Hybridity in MT: Experiments on the Europarl Corpus

Declan Groves and Andy Way

National Centre for Language Technology
School of Computing
Dublin City University
Dublin 9, Ireland
{dgroves,away}@computing.dcu.ie

Abstract

(Way & Gough, 2005) demonstrate that their Marker-based EBMT system is capable of outperforming a word-based SMT system trained on reasonably large data sets. (Groves & Way, 2005) take this a stage further and demonstrate that while the EBMT system also outperforms a phrase-based SMT (PBSMT) system, a hybrid 'example-based SMT' system incorporating marker chunks and SMT sub-sentential alignments is capable of outperforming both baseline translation models for French–English translation.

In this paper, we show that similar gains are to be had from constructing a hybrid 'statistical EBMT' system capable of outperforming the baseline system of (Way & Gough, 2005). Using the Europarl (Koehn, 2005) training and test sets we show that this time around, although all 'hybrid' variants of the EBMT system fall short of the quality achieved by the baseline PBSMT system, merging elements of the marker-based and SMT data, as in (Groves & Way, 2005), to create a hybrid 'example-based SMT' system, outperforms the baseline SMT and EBMT systems from which it is derived.

Furthermore, we provide further evidence in favour of hybrid systems by adding an SMT target language model to all EBMT system variants and demonstrate that this too has a positive effect on translation quality.

1 Introduction

Almost all research in MT being carried out today is corpus-based. Within this field, by far the most dominant paradigm is PB-SMT, but much important work continues to be carried out in EBMT. Until the recent work of (Way & Gough, 2005), no comparative studies of any flavour of SMT and EBMT had appeared in print. While this work demonstrated that the Marker-based EBMT system of (Way & Gough, 2005) was capable of outperforming a word-based SMT system, (Groves & Way, 2005) showed that the EBMT system was also capable of higher translation quality than a PB-SMT system constructed from freely available resources. However, perhaps more importantly for the MT research community as a whole, this paper also demonstrated that a novel hybrid 'example-based SMT' system incorporating marker chunks and SMT sub-sentential alignments was capable of outperforming both baseline translation models, for French–English.

In this paper, we continue that line of research by developing a new hybrid 'statistical EBMT' system which outperforms the equivalent baseline system and a hybrid 'example-based' SMT system which outperforms the baseline EBMT and SMT systems from which it is derived. Crucially, therefore, the most important message arising from our research is that MT developers need to combine aspects from both SMT and EBMT if further gains are to be made; that is, despite the obvious convergence of the two paradigms, the remaining differences between SMT and EBMT (Way & Gough, 2005) are crucial for improved system development.

Both (Groves & Way, 2005) and (Way & Gough, 2005) use training and test data derived from a corpus of *Sun Microsystems*' documentation, consisting of 203K

French–English sentence pairs (max. sentence length of 112 words for English, 134 words for French, with average sentence lengths of 10.85 words and 12.05 words respectively). In this paper, we switch to the French–English Europarl (Koehn, 2005) training and test sets, which are fast becoming the standard data in the field. Interestingly, and in contrast to the research of (Groves & Way, 2005), on these data sets, we show that the PBSMT system outperforms the EBMT system of (Way & Gough, 2005). We observe that the coverage of the marker chunks is much greater on the *Sun Microsystems* test set than on the Europarl test sets, indicating that many more 'close' and exact chunk matches were found by the EBMT system on the *Sun* corpus, while translation of the more heterogeneous Europarl data required more word-for-word translation. This is also reflected in subsequent experiments exploring the efficacy of the SMT and EBMT lexicons in dealing with the Europarl data. By making use of incremental training sets of 78K (1.49M words), 156K (2.98M words) and 322K sentence pairs (6.12M words) we also demonstrate clearly that adding more training data improves all system variants.

To investigate the effect of adding hybrid sub-sentential fragments, we seeded both baseline systems with chunks from the EBMT system and the PBSMT system (extracted following the method of (Och & Ney, 2003)) to create our various 'hybrid' EBMT and PBSMT systems. Again we performed translation in both language directions. Incorporating SMT word alignments into the EBMT system results in an average BLEU score increase of 2.9% for French–English and 3.32% for English–French. When merging all data resources available, for French–English, we see an average relative improvement of 19.06% BLEU score for our hybrid EBMT system and 6.37% for our hybrid PBSMT system over their equivalent baselines, and 15.9% (EBMT) and 6% (PBSMT) for English–French.

In addition, we integrate an SMT target language model with the EBMT system of (Way & Gough, 2005), in a somewhat similar fashion to (Bangalore, Murdock, & Riccardi, 2002) who select the final translation output from multiple candidates, based on what fits a posterior trigram language model best, and thus improved results (although no actual figures identifying the contribution of the language model are given). We too demonstrate the positive effect that this has on translation quality. This we take to be further evidence in favour of our hypothesis that translation quality can be improved by combining aspects of SMT and EBMT.

The remainder of this paper is organised as follows: we briefly outline relevant previous research in the area of hybrid data-driven MT in section 2. In section 3, we describe the basic ideas behind EBMT, and detail the EBMT system used in these experiments. We summarise the main principles of SMT, and describe the techniques we use to derive phrasal alignments in section 4. Section 5 presents a series of experiments, including a description of the data resources, the performance of the baseline phrase-based SMT and EBMT systems, and the improved performance of the different hybrid systems. Finally we conclude, summarising our novel hybrid 'statistical EBMT' system, and our contribution to the area of data-driven MT in general, together with some avenues for further research in this area.

2 Related Work

While not directly related to the work we present here, there exists a body of work which merges translation memory (TM) resources with SMT. (Vogel & Ney, 2000) automatically derive a hierarchical TM from a parallel corpus, comprising a set of transducers encoding a simple grammar. In a similar manner, (Marcu, 2001) uses an SMT model (Brown et al., 1993) to automatically derive a statistical TM. In addition, he adapts the SMT decoder of (Germann et al., 2001) to avail of both the statistical TM resources and the translation model itself. Unlike the system of (Vogel & Ney, 2000), for which no evaluation is provided, Marcu demonstrates that his hybrid system outperforms two (unnamed) commercial systems: the hybrid French–English system translated 58%

of a 505-sentence test set perfectly, while the commercial systems did so for only 40–42% of the sentences.

In similar work, (Langlais & Simard, 2002) also attempt to merge EBMT and SMT resources. Despite the increase in WER when the SMT system is augmented with TM data, the authors observe "many cases where the translation obtained by merging the extracted examples with the decoder clearly improved the results obtained by the engine alone".

There also exist previous attempts to link TMs with EBMT. (Carl & Hansen, 1999) show that when the fuzzy match score of a TM falls below 80%, translation quality is likely to be higher using EBMT than with TM. (Planas & Furuse, 2003) extend TMs in the direction of EBMT by allowing sub-sentential matches, and providing a multi-level structuring of TMs.

However, to our knowledge the first research which sets out in detail a comparison between the leading data-driven approaches to MT is (Way & Gough, 2005). Here they provide an in-depth comparison of their EBMT system with a word-based SMT system constructed from freely available tools. According to a wide variety of automatic evaluation metrics, they demonstrated that their EBMT system outperformed the SMT system for both French–English and English–French translation.

Given that they did not test their EBMT system against a phrase-based SMT system, the findings of (Way & Gough, 2005), while interesting, are of rather limited value. Accordingly, in (Groves & Way, 2005), we replicated their experiments using the Pharaoh phrase-based SMT Decoder (Koehn, 2004)[1] instead of the word-based ISI ReWrite Decoder.[2] In general, in (Groves & Way, 2005) we showed that the baseline phrase-based SMT system still fell short of the quality obtained via EBMT for these evaluation metrics for English–French and French–English. However, when Pharaoh was seeded with the data sets automatically induced by both Giza++ (Och & Ney, 2003)[3] and the EBMT system of (Way & Gough, 2005), better translation quality results are seen for French–English (0.489 BLEU score) than for the EBMT system *per se* (0.4611).

While (Groves & Way, 2005) show that a hybrid example-based SMT system can outperform both an SMT system and an EBMT system from which it is built, our primary goal in this paper is to see whether a new hybrid model of 'statistical EBMT' can similarly outperform the baseline systems.

Finally, (Aue et al., 2004) observe that their approach of merging dependency treelets with phrase-based SMT may be considered as an instance of "the convergence of statistical and example-based machine translation". By learning non-contiguous word sequences directly and making use of syntactic information in source language dependency trees and during decoding, they are able to more accurately predict the target language position of words. For French–English they obtain a 3.5% relative increase in BLEU score over Pharaoh, while for English-Japenese the treelet approach scores 0.332 BLEU, compared to 0.306 for Pharaoh.

3 Example-Based MT

Assuming a corpus of source–target sentence pairs, EBMT models of translation perform three distinct processes in order to transform a new input string into a target language translation:

1. Searching the source side of the bitext for 'close' matches and their translations;

2. Determining the sub-sentential translation links in those retrieved examples;

3. Recombining relevant parts of the target translation links to derive the translation.

Searching for the best matches involves determining a similarity metric based on word occurrences and part-of-speech labels, generalised templates and bilingual dictionaries. The recombination process depends on the nature of the examples used in the first place, which may include aligning phrase-structure (sub-)trees (Hearne &

[1] http://www.isi.edu/licensed-sw/pharaoh/

[2] http://www.isi.edu/licensed-sw/rewrite-decoder/

[3] http://www.isi.edu/~och/Giza++.html

Way, 2003) or dependency trees (Watanabe, Kurohashi, & Aramaki, 2003), or using placeables (Brown, 1999) as indicators of chunk boundaries.

3.1 Marker-Based EBMT

An alternative approach used in the EBMT system used in our experiments (Gough & Way, 2004; Gough, 2005; Way & Gough, 2005) is to use a set of closed-class words to segment aligned source and target sentences and to derive an additional set of lexical and phrasal resources. This series of research papers is based on the 'Marker Hypothesis' (Green, 1979), a universal psycholinguistic constraint which posits that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes. In a preprocessing stage, the source–target aligned sentences are segmented at each new occurrence of a marker word, and together with cognate matches and mutual information scores, aligned marker chunks, generalised templates and a word-level lexicon are derived.

In order to describe this resource creation in more detail, consider the English–French example in (1) (from (Koehn, 2005), Figure 2):

(1) that is almost a personal record for me this autumn!
⟶c' est pratiquement un record personnel pour moi , cet automne!

The first stage involves automatically tagging each closed-class word in (1) with its marker tag, as in (2):

(2) <DET> that is almost <DET> a personal record <PREP> for <PRON> me <DET> this autumn!
⟶<DET> c' est pratiquement <DET> un record personnel <PREP> pour <PRON> moi , <DET> cet automne!

Taking into account marker tag information (label, and relative sentence position), and lexical similarity (via mutual information), the marker chunks in (3) are automatically generated from the marker-tagged strings in (2):

(3) a. <DET> that is almost : <DET> c' est pratiquement
b. <DET> a personal record : <DET> un record personnel
c. <PREP> for me this autumn : <PREP> pour moi cet automne

A set of generalised templates which (Gough, 2005) demonstrates improve both coverage and translation quality are automatically derived from the marker chunks in (3) by simply replacing the marker word by its relevant tag. From the examples in (3), the generalised templates in (4) are derived:

(4) a. <DET> is almost : <DET> est pratiquement
b. <DET> personal record : <DET> record personnel
c. <PREP> me this autumn : <PREP> cet automne

Generalised templates enable more flexibility in the matching process, as now any marker word can be inserted after the relevant tag if it appears with its translation in the lexicon. For example, assuming it to be absent from the set of marker chunks, the string *it is almost* can now be translated by recourse to the template in (4a) and by inserting a (or all) translation(s) for *it* in the system's lexicon.

Such a lexicon can be constructed in two ways. Firstly, deleted marker words in generalised templates are assumed to be translations of each other. Secondly, in source–target marker chunks or generalised templates, where there is just one content word in both source and target, these are assumed to be translationally equivalent. Taking (3c) as an example, the lexical entry in (5) is automatically created:

(5) <LEX> autumn : <LEX> automne

When a new sentence is submitted for translation, it is segmented into all possible n-grams that might be retrieved from the system's memories. For each n-gram these resources are searched from maximal context (specific source–target sentence-pairs) to minimal context (word-for-word translation).

4 Statistical MT

While EBMT models of translation have since their very inception (Nagao, 1984) incorporated both lexical and phrasal information, it is only quite recently that SMT practitioners have obtained higher translation quality via phrase-based models (e.g. (Koehn, Och, & Marcu, 2003; Och, 2003)) compared to the older word-based systems (Brown et al., 1990, 1993). This inclusion of chunks as well as word alignments has been so successful that PBSMT has become, by some distance, the most dominant approach in MT research today.

4.1 Phrasal Alignment Techniques

A number of methods are available in order to extract phrase correspondences from a bilingual training corpus. The most common method is to first perform word alignment using EM methods, such as that performed by Giza++. Following the method of (Och & Ney, 2003), word alignment is performed in both source–target and target–source directions. The intersection of these unidirectional alignments is taken (producing a set of highly confident word alignments) and is extended iteratively by adding adjacent alignments present within the union of the unidirectional alignment sets. In a final step, alignments are added to the set that occur in the union, where both the source and target words are unaligned. The resulting set of alignments can then be used to extract pairs of source–target phrases which correspond to these alignments, with translation probabilities estimated from relative frequencies.

5 Experiments

For the various translation experiments we performed, we used the training and test sets of the Europarl corpus (Koehn, 2005). From the designated French–English training section of the corpus, we extracted training sets consisting of 78K, 156K and 322K sentence pairs. The training sentence pairs had a maximum sentence length of 40 words and a maximum relative sentence length ratio of 1.5. The sentences contained in the various training sets were randomly selected and were therefore not necessarily supersets of each other.

For testing, we randomly selected 5000 sentences from the Europarl common test set, again limiting sentence length to 40 words. For this test set, the average sentence length was 20.50 words for French and 18.99 words for English. We performed translation for both French–English and English–French, automatically evaluating translation performance over all systems in terms of Word-Error Rate (WER), Sentence-Error Rate (SER), BLEU score (Papineni, Roukos, Ward, & Zhu, 2002), and Precision and Recall (Turian, Shen, & Melamed, 2003). Our experiments together with their results are described in more detail in the following sections.

5.1 EBMT vs. PBSMT

In order to evaluate the performance of PBSMT against our Marker-based EBMT system, we built a baseline PBSMT system using the Pharaoh phrase-based SMT decoder along with the SRI language modeling toolkit.[4] The translation model used in the system was created using the phrasal extraction technique as described in section 4.1. Our EBMT system used the Marker-based techniques as described in section 3.1 to create chunks, generalised templates and lexical resources.

5.1.1 French–English Results

The results for French–English translation are given in Table 1. Note that doubling the amount of training data improves system performance across the board. However, it is clear to see that the PBSMT system considerably outperforms the EBMT system on the Europarl data sets, on average achieving 0.07 BLEU score higher that the EBMT system and achieving a significantly lower WER (68.55 vs. 82.43 for the 322K data set).

Increasing the amount of training data results in a 3% to 5% increase in rela-

[4] http://www.speech.sri.com/projects/srilm/

tive BLEU score for the PBSMT system, whereas we see a higher increase for EBMT, with a 6.2% to 10.3% relative BLEU score improvement.

		BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	.1217	.4556	.5315	85.63	98.94
	PBSMT	.1943	.5289	.5477	70.74	98.42
156K	*EBMT*	.1343	.4645	.5368	83.55	99.02
	PBSMT	.2040	.5369	.5526	69.41	98.30
322K	*EBMT*	.1427	.4734	.5419	82.43	99.06
	PBSMT	.2102	.5409	.5539	68.55	98.72

Table 1: Comparing the EBMT system of (Gough & Way, 2004) with a PBSMT system for French–English.

5.1.2 English–French Results

		BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	.1240	.4422	.4365	79.09	99.1
	PBSMT	.1771	.5046	.4696	70.44	98.54
156K	*EBMT*	.1374	.4548	.4476	77.66	98.96
	PBSMT	.1855	.5120	.4724	69.37	98.20
322K	*EBMT*	.1488	.4587	.4530	77.73	99.22
	PBSMT	.1933	.5180	.4751	68.30	98.12

Table 2: Comparing the EBMT system of (Gough & Way, 2004) with a PBSMT system for English–French.

The results for the same experiment set up for the reverse language direction are given in Table 2. The PBSMT system continues to outperform our EBMT system by some distance across all metrics (e.g. 0.1933 vs. 0.1488 BLEU score, 0.518 vs. 0.4587 Recall, for the 322K training set). As with French–English, WER is lower for the PBSMT system than the EBMT system for English–French (68.30 vs. 77.73 on the 322K data set), but the difference is somewhat less than for French–English.

Doubling the amount of training data improves BLEU score by about 0.8 absolute (i.e. between 4% and 4.7% relative improvement) for the PBSMT system. Precision and Recall rise and WER and SER fall linearly to the amount of training data. For EBMT, as with French–English, we see a greater increase in BLEU score as we increase the amount of training data, with relative BLEU score improving between 5.4% and 10.8%.

However, we consider it noteworthy that, as in our experiments with the *Sun* data, the performance of the EBMT system remains much more consistent for both language directions than the baseline PBSMT system, which performs about 2% BLEU score worse (about 10% relative) for English–French than for French–English translation across training sets (also reflected in the differences in WER between the two systems for both language directions). The previous work of (Groves & Way, 2005; Way & Gough, 2005) suggests that translating from French–English is inherently 'easier' than for English–French as far fewer agreement errors and cases of boundary friction are likely. For instance, translating *le* as a determiner into English can only realise the word *the*, but in the reverse direction *the* has the possible translations *le, la, l'* and *les*, only one of which will ever be correct in a particular context.

5.2 Hybrid System Experiments

Following on from the experiments described in section 5.1, we merged elements of the EBMT Marker-based alignments and the PBSMT phrases (and words) induced from the GIZA++ word alignments in a number of ways in order to improve the performance of our baseline systems:

LEX-EBMT - *Making use of the PBSMT lexicon*: For these experiments we replaced the lexicon of the baseline EBMT system with the higher-quality PBSMT word-alignments, in an attempt to improve coverage and lower WER.

H-EBMT vs H-PBSMT- *Hybrid EBMT System vs. Hybrid PBSMT System*: For these experiments we merged all of the data (words and phrases) induced from the GIZA++ alignments (as described in section 4.1) with the data extracted via the marker hypothesis, in order to see if these 'fully-hybrid' systems could outperform their baseline equivalents.

EBMT-LM and H-EBMT-LM - *Baseline and Hybrid EBMT systems with language model reranking*: For these experiments we re-ranked the n-best output of the various EBMT systems using the PBSMT system's equivalent language model (LM) (i.e. the 78K LM to rerank the output of the EBMT systems trained on the 78K training set).

5.2.1 LEX-EBMT - *Improving the EBMT lexicon*

In contrast to the *Sun Microsystems* corpus which consists of rather homogeneous data,

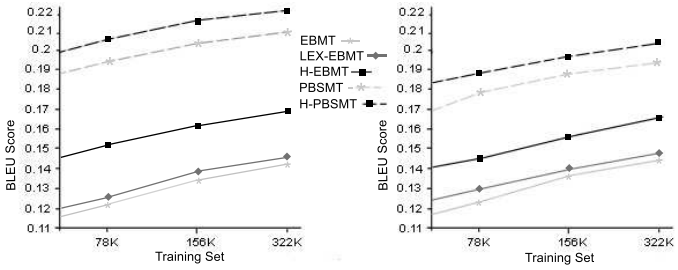
the Europarl corpus consists of much more diverse and complex language. When translating the Europarl data, the EBMT system had to revert to translating individual words much more often (on average 13 words per sentence were considered for direct translation) than when translating using the *Sun* data (on average only 7 words per sentence were candidates for direct translation). The EBMT system seems to perform most poorly when it needs to resort to its lexicon to perform word-for-word translations, reflected in the poor WER scores in the tables (in particular for French–English). In order to address this problem we decided to use the higher-quality SMT word-alignments in place of the EBMT lexicon and repeated the previous experiments for French–English and English–French. The resulting system is referred to as LEX-EBMT in what follows.

French–English Results

The results for French–English are indicated in the graph in Figure 1(a). Here we can see how the use of the improved lexicon leads to only small improvements in translation quality over the equivalent baseline system. We see a relative average increase of 2.9% BLEU score across all training sets. Disappointingly, we also observe a very slight decrease in WER with an average reduction of just 0.5%, indicating that only improving the lexicon is not enough to improve translation quality sufficiently.

English–French Results

From the graph in Figure 1(b) we can see that again BLEU scores increase slightly for the LEX-EBMT system over the baseline, with relative BLEU score increasing by 3.32% on average. Again WER only dropps slightly, with an average decrease of just 0.48%.



(a) French–English (b) English–French

Figure 1: BLEU scores for various baseline and hybrid EBMT and PBSMT systems, for French–English and English–French

5.2.2 H-EBMT vs. H-PBSMT- *Hybrid Systems*

Following on from the LEX-EBMT experiments, we decided to merge all of the data induced via Marker-based methods with the PBSMT phrases and words induced from the GIZA++ word alignments in a similar vein to (Groves & Way, 2005), in order to investigate the possible differences in the quality of these data resources and their effect on translation performance. The resulting systems are referred to as H-EBMT and H-PBSMT in what follows.

French–English Results

Looking at the results for French–English as indicated in Figure 1(a) it is clear that adding the hybrid data improves over all baseline results even further. Most importantly, as (Groves & Way, 2005) showed for the *Sun* data, incorporating the EBMT marker chunks and the PBSMT subsentential alignments in a hybrid 'example-based SMT' system improves on the baseline PBSMT system. Relative BLEU score increases on average by 6.8% and Precision and Recall also rise, with WER falling.

One very interesting result is that the H-PBSMT system achieves a higher BLEU score when trained on the 78K data set compared with the baseline system trained on twice as much data (0.207 vs. 0.204). We get a similar result for the 156K set (0.2176 vs. 0.2120 for the baseline system trained on the 322K set). For H-EBMT, we see a greater increase over the baseline system, with an average relative increase of 21.4% BLEU score). The improvements

in the performance of the H-EBMT system are also reflected in the increase in chunk coverage. A further 6% of test sentences are successfully translated by the H-EBMT system using chunks alone and we get an average relative increase of 76% for the number of possible chunk matches contained in the hybrid EBMT database.

English–French Results

From the graph in Figure 1 (b), again we can see that adding the hybrid data improves over all baseline results, with relative BLEU score increasing by 15% on average over the baseline EBMT system. It is interesting to note that the H-EBMT system trained on only 78K sentence pairs performs almost as well as the baseline system trained on over 4 times as much data. As with the H-EBMT system, the H-PBSMT system improves over its baseline equivalent, achieving a relative increase in BLEU score of 6.2%. We also observe a decrease in WER for the hybrid system compared to its baseline equivalent (an average decrease of 4.05% for H-EBMT and 1.09% for H-PBSMT).

5.2.3 EBMT-LM and H-EBMT-LM

From the results in section 5.2.1, we realised that improving the lexicon was not sufficient to help the EBMT system when it has to resort to performing word-for-word translation. Whithout any guide as to the correct target language word order, the EBMT engine simply follows the order of the words in the original input sentence and thus often fails to produce a syntactically well-formed output translation in these cases. Following these observations we re-ranked the output of both the baseline and hybrid EBMT systems, using the same language model as was used in the equivalent PBSMT experiments. (Bangalore et al., 2002) make use of a trigram language model to help select the best translation from multiple candidates, but do not explicitly report on the actual contribution of the language model and deal with outputs from multiple MT engines rather than from a single system. (Aramaki, Kurohashi, Kashioka, & Kato, 2005) use a language model, but only to re-order the words in the final translation produced by their system, rather than during the re-ranking of translation candidates.

French–English Results

The results from these experiments for French–English are shown in Table 3.

Comparing these results to those of the EBMT and H-EBMT systems, we can see that using the language model improves the performance of both the baseline and hybrid 'statistical EBMT' systems. These results illustrate how the language model guides the reordering of these word-to-word translations to improve overall translation quality. For the baseline EBMT system the BLEU score rises by about 10% relative across training sets, Precision rises, Recall stays about the same, but WER improves by about 4% absolute (5% relative) across the board.

For the H-EBMT system the BLEU score rises by 6–7% relative across training sets, and in a similar pattern to the baseline EBMT system, relative across training sets, Precision rises, Recall stays about the same, and WER improves by about 2% absolute (about 2.7% relative) for the 78K training set.

English–French Results

Similar improvements can also be seen for English–French (Table 3). We get an average relative increase of 6.2% BLEU score across the various training sets. WER also improves, falling from 77.73% to 74.69% for the 322K data set, averaging at a decrease of 2.62% across all training sets. The language model also improves over the H-EBMT system. We get an average relative increase of 4.8% BLEU score and again WER scores improve, falling from and average of 74.11% for the H-EBMT system to 73.55% with the addition of a language model.

		French-English					English–French				
		BLEU	**Prec.**	**Recall**	**WER**	**SER**	**BLEU**	**Prec.**	**Recall**	**WER**	**SER**
78K	*EBMT-LM*	.1335	.4718	.5282	81.46	98.76	.1335	.4596	.4429	76.51	99.00
	H-EBMT-LM	.1624	.5091	.5341	74.15	98.80	.1527	.4871	.4611	73.23	99.08
156K	*EBMT-LM*	.1474	.4832	.5381	79.33	98.56	.1460	.4688	.4529	75.42	98.76
	H-EBMT-LM	.1722	.5177	.5463	75.14	99.16	.1635	.4955	.4709	73.86	98.88
322K	*EBMT-LM*	.1567	.4930	.5428	77.97	98.72	.1557	.4755	.4599	74.69	98.68
	H-EBMT-LM	.1773	.5224	.5530	72.85	98.80	.1744	.5014	.4818	71.96	98.80

Table 3: Re-ranking the output of the baseline EBMT system and 'hybrid' EBMT system for French–English and English–French

6 Conclusions

(Way & Gough, 2005) demonstrate that their Marker-based EBMT system is capable of outperforming a word-based SMT system. (Groves & Way, 2005) show that the EBMT system also outperforms a PBSMT system constructed from freely available resources. However, perhaps more importantly we showed that a hybrid 'example-based SMT' system incorporating marker chunks and SMT sub-sentential alignments is capable of outperforming both baseline translation models on which it is based.

On a different data set—the Europarl corpus—we demonstrate in this paper that the baseline PBSMT system achieves higher translation quality than the EBMT system of (Way & Gough, 2005). For the most part, we feel that this is due to the heterogeneous nature of the training data compared to the *Sun* TM that was used in previous experiments.

We demonstrate that in a number of novel improvements, the baseline EBMT system can be improved by adding in resources derived from a PBSMT system: adding SMT chunks helps, as does the set of word alignments induced by Giza++, and incorporating a target language model in a *post hoc* reranking stage improves translation quality still further. However, the novel hybrid 'statistical EBMT' systems continue to fall short of the translation quality achieved by the PBSMT system. Nevertheless, as shown in (Groves & Way, 2005), we show in a further experiment that adding the EBMT marker chunks to the baseline SMT system derived an 'example-based SMT' system that was capable of improving translation quality compared to the baseline PBSMT system, for a range of automatic evaluation metrics.

The consequences for the field of data-driven MT are clear; by incorporating sub-sentential resources from both SMT and EBMT into novel hybrid systems, translation quality will improve compared to the baseline variants (with much less training data being required). That is, while there is an obvious convergence between both paradigmatic variants, more gains are to be had from combining their relative strengths in novel hybrid systems.

7 Future Work

A number of avenues remain for future research. Given that 110 MT systems can be created using the Europarl resources, we aim to extend our analysis to different language pairs. In addition, we hope to use EBMT techniques, such as integrating marker words into SMT language and translation models, and incorporating generalised templates in SMT, for the extraction of phrasal alignments for PBSMT. We also wish to develop a more intelligent way to incorporate punctuation markers in the EBMT system, perhaps by including <PNCT> as an *end*-of-sequence marker. We also wish to test the systems with regards to out-of-vocabulary words and to repeat the experiments outlined in this paper, but this time using the Europarl common test set containing 1,756 sentences with sentence lengths of 5-15 words.

Furthermore, rather than just using the SMT language model for reranking, we hope to integrate a language model during the EBMT recombination phase.

References

Aramaki, E., Kurohashi, S., Kashioka, H., & Kato, N. (2005). Probabilistic Model for Example-Based Machine Translation. In *Machine Translation Summit X* (pp. 219–226). Phuket, Thailand.

Aue, A., Menezes, A., Moore, R., Quirk, C., & Ringger, E. (2004). Statistical Machine Translation Using Labeled Semantic Dependency Graphs. In *Proceedings of TMI-04.*

Bangalore, S., Murdock, V., & Riccardi, G. (2002). Boostrapping Bilinugal Data using Consensus Translation for a Multilingual Instant Messaging System. In *COLING 2002* (pp. 1–7). Tapei, Taiwan.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., & Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, *16*, 79–85.

Brown, P., Pietra, S. D., Pietra, V. D., & Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, *19*(2), 263–311.

Brown, R. (1999). Adding Linguistic Knowledge to a Lexical Example-based Translation System. In *Proceedings of TMI-99.*

Carl, M., & Hansen, S. (1999). Linking Translation Memories with Example-Based Machine Translation. In *Machine Translation Summit VII* (pp. 617–624). Singapore.

Germann, U., Jahr, M., Knight, K., Marcu, D., & Yamada, K. (2001). Fast Decoding and Optimal Decoding for Machine Translation. In *Proceedings of the 39th ACL and 10th Conference of the European Chapter* (pp. 228–235). Toulouse, France.

Gough, N. (2005). *Example-Based Machine Translation Using the Marker Hypothesis.* Unpublished doctoral dissertation, Dublin City University, Dublin, Ireland.

Gough, N., & Way, A. (2004). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of TMI-04.*

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

Groves, D., & Way, A. (2005). Hybrid Example-Based SMT: the Best of Both Worlds? In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond* (pp. 183–190). Ann Arbor, MI.

Hearne, M., & Way, A. (2003). Seeing the Wood for the Trees: Data-Oriented Translation. In *Machine Translation Summit IX* (pp. 165–172). New Orleans, LA.

Koehn, P. (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In R. Frederking & K. Taylor (Eds.), *Machine Translation: From Real Users to Research; AMTA 2004, LNAI 3265* (pp. 115–124). Berlin/Heidelberg, Germany: Springer Verlag.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X* (pp. 79–86). Phuket, Thailand.

Koehn, P., Och, F., & Marcu, D. (2003). Statistical Phrase-Based Translation. In *Human Language Technology Conference (HLT-NAACL)* (pp. 48–54). Edmonton, Canada.

Langlais, P., & Simard, M. (2002). Merging Example-Based and Statistical Machine Translation. In S. Richardson (Ed.), (pp. 104–113). Berlin/Heidelberg, Germany: Springer Verlag.

Marcu, D. (2001). Towards a Unified Approach to Memory- and Statistical-Based Machine Translation. In *Proceedings of the 39th ACL and 10th Conference of the European Chapter* (pp. 378–385). Toulouse, France.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In A. Elithorn & R. Banerji (Eds.), (pp. 173–180). Amsterdam, The Netherlands: North-Holland.

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of 41^{st} ACL* (pp. 160–167). Sapporo, Japan.

Och, F., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *29*, 19–51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th acl)* (pp. 311–318). Philadelphia, PA.

Planas, E., & Furuse, O. (2003). Formalizing Translation Memory. In M. Carl & A. Way (Eds.), *Recent Advances in Example-Based Machine Translation* (pp. 157–188). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Turian, J., Shen, L., & Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. In *Machine Translation Summit IX* (pp. 386–393). New Orleans, LA.

Vogel, S., & Ney, H. (2000). Construction of a Hierarchical Translation Memory. In *Proceedings of the 18th International Conference on Computational Linguistics: COLING 2000 in Europe* (pp. 1131–1135). Saarbrücken, Germany.

Watanabe, H., Kurohashi, S., & Aramaki, E. (2003). Finding translation patterns from paired source and target dependency structures. In M. Carl & A. Way (Eds.), *Recent Advances in Example-Based Machine Translation* (pp. 397–420). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Way, A., & Gough, N. (2005). Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, *11*(3), 295–309.