM/

Furthermore, as *Logomedia* is a general-purpose system, it does not have recourse to all the domain-specific vocabulary present in our training data. [Way & Gough, 2004] report a higher Bleu score when translating from French to English. Their figures for Precision and Recall and WER/SER conflict with this result as they are higher when translating from English to French. All figures obtained for our enhanced system via automatic metrics suggest that it performs better when translating from English to French.

3.1.2 Evaluating Translations for non-exact matches (En-Fr)

Bleu may favour the similarity of the testset to the training data. For 35.1% of the sentences in our English testset an exact match can be found in the training data. To this end, we eliminated these sentences from the testset. We then obtained an average Bleu score for the translations produced for the remaining sentences. For comparison, we also obtained an average Bleu score for the translations of these same sentences via *Logomedia*. These results are shown in Table 3.

System	Bleu	Precison	Recall	WER	SER
Our System	0.1766	0.4831	0.4696	73.5	98.9
Logomedia	0.1419	0.4772	0.4890	80	98.7

Table 3: Comparing our large-scale EBMT system (English-French:exact matches eliminated) with *Logomedia* using Automatic Evaluation metrics

Although the average Bleu score for our enhanced system presented in this paper falls to 0.1766 when translating from English to French, it is still higher than *Logomedia* at 0.1419. This indicates that in terms of Bleu score we still outperform *Logomedia*, albeit less significantly when exact matches cannot be found. We also outperform *Logomedia* with regard to Precision and WER. We do slightly worse than *Logomedia* with regard to Recall and SER. It should be noted that *Logomedia* obtains a higher Bleu score in this particular experiment than it does when translating the entire testset.

3.1.3 Filtering the Data (En-Fr)

Although the refined sub-sentential alignment method of [Way & Gough, 2004] improves on that of [Gough *et al*., 2002, Way & Gough, 2003, Gough & Way, 2003], some incorrect alignments remain in the system's databases. In a final experiment, we perform a filtering of the data generated using the Marker Hypothesis. We translate each English chunk in our marker lexicon via *Logomedia*. We then compare each translation with the corresponding French chunk in our lexicon. The comparison is length-based. We eliminate all aligned chunks that differ by more than one word in length from the translation produced by *Logomedia*. For example, the incorrect chunk alignment *your password : votre mot* appears in the marker lexicon. When *your password* is translated via *Logomedia*, its translation is *votre mot de passe*. As *votre mot* and *votre mot de passe* differ in length by more than one word, the faulty alignment is eliminated from the marker-lexicon.

In our initial English-French experiment, 275,822 unique aligned chunks were produced. This figure now falls to 134,752 when the data is filtered, a loss of over 51%. The number of unique word alignments produced in our first experiment was 2828. This figure falls by 10.5% to 2531. The number of generalised templates is reduced by 49.6%.

Any words not present in the word-level lexicon were translated via *Logomedia*. The same testset used in the first English-French experiment was then translated using the filtered lexical resources. The results are presented in Table 4. The results for the first experiment and for *Logomedia* are included for comparison.

Using the filtered data the overall Bleu score for our enhanced system increases by 17.9%. Figures for Precision and Recall also improve by 11.9% and 6.5% respectively.

System	Bleu	Precison	Recall
Our System (original)	0.3435	0.5318	0.6648
Our System (filtered)	0.4049	0.5953	0.7081
Logomedia	0.1292	0.4500	0.4766

Table 4: Comparing our large-scale EBMT system (English-French) with *Logomedia* using our filtered data

3.2 Manual Evaluation

In addition to these automatic evaluations, we also performed a manual evaluation using the notions of intelligibility and accuracy. Accuracy measures how faithfully the translation represents the source. Intelligibility depends on the number of grammatical errors or mistranslations in the string. The purpose of the manual evaluation was to provide a more detailed analysis of the effect which filtering the data has on translation quality. We also wanted to compare our enhanced system with *Logomedia* in an effort to confirm the findings of the automatic evaluation.

As in [Gough & Way, 2003, Way & Gough, 2004], we measured accuracy on a 5-point scale. 'Score 4' is attributed to a very accurate translation which represents the source faithfully and 'Score 0' to a completely inaccurate translation. Scores for Intelligibility are defined at four levels, from 'Score 3' (a very intelligible translation with no syntactic errors) to 'Score 0' (an unintelligible translation). We randomly extracted 50 French translations produced for non-exact matches. A native English speaker with good French language competence carried out a manual evaluation on these strings. The translations for the same strings produced using the filtered data were then evaluated for comparison. The results for Accuracy are given in Table 5.

System	Score 0	1	2	3	4
Our System (original)	4	2	4	22	18
Our System (filtered)	2	2	0	10	36
Logomedia	0	4	20	6	20

Table 5: Comparing our large-scale EBMT system (English-French) with *Logomedia* in a Human Evaluation: Accuracy

Using our original data, 80% of the translations produced by our enhanced system presented in this paper obtain a score of 3 or 4. The same scores are only assigned to 52% of translations produced by *Logomedia* for the same strings. Using the filtered data our enhanced system considerably outperforms *Logomedia*. The number of translations with score 3 or 4 increases to 92%. The improvement in translation quality using the filtered data was also noted in the automatic evaluation and can be attributed to the overall impact of the filtering process on the ranking of the 'best' translation. Only the highest ranked translation is submitted for evaluation using automatic metrics. Of the sentences evaluated manually, the 'best' translation was ranked first in 26% of cases using the original data. When the filtered data was integrated, the 'best' translation was ranked first in 92% of cases.

The results for Intelligibility are given in Table 6. Using the original data *Logomedia* slightly outperforms our enhanced system. 92% of its translations obtain a score of 2 or 3. This figure is slightly lower for our enhanced system at 88%. However, using the filtered data our enhanced system also produces 92% intelligible translations.

System	Score 0	1	2	3
Our System (original)	0	6	16	28
Our System (filtered)	0	4	12	34
Logomedia	0	4	22	24

Table 6: Comparing our large-scale EBMT system (English-French) with *Logomedia* in a Human Evaluation: Intelligibility

4 Conclusions

In this paper we presented an EBMT system which uses the Marker Hypothesis to induce additional lexical resources from a large-scale sententially aligned example base. We applied the marker-based sub-sentential alignment algorithm of [Way & Gough, 2004] to a training set containing 203,529 pairs. As far as we are aware, this is the largest English-French EBMT system in existence, and is certainly the largest marker-based EBMT system for any language pair.

We show that by increasing our example base and heightening the similarity between the training and test data, our enhanced system can outperform a good on-line MT system such as *Logomedia* using automatic metrics. We also compare our results to other marker-based EBMT systems [Gough & Way, 2003, Way & Gough, 2004] which were trained on a far smaller dataset and relied on *Logomedia* to generate the target strings. We find that using automated evaluation metrics our large-scale system outperforms *Logomedia* by 165.9% *(En-Fr)* and 44.2% *(Fr-En)* on a 3,939 sentence testset. The Bleu score obtained for our enhanced system presented in this paper is higher than other marker-based systems (154% *(En-Fr)*), and (42% *(Fr-En)*). We also outperform previous systems of this type in terms of figures for Precision, Recall, WER and SER. By eliminating all exact matches from our testset we observe an anticipated deterioration in these figures but continue to outperform *Logomedia*.

In a final experiment, we performed a novel filtering of the training data. Although the size of the marker-based lexicon is reduced by over 50% and the generalized lexicon and the word-level lexicon are also pruned, we note a 19% improvement in translation quality on the same testset according to the Bleu score. Precision and Recall figures increase by 11.9% and 6.5% respectively. A manual evaluation carried out on a number of translations also indicates that filtering the data in this way leads to improved translation performance and increases the number of instances where the 'best' translation is ranked first among a set of candidates. This suggests that a marker-based EBMT system can produce better translations when it is trained on a large-scale dataset and the quality of the induced marker lexicon is much improved.

In terms of further work, we aim to provide a more detailed evaluation of the filtering process and its effect on translation quality. We have also obtained TM data from IBM, covering 28 different language pairs. We intend to use these resources to extend our experiment to other language pairs, and are currently extending our marker-based EBMT system to Chinese. Finally, while it is generally acknowledged that one of the advantages of EBMT over SMT is that EBMT systems require far less training text, there is a widespread perception that as soon as larger training data is taken into account, SMT wins out. In our view, this is somewhat contentious, and certainly not proven, and we hope in the near future to shed some (much needed) light on this area by comparing our EBMT system with a high-quality statistical machine translation (SMT) system using the *Giza++* modelling tool. [Och & Ney, 2003][4]

[4] http://www.isi.edu/~och/GIZA++.html

Acknowledgements

This work was partially funded by an IBM Fellowship. Thanks are also due to three anonymous reviewers whose insightful comments served to improve this paper.

References

[Furuse & Iida, 1992] Furuse, O. and H. Iida. 1992. An Example-based Method for Transfer-Driven Machine Translation. In *TMI* (1992), pp.139–150.

[Gough *et al*., 2002] Gough, N., A. Way. and M. Hearne. 2002. Example-based Machine Translation via the Web. In S. Richardson (ed.) *Proceedings of AMTA-02*, Tiburon, CA., pp.74–83.

[Gough & Way, 2003] Gough, N. and A. Way. 2003. Controlled Generation in Example-based Machine Translation. In *MT Summit IX* (2003) New Orleans, LA., pp.133–140.

[Green, 1979] Green, T. 1979. The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18**:481–496.

[Juola, 1994] Juola, P. 1994. A Psycholinguistic Approach to Corpus-Based Machine Translation. In *CSNLP 1994*, Dublin, Ireland, [pages not numbered].

[Kaji *et al*., 1992] Kaji, H., Y. Kida. and Y. Morimoto. 1992. Learning Translation Templates from Bilingual Text. In *Coling* (1992), Nantes, France, pp.672–678.

[Matsumoto & Kitamura, 1995] Matsumoto, Y. and M. Kitamura. 1995. Acquisition of Translation Rules from Parallel Corpora. In R. Mitkov and N. Nicolov (eds.) *RANLP*, Amsterdam: John Benjamins, pp.405–416.

[Mima *et al*., 1998] Mima, H., H. Iida. and O. Furuse. 1998. Simultaneous Interpretation Utilizing Example-based Incremental Transfer. In *Coling-ACL* (1998), Montreal, Canada, pp.855–861.

[Murata *et al*., 1999] Murata, M., Q. Ma., K. Uchimoto. and H. Isahara. 1999. An Example-based Approach to Japanese-to-English Translation of Tense, Aspect, and Modality. In *TMI* (1999), Chester, UK, pp.66–76.

[Och & Ney, 2003] A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* **29**(1), pp. 19–51.

[Öz & Cicekli, 1998] Öz, Z. and I. Cicekli. 1998. Ordering Translation Templates by Assigning Confidence Factors. In *Proceedings of AMTA-98*, Langhorne, PA., pp.51–61.

[Sato, 1993] Sato, S. 1993. Example-based Translation of Technical terms. In *TMI* (1993), Kyoto, Japan, pp.58–68.

[Somers *et al*., 1994] Somers, H., I. McLean. and D. Jones. 1994. Experiments in Multilingual Example-based Generation. In *CSNLP 1994* Dublin, Ireland, [pages not numbered].

[Somers, 2003] Somers, H. 2003. An Overview of EBMT In M. Carl. and A. Way. (eds.) *Recent Advances in Example-based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.3—57.

[Sumita & Iida, 1991] Sumita, E. and H. Iida. 1991. Experiments and Prospects of Example-based Machine Translation. In *29th Annual Meeting of the ACL*, Berkeley, CA., pp.185–192

[Turian *et al*., 2003] Turian, J., L. Shen. and D. Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *MT Summit IX*, New Orleans, LA., pp.386–393.

[Veale & Way, 1997] Veale, T. and A. Way. 1997. *Gaijin:* A Bootstrapping, Template-Driven Approach to Example-based Machine Translation. In *RANLP-97*, Tzigov Chark, Bulgaria, pp.239–244.

[Way & Gough, 2003] Way, A. and N. Gough. 2003. *wEBMT*: Developing and Validating an Example-based Machine Translation System using the World Wide Web. *Computational Linguistics* **29**(3).

[Way & Gough, 2004] Way, A. and N. Gough. 2004. Example-based Controlled Translation In *Proceedings of the 9th Workshop of The EAMT*, Malta, pp.73–81.