Natural Language Engineering **11** (3): 295–309. © 2005 Cambridge University Press
doi:10.1017/S1351324905003888 Printed in the United Kingdom

Comparing example-based and statistical machine translation

ANDY WAY and NANO GOUGH
School of Computing, Dublin City University, Dublin 9, Ireland
e-mail: {away,ngough}@computing.dcu.ie

(*Received May 1 2004; revised November 30 2004*)

Abstract

In previous work (Gough and Way 2004), we showed that our Example-Based Machine Translation (EBMT) system improved with respect to both coverage and quality when seeded with increasing amounts of training data, so that it significantly outperformed the on-line MT system *Logomedia* according to a wide variety of automatic evaluation metrics. While it is perhaps unsurprising that system performance is correlated with the amount of training data, we address in this paper the question of whether a large-scale, robust EBMT system such as ours can outperform a Statistical Machine Translation (SMT) system. We obtained a large English-French translation memory from *Sun Microsystems* from which we randomly extracted a near 4K test set. The remaining data was split into three training sets, of roughly 50K, 100K and 200K sentence-pairs in order to measure the effect of increasing the size of the training data on the performance of the two systems. Our main observation is that contrary to perceived wisdom in the field, there appears to be little substance to the claim that SMT systems are *guaranteed* to outperform EBMT systems when confronted with 'enough' training data. Our tests on a 4.8 million word bitext indicate that while SMT appears to outperform our system for French-English on a number of metrics, for English-French, on all but one automatic evaluation metric, the performance of our EBMT system is superior to the baseline SMT model.

1 Introduction

Gough and Way (2004) showed that our Example-Based Machine Translation (EBMT) system improved with respect to both coverage and quality when seeded with increasing amounts of training data, so that it significantly outperformed the on-line MT system *Logomedia*[1] according to a wide variety of automatic evaluation metrics. While it is perhaps unsurprising that system performance is correlated with the amount of training data, we address in this paper the far more interesting question of whether our large-scale, robust EBMT system can outperform a Statistical Machine Translation (SMT) system.

[1] http://www.logomedia.net. In Gough and Way (2003), we observed that when phrases derived from *Logomedia* were used to seed our example-base, the resulting translations were deemed better in comparison with those derived via two other such on-line systems.

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

There is no doubt that most MT research being undertaken today is corpus-based, and that of the two main data-driven approaches, by far the more dominant paradigm is SMT. Nonetheless, it is a fact that it is harder than it has ever been to describe the differences between the two methods. This used to be easy: from its very inception (Nagao 1984) to the present-day approaches (Carl and Way 2003), EBMT has sought to translate new texts by means of a range of sub-sentential data – both phrasal and lexical – stored in the system's memory. Until quite recently, SMT models of translation were based on the simple word alignment models of Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Mercer and Roossin (1990).

Nowadays, SMT practitioners also get their systems to learn phrasal as well as lexical alignments (Koehn, Och and Marcu 2003; Och 2003). Unsurprisingly, the quality obtained by today's phrase-based SMT systems is considerably better than that obtained by the poorer word-based models. Despite the fact that EBMT models have been modelling both lexical and phrasal correspondences for twenty years, we are unaware of any papers on SMT which acknowledge this debt to EBMT, nor which describe their approach as 'example-based'. Nonetheless, commentators can be found doing just this, for example, in the debate on whether the MT community had reached the 'Holy Grail' at the MT Summit in 2003 (Hovy & Ravichandran 2003)[2] describe the system of Och (2003) as "[adding] EBMT phrase patterns".[3]

We are unsure as to why SMT practitioners omit to attribute certain ideas to EBMT; perhaps, as Simard observes: "I suspect that the EBMT label has a bit of a dubious reputation among the machine-learning crowd."[4] In our experience, there is a good deal of ignorance in the community regarding the historical provenance of 'phrase-based' MT. For example, in submitting a paper to a recent (non-MT) conference, we received the following comment from a reviewer: "the paper … completely ignores the current mainstream empirical approach to machine translation: phrase-based or template-based statistical machine translation". We take this comment to be an indication that the recent developments in SMT are (wrongly) seen as seminal, rather than their debt to EBMT being acknowledged.

In such a climate where one paradigm is clearly dominant, a number of errors, myths and misconceptions such as these abound. In the same debate on EBMT and SMT we started on the MT-List, one eminent scholar in our field wrongly observed that in EBMT systems, all sub-sentential alignments were obtained by hand, whereas SMT systems acquired their phrasal and lexical knowledge automatically. This is, sadly, not atypical of the lack of understanding in the community as regards EBMT, despite the recent appearance of Carl and Way (2003).

In addition, all three reviewers of that (rejected) paper stated that we should compare our approach with an SMT approach. While this is a good idea, and indeed forms the basis of this paper, we suspect that no SMT paper has ever been asked to compare its results to EBMT. Ignoring here the fact that no freely available

[2] Available at http://www.amtaweb.org/summit/MTSummit/FinalPapers/panel-hovy.pdf.
[3] Note that we disagree that these are "EBMT phrase patterns" *per se*, cf. comment after (2).
[4] MT-List digest, Vol 1 #38-5, 12/07/04.

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

EBMT system exists, we consider these comments to be typical of the current bias in the MT research paradigm in favour of SMT.

In comparing our large-scale EBMT system (Gough and Way 2004) to SMT, we wanted to shed some light on another matter. It is generally agreed that the amount of training text required to develop an SMT system is much larger than that needed to seed the memories of an EBMT system. This has, we feel, probably been shown to be correct: if one consults Somers (2003), EBMT systems ranging from just a few to over 700K training pairs can be seen. By contrast, SMT systems have been created based on their exposure to hundreds of millions of training examples (e.g. the Chinese-English SMT system of Vogel, Zhang, Huang, Tribble, Venugopal, Zhao and Waibel (2003) is trained on 150 million words). Probably as a result of this, the perceived wisdom in our community is that all SMT systems need is 'enough' training text for them to outperform an EBMT system. As far as we know, this has never been shown to be the case definitively: it may well be true, but it cannot be allowed to persist unless rigorous scientific experimentation has shown that it is indeed the case.

This, therefore, is the departure for this paper. We set out to test whether our marker-based EBMT system could outperform an SMT system compiled from generally available tools. We obtained a large translation memory from *Sun Microsystems* containing 207,468 English-French sentence pairs in total. From this, we randomly extracted 3,939 sentences which constitutes our test set. In order to test out this hypothesis, we split the remaining 203,529 sentences into three training sets, of roughly 50K, 100K and 200K sentence-pairs. In this way we could observe the effect of increasing the size of the training data on the performance of the two systems. In our previous work, we ensured that each word in the test set was contained somewhere within the word-level lexicon in our system, but in order to maintain a fair and unbiased comparison between EBMT and SMT, we did not impose such a constraint in this experiment.

We used the following tools to develop a language and translation model for the SMT system used in our comparative experiments:

- *Giza++* (Och and Ney 2003);[5]
- the CMU-Cambridge statistical toolkit;[6]
- the ISI ReWrite Decoder.[7]

Translation was performed at each stage from both English-French and French-English, and the resulting translations were evaluated using a range of automatic metrics. At each stage, our EBMT system and the SMT system were trained on the three data sets. Note that at each stage the amount of training data is roughly doubled, ranging from 1.1M words, to 4.8M words.

The remainder of the paper is constructed as follows. In section 2, we summarize how typical SMT and EBMT models of translation work, and despite these

[5] http://www.isi.edu/~och/Giza++.html
[6] http://mi.eng.cam.ac.uk/~prc14/toolkit.html
[7] http://www.isi.edu/licensed-sw/rewrite-decoder/

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

approaches becoming ever more similar, we offer up some areas where we think they continue to differ. In section 3, we present our robust, large-scale EBMT system discussed in Gough and Way (2004). In section 4, we describe the experiments carried out to try to shed some light on the otherwise unexplored area of how EBMT compares to SMT. We present the results and discuss the main findings. Finally, we conclude and offer some thoughts for further research.

2 EBMT and SMT

We briefly summarize here how example-based and statistical approaches attempt to solve the problems of MT. A prerequisite for both methods is a bilingual corpus (or 'bitext') of sententially-aligned examples. Although similar in that both models of translation induce translation knowledge from this resource, there are significant differences regarding both the type of information learnt and how this is brought to bear in dealing with new input.

SMT systems deduce language and translation models from very large quantities of monolingual and bilingual data using a range of theoretical approaches to probability distribution and estimation. The translation model establishes the set of target language words (and more recently, phrases) which are most likely to be useful in translating the source string, while the language model tries to assemble these words (and phrases) in the best order possible. The language model is trained by determining all bigram and/or trigram frequency distributions occurring in the training data, while the translation model takes into account source and target word (and phrase) co-occurrence frequencies, sentence lengths and the relative sentence positions of source and target words.

EBMT systems translate new sentences by adapting the aligned sentences in the example-base. Input strings are matched against the source side of the corpus and 'close' matches are retrieved with their translations, translationally relevant chunks are determined from these similar matches, which are then recombined to generate a target string. In searching for the best matches, similarity is established by means of word-cooccurrence and part-of-speech information, generalised templates and bilingual dictionaries. In order for target chunks to be recombined, sub-sentential alignment algorithms are required, using a variety of methods, including:

- deriving transfer rules from examples (Furuse and Iida 1992);
- aligning phrase-structure (sub-)trees (Hearne and Way 2003);
- aligning dependency structures (Watanabe, Kurohashi and Aramaki 2003);
- using placeables (Brown 1999).

Another method (Way and Gough 2003; Gough and Way 2004) uses a set of marker-words to segment aligned source and target sentences and to derive an additional set of lexical and phrasal resources. This is decribed in detail in the next section.

As stated in the previous section, we observed that the mainstream phrase-based models of SMT make it more difficult than it has ever been to distinguish clearly

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

such approaches from models of EBMT. Nonetheless, in our view there remain a couple of differences between EBMT and phrase-based SMT models of translation:

1. To a greater (e.g. Hearne and Way 2003; Melamed, Satta and Wellington 2004) or lesser (e.g. our work on marker-based EBMT, cf. section 3) extent, an EBMT model of translation builds in syntax *at its core* – most SMT systems, by contrast (the preliminary work of Charniak, Knight and Yamada (2003) being a significant exception), only utilise models of syntax in a *post hoc* reranking process, and even here, Koehn *et al.* (2003) demonstrated that 'bolting on' syntax in this manner did not help improve translation quality;
2. given the above, phrase-based SMT systems are likely to 'learn' chunks that EBMT systems would not.

Where we say that EBMT approaches to translation build in some model of syntax, we do not necessarily mean PS-trees (although both Hearne and Way (2003) and Melamed *et al.* (2004) do use trees as their example-base), nor do we exclusively mean dependency trees (Watanabe *et al.* 2003). Even the approaches of Brown (1999) or Way and Gough (2003) and Gough and Way (2004) (cf. section 3) contain some partial notion of surface syntax, so that in the recombination stage, the chances of chunks being stitched together in the correct order are improved.

Regarding the observation that phrase-based SMT learns chunks that EBMT systems would not, consider the example in (1) (from Koehn and Knight, 2004):

(1) Mary did not slap the green witch ⟹Maria no dió una botefada a la bruja verde.
(Lit: 'Mary not gave a slap to the witch green')

From this aligned example, an SMT system would potentially learn the 'phrases' (along with many others) in (2):

(2) a. slap the ⟹dió una botefada a
b. slap the ⟹dió una botefada a la
c. the green witch ⟹a la bruja verde

Note that the SMT system essentially learns *n*-gram sequences, rather than phrases *per se*. Even an EBMT system containing somewhat 'surfacey' syntax such as our Marker-Based approach would not accept the sub-sentential alignments in (2), as it would not be possible (nor do we consider it to be terribly sensible) to break an NP between a Determiner and its head Noun, or a PP between the Preposition and the following NP.

3 Our marker-based EBMT system

In Gough and Way (2004), we presented the latest version of our 'linguistics-lite' EBMT system based on the 'Marker Hypothesis' (Green 1979). This is a universal psycholinguistic constraint, stating that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes.

We define seven sets of marker words for English and French, namely determiners (⟨DET⟩), prepositions (⟨PREP⟩), quantifiers (⟨QUANT⟩), conjunctions (⟨CONJ⟩), wh-adverbs (⟨WRB⟩), possessive (⟨POSS⟩) and personal pronouns (⟨PRON⟩). These

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

are then used to segment the source-target aligned sentences. As an example, consider the strings in (3) appearing in the *Sun Microsystems* TM:

(3) 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 (4):

(4) ⟨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 (4) 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 (4), the marker chunks in (5) are generated:

(5) 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

Marker chunks such as those in (5) are automatically derived using an improved sub-sentential alignment algorithm which (a) enables much more data to be retained in the system's memories, and (b) leads to a significant increase in translation quality. This method presented in Gough and Way (2004) checks that chunks are marked with similar tags, but in addition uses a base-dictionary created via *Logomedia* to check for equivalences between content words across chunks. Along with cognate matches and mutual information scores, 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 greater the distance between two chunks, the less likely they are to align. In our previous work (e.g. Way and Gough 2003), it was only possible to produce 1:1 alignments, whereas our new, smarter algorithm facilitates the merging of chunks to produce 2:1 and 3:1 alignments.

Further lexical information can be extracted from the marker chunks in (5). We take advantage of the fact that chunks containing just one non-marker word in both source and target are often translations of each other. In this way we can extract the 'word-level' translations in (6):

(6) ⟨PREP⟩ to : pour ⟨LEX⟩ view : visualiser ⟨LEX⟩ effect : effet
⟨PRON⟩ you : vous ⟨DET⟩ the : l' ⟨PREP⟩ of : de

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

More general examples can add flexibility to the matching process and improve coverage. In a final pre-processing stage, therefore, we produce a set of generalised marker templates by replacing marker words with their associated tags. For example, from the entries in (5), the templates in (7) can be generated:

(7) 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 (7) 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 can now be found, its translation ⟨DET⟩ *effet* can be retrieved. The final translation can be produced by inserting the translation of *an* in place of the ⟨DET⟩ tag.

In sum, a total of four lexical resources are automatically created, namely:

- the original sententially-aligned source-target pairs (cf. (3) above);
- the marker-aligned chunks (5);
- the generalised marker chunks (7);
- the word-level lexicon (6).

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 (with the exception of those ending with a marker word–given our marker-based segmentation method, new chunks are created when marker words are found), these resources are searched in the order provided in the bulleted list above, from maximal (specific source-target sentence-pairs) to minimal context (word-for-word translation). Each translation retrieved is assigned a weight using the formula in (8):

(8) $$\text{weight} = \frac{\textit{no. occurrences of the proposed translation}}{\textit{total no. translations produced for SL phrase}}$$

For example, given the source language phrase *the house*, assuming that *la maison* is found 8 times and *le domicile* is found twice, then *weight*("*la maison*" "*the house*") = 8/10 and *weight*("*le domicile*" "*the house*") = 2/10.

4 Experiments and results

In Gough and Way (2004), we showed that our enhanced system was capable of outperforming the on-line MT system *Logomedia* in both directions (French-English and English-French). Despite the fact that our system was trained on data similar to the test data, and that by contrast *Logomedia* is a general-purpose system, the scale by which we outperformed *Logomedia* quite clearly showed that when trained on over 200K sentence pairs, our system derived significantly better translations as adjudged by a range of automatic evaluation metrics.

In this section, we describe an experiment using the same training data where we attempt to shed some light on the widely held (and barely contested – yet in our

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

Table 1. *Size of the training sets in words*

Training set	# Eng. Words	# Fr. Words	Total
TS1	537,309	631,770	1,169,079
TS2	1,115,218	1,305,753	2,420,971
TS3	2,215,834	2,585,281	4,801,115

view unproven) opinion that when it comes to a large amount of training text, SMT will clearly win out over EBMT.

The experimental set-up was as follows: from the original *Sun Microsystems* TM of 207,468 English-French sentence pairs, we randomly extracted a 3,939-sentence test set. The remaining 203,529 sentences were used as training data, split three ways:

- Training Set 1 (TS1): 50,882 English-French sentence pairs;
- Training Set 2 (TS2): 101,765 English-French sentence pairs (inc. TS1);
- Training Set 3 (TS3): the full set of 203,529 English-French sentence pairs (inc. TS1 and TS2).

By doubling the size of the training sets, even if the results obtained were unclear, we wanted to see whether the respective plots for SMT and EBMT converged at all. In terms of words, the size of the training sets is given in Table 1.

In our previous work, we ensured that each word in the test set was contained somewhere within the word-level lexicon in our system, but in order to maintain a fair and unbiased comparison between EBMT and SMT, we did not impose such a constraint in this experiment.

We constructed the SMT system used in our experiments via the freely available tools *Giza++*, the CMU-Cambridge statistical toolkit and the ISI Rewrite Decoder. The results shown in the next two sections are for a range of automatic evaluation metrics, including BLEU (Papineni, Roukos, Ward and Zhu 2002), Precision and Recall (Turian, Shen and Melamed 2003), and Word- and Sentence-Error Rates.[8]

4.1 Comparing SMT and EBMT: English–French

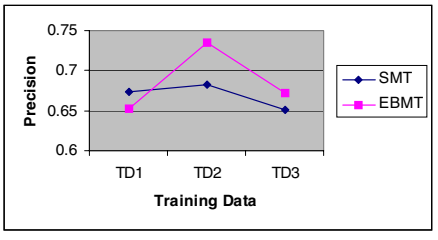
The results obtained for English-French for both our EBMT system and SMT are presented in Table 2 and Figure 1. For the most part, all the automatic evaluation metrics bar one suggest that EBMT can outperform SMT from English-French. The only exception to this is for the first training set, where SMT outperforms EBMT in terms of precision (0.6739 compared to 0.6525). In terms of BLEU score, the EBMT system incrementally improves as the system is augmented with additional training

[8] BLEU scores are calculated using the NIST MT Evaluation Toolkit–http://www.nist.gov/speech/tests/mt/mt2001/index.htm–while Precision and Recall scores are derived via the NYU Generalized Text Matcher–http://nlp.cs.nyu.edu/GTM/. All measures have a range of 0-1, the higher the better for BLEU, Precision and Recall, and the lower the better for WER and SER.

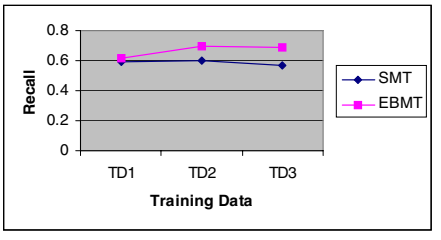
https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

Table 2. *Comparing our EBMT system (English-French) with an SMT system trained on the same data using automatic evaluation metrics*

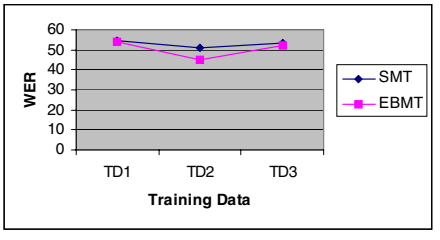
Training set	System	BLEU	Precision	Recall	WER	SER
TS1	*SMT*	0.2971	0.6739	0.5912	54.9	90.8
	EBMT	0.3318	0.6525	0.6183	54.3	89.2
TS2	*SMT*	0.3375	0.6824	0.5962	51.1	89.9
	EBMT	0.4534	0.7355	0.6983	44.8	77.5
TS3	*SMT*	0.3223	0.6513	0.5704	53.5	89.1
	EBMT	0.4409	0.6727	0.6877	52.4	65.6



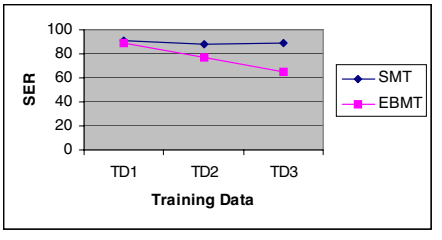
(a) Precision



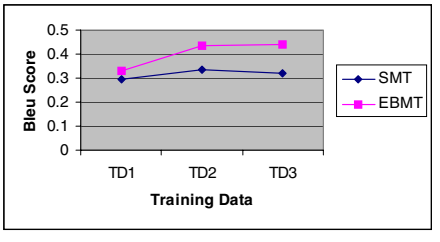
(b) Recall



(c) Word Error Rate



(d) Sentence Error Rate



(e) BLEU scores

Fig. 1. Comparing EBMT and SMT *(English-French)*.

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

data. However, apart from SER, the remaining metrics suggest that training on just over 100,000 sentences pairs yields better results than training on just over 200,000 sentences.

The results for SMT display a similar trend. An improvement is observed from TS1 to TS2. From TS2 to TS3, an improvement is observed with regard to SER. However, all other metrics suggest that translation quality falls when the system is trained on 203,529 sentence pairs, so again, overfitting may be an issue. Training on just over 100,000 sentences, therefore, seems to provide optimal results for both SMT and EBMT for this particular test set, in terms of BLEU score, Precision, Recall and WER. These results are surprising, especially so for SMT, as it is generally assumed that increasing the size of the training data in a system of this type will improve the quality of the output translations. While a degree of overfitting may be observed, we intend to carry out some variance analysis on these results (e.g. performing bootstrap-resampling on the test set (Koehn 2004)), or re-test with different sample test sets in order to investigate whether the same phenomenon is observed.

The figures for SER improve for both SMT and EBMT from TS1 to TS3. However, the improvement is much more significant for EBMT (20.6%) than for SMT (0.1%). This is more than likely due to the fact that when the amount of training data is increased, the number of test sentences for which an exact match can be obtained also increases. When the system is trained on 203,529 sentence pairs, the number of test sentences which can be retrieved directly from our example-base is approximately 10%. This means that for the 34.4% of translations which obtain a 0% SER score, 24.4% of these are produced using chunks derived from the Marker Hypothesis. The graphs in Figure 1 illustrate the results for both EBMT and SMT for each stage in terms of BLEU score, Precision, Recall, WER and SER. Unsurprisingly, the divergence between the lines representing SER in Figure 1(d) becomes more obvious as the amount of training data increases. The other metrics also show that EBMT continues to outperform SMT on the larger training set. It remains to be seen if adding even more training data will affect this trend.

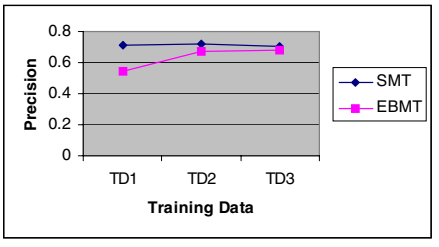
4.2 Comparing SMT with EBMT: French–English

The results for French-English translations are presented in Table 3 and Figure 2. Intuitively, translating from French-English should yield better results than translating from English–French as there are far fewer potential problems concerning agreement errors and boundary friction. That is, where chunks from different translation examples are recombined, the quality of the resulting translations may be compromised (cf. Way and Gough (2003), for instance). The results for SMT provide evidence in favour of this as all the automatic evaluation metrics show better results for French-English. The EBMT system also produces better translations from French-English in terms of BLEU, Recall and SER. However, with regard to WER, the French-English translations obtain an optimal score of 50.8% compared to the better score of 44.8% for English-French. Similarly, the EBMT system obtains a maximum score of 0.7355 for Precision from English-French compared to a top score of 0.6782 for French-English.

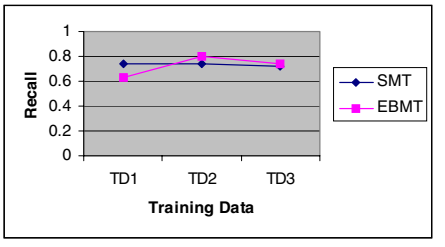
https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

Table 3. *Comparing our EBMT (French-English) system with an SMT system trained on the same data using automatic evaluation metrics*

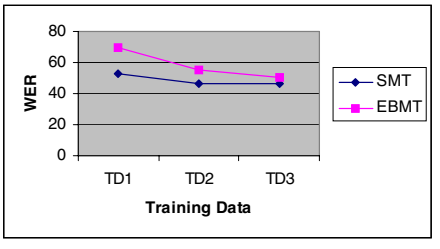
Training set	System	BLEU	Precision	Recall	WER	SER
TS1	*SMT*	0.3794	0.7096	0.7355	52.5	86.5
	EBMT	0.2571	0.5419	0.6314	69.7	89.2
TS2	*SMT*	0.3924	0.7206	0.7433	46.2	81.3
	EBMT	0.4262	0.6731	0.7962	55.2	66.2
TS3	*SMT*	0.4462	0.7035	0.7240	46.8	80.8
	EBMT	0.4611	0.6782	0.7441	50.8	51.2



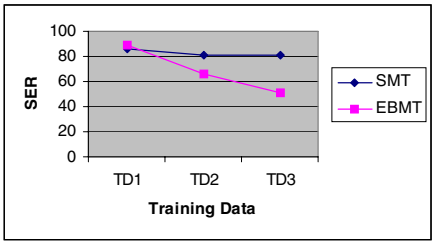
(a) Precision



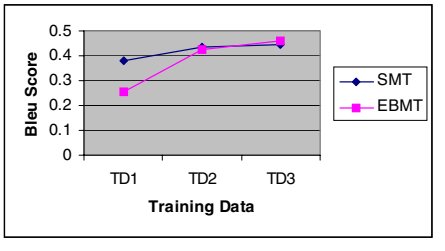
(b) Recall



(c) Word Error Rate



(d) Sentence Error Rate



(e) BLEU scores

Fig. 2. Comparing EBMT and SMT *(French-English)*.

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

While the majority of metrics obtained for English-French indicate that EBMT outperforms SMT, the results for French-English are by no means as conclusive. EBMT does not outperform SMT on the training set TS1 from French-English for *any* of the five metrics. When trained on 101,764 sentences, however, EBMT outperforms SMT in terms of BLEU score, Recall and SER (66.5% compared to 81.3% for SMT), while SMT produces better translations according to Precision and WER (46.2% compared to 55.2%). On training set TS3, SMT continues to outperform EBMT in terms of Precision (2.5%) and WER (4%–both less significant differences than for TS1 and TS2), but EBMT wins out according to the other three metrics–notably, by a huge 29.6% for SER.

The graphs in Figure 2 illustrate the performance of SMT and EBMT when both systems are trained on the same data sets and translation is performed from French-English. Again, the improvement for SER is more marked for EBMT than for SMT. While SMT appears to outperform EBMT when translating from French-English, there is less disparity between the two systems as the amount of training data is increased.

In terms of BLEU score, SMT obtains significantly higher scores for French-English compared to English-French: 8% higher for TS1, 6% higher for TS2, and 12% higher for TS3. Apart from TS1, the EBMT scores for the two different language directions are much more in line, indicating perhaps that EBMT may be more consistent even for the same language pair in different directions.

4.3 Summary

As expected, our experimental results show that both the EBMT system and the SMT model of translation achieve better translation quality from French-English compared to English–French. Of the five automatic evaluation metrics for each of the three training sets, in nine of the fifteen cases SMT wins out over our EBMT system.

However, when compared against SMT, the results for our English-French system are much more significant as fourteen of the fifteen scores indicate that EBMT can outperform SMT. Summing these results together, EBMT outperforms SMT in 20 tests, while SMT does better in 10 experiments.

In all our previous work, we ensured that all words in our test set were contained somewhere within our word-level lexicon. To ensure a fair test against SMT, we do not enforce this constraint, so that certain words cannot be translated by both systems. We deal with these unknown words in our EBMT system by inserting the original source word in the output translation. Given that there are more words in the French test set than in the English equivalent, it is likely that more French words are absent from the word-level lexicon than is the case for English words. As a result, when translating from French–English, an increased number of French words will not be matched in the word-level lexicon. This in turn may adversely affect the quality of the French-English translations.

Nonetheless, while it is interesting to compare the performance of each system for the same language pair in different directions, the main point of the exercise

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

was to compare our EBMT system against SMT, for iteratively large amounts of training data. While the results are a little mixed, it is clear that EBMT tends to outperform SMT on this sublanguage and on these training sets: to return to the original motivation for this paper, far from being the case that EBMT developers ought to compare their systems to SMT, our results indicate that if anything, it ought to be the other way round.

5 Concluding remarks

Our previous work showed that the coverage and quality of our EBMT system improved significantly when more training data was available. We consistently and significantly outperformed the on-line MT system *Logomedia* in a number of tests.

In this paper, we set out to show empirically how our EBMT system fared when compared with the baseline SMT model of translation. Specifically, we wanted to see whether there was any truth in the widely held belief – though not one that we have ever seen tested in the literature – that when 'enough' training data is available, SMT would clearly win out over EBMT.

Across three training sets of 50K, 100K and 200K sentence pairs (1.1M words to 4.8M words), we tested the performance of our EBMT system against an SMT system constructed from freely available resources. For English-French, against a near 4K test set, EBMT outperformed SMT in 14/15 tests. For French–English, by contrast, SMT did better in 9/15 tests. Summing these together to give an indication of overall performance, EBMT appears to outperform SMT by a factor of two to one.

Critics of our approach will simply say that if we were to use still more training data, then SMT would indeed outperform our EBMT system. Significantly, however, both systems do better when trained on 100K sentences rather than the full 203K-sentence training set.

Ultimately, given the fact that our system has some surface syntax designed into the system, while SMT does not, we fail to be convinced that exposure to further amounts of training data will see a drop-off in performance of our system compared to SMT. By extension, therefore, approaches such as Hearne and Way (2003) and Melamed *et al.* (2004) which are based on richer models of syntax ought to outperform SMT even more significantly. With regard to our EBMT system, this shows up best, we feel, when the SER rates are examined: for English-French, we outperform SMT by 23.5%, and by 29.6% for French-English. That is, while the WER rates are comparable (those for SMT are even a little better than for EBMT), indicating that SMT is able to select the right target words compared to the reference translation, without some (even rudimentary, as in our approach) knowledge of syntax (over and above *n*-gram sequences), there is little prospect of getting these in the right order. Accordingly, therefore, we consider the results for our 'linguistics-lite' EBMT model of translation to be promising and are confident that with further improvements to the current EBMT model we can continue to outperform SMT even on larger training sets.

Finally, we wish to continue to experiment with other training data in other sublanguage domains, especially those where larger corpora are available (such as

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

the Canadian Hansards, for instance). In addition, when the *Pharaoh* phrase-based model of SMT becomes available, we plan to re-run these experiments to see what improvement might be seen in SMT over the word-based model used here. Rigorous scientific experimentation is the only way to settle debates such as the one we have outlined here – we do not consider it useful for unproven opinion to circulate in our community as to the relative merits of paradigmatic approaches to translation, and hope that our initial work in this area has shed some light on what to our knowledge was not tested prior to our research.

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. We also thank our colleague, Declan Groves, for helping to assemble the SMT system.

References

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

Brown, R. (1999) Adding Linguistic Knowledge to a Lexical Example-Based Translation System. *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pp. 22–32. Chester, UK.

Carl, M. and Way, A. (2003) *Recent Advances in Example-Based Machine Translation*. Kluwer.

Charniak, E., Knight, K. and Yamada, K. (2003) Syntax-based Language Models for Statistical Machine Translation. *MT Summit 2003*, pp. 40–46. New Orleans, LA.

Furuse, O. and Iida, H. (1992) An Example-Based method for transfer-driven machine translation. *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation: Empiricist vs. Rationalist Methods in MT (TMI-92)*, pp. 139–150. Montréal, Canada.

Gough, N. and Way, A. (2004) Robust Large-Scale EBMT with Marker-Based Segmentation. *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pp. 95–104. Baltimore, MD.

Green, T. (1979) The Necessity of Syntax Markers. Two Experiments with Artificial Languages. *Journal of Verbal Learning and Behavior* **18**: 481–496.

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

Koehn, P. (2004) Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pp. 388–395. Barcelona, Spain.

Koehn, P. and Knight, K. (2004) Introduction to Statistical Machine Translation: Tutorial Notes. *Sixth Biennial Conference of the Association for Machine Translation in the Americas*, Georgetown, Washington, DC.

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

Melamed, D., Satta, G. and Wellington, B. (2004) Generalized Multitext Grammars. *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 662–669. Barcelona, Spain.

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

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press

Och, F. J. (2003) Minimum Error Rate Training in Statistical Machine Translation. *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pp. 160–167. Sapporo, Japan.

Och, F. J. and Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* **29**: 19–51.

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pp. 311–318. Philadelphia, PA.

Somers, H. (2003) An Overview of EBMT In: M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pp. 3–57. Kluwer Academic.

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

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B. and Waibel, A. (2003) The CMU Statistical Machine Translation System. *MT Summit IX*, pp. 402–409. New Orleans, LA.

Watanabe, H., Kurohashi, S. and Aramaki, E. (2003) Finding Translation Patterns from Paired Source and Target Dependency Structures. In: M. Carl and A. Way, editors, *Recent Advances in Example-Based Machine Translation*, pp. 397–420. Kluwer Academic.

Way, A. and Gough, N. (2003) *wEBMT*: Developing and Validating an Example-Based Machine Translation System using the World Wide Web. *Computational Linguistics* **29**(3): 421–457.

https://doi.org/10.1017/S1351324905003888 Published online by Cambridge University Press