Hybrid Example-Based SMT: the Best of Both Worlds?

Declan Groves
School of Computing
Dublin City University
Dublin 9, Ireland
dgroves@computing.dcu.ie

Andy Way
School of Computing
Dublin City University
Dublin 9, Ireland
away@computing.dcu.ie

Abstract

(Way and Gough, 2005) provide an in-depth comparison of their Example-Based Machine Translation (EBMT) system with a Statistical Machine Translation (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 by a factor of two to one.

Nevertheless, they did not test their EBMT system against a phrase-based SMT system. Obtaining their training and test data for English–French, we carry out a number of experiments using the Pharaoh SMT Decoder. While better results are seen when Pharaoh is seeded with Giza++ word- and phrase-based data compared to EBMT sub-sentential alignments, in general better results are obtained when combinations of this 'hybrid' data is used to construct the translation and probability models. While for the most part the EBMT system of (Gough & Way, 2004b) outperforms any flavour of the phrase-based SMT systems constructed in our experiments, combining the data sets automatically induced by both Giza++ and their EBMT system leads to a hybrid system which improves on the EBMT system *per se* for French–English.

1 Introduction

(Way and Gough, 2005) provide what are to our knowledge the first published results comparing Example-Based and Statistical models of Machine Translation (MT). Given that most MT research carried out today is corpus-based, it is somewhat surprising that until quite recently no qualitative research existed on the relative performance of the two approaches. This may be due to a number of factors: the relative unavailability of EBMT systems, the lack of participation of EBMT researchers in competitive evaluations or the dominance in the MT research community of the SMT approach—whenever one paradigm finds favour with the clear majority of MT practitioners, the assumption made by most of the community is that this way of doing things is clearly better than the alternatives.

Like (Way and Gough, 2005), we find this regrettable: the only basis on which such views should be allowed to permeate our field is following extensive testing and evaluation. Nonetheless, given that no EBMT systems are freely available, very few research groups are in the position of being able to carry out such work.

This paper extends the work of (Way and Gough, 2005) by testing EBMT against phrase-based models of SMT, rather than the word-based models used in this previous work. In so doing, it provides a more complete evaluation of the main question at hand, namely whether an SMT system outperforms an EBMT system on reasonably large training and test sets.

We obtained the same training and test data used

Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 183–190,
Ann Arbor, June 2005.

in (Way and Gough, 2005), and evaluated a number of SMT systems which use the Pharaoh decoder[1] against the Marker-Based EBMT system of (Gough & Way, 2004b), for French–English and English–French. We provide results using a range of automatic evaluation metrics: BLEU (Papineni et al., 2002), Precision and Recall (Turian et al., 2003), and Word- and Sentence Error Rates. (Way and Gough, 2005) observe that EBMT tends to outperform a word-based SMT model, and our experiments show that a number of different phrase-based SMT systems still tend to fall short of the quality obtained via EBMT for these evaluation metrics. However, when Pharaoh is seeded with the data sets automatically induced by both Giza++ and their EBMT system, better results are seen for French–English than for the EBMT system *per se*.

The remainder of the paper is constructed as follows. In section 2, we summarize the main ideas behind typical models of SMT and EBMT, as well as the EBMT system of (Gough & Way, 2004b) used in our experiments. In section 3, we revisit the experiments and results carried out by (Way and Gough, 2005). In section 4, we describe our extensions to their work, and compare their findings to ours, and in section 5, present a number of hybrid SMT models. Finally, we conclude and offer some thoughts for future work in section 6, and in section 7 present some further comments on the narrowing gap between EBMT and phrase-based SMT.

2 Example-Based and Statistical Models of Translation

A *sine qua non* for both EBMT and SMT is a set of sentences in one language aligned with their translations in another. Although similar in that both models of translation automatically 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.

2.1 EBMT

Given a new input string, EBMT models use three separate processes in order to derive translations:

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

Another method—and the one used in the EBMT system used in our experiments—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. (Gough & Way, 2004b) base their work 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, (Gough & Way, 2004b) use 7 sets of marker words for English and French (e.g. determiners, quantifiers, conjunctions etc.), which together with cognate matches and mutual information scores are used to derive three new data sources: sets of marker chunks, generalised templates and a lexicon.

In order to describe this in more detail, we revisit an example from (Gough & Way, 2004a), namely:

(1) each layer has a layer number $\Longrightarrow$chaque couche a un nombre de la couche

From the sentence pair in (1), the strings in (2) are generated, where marker words are automatically tagged with their marker categories:

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

(2) <QUANT> each layer has <DET> a layer number $\Longrightarrow$<QUANT> chaque couche a <DET> un nombre <PREP> de la couche

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

(3) a. <QUANT> each layer has: <QUANT> chaque couche a
b. <DET> a layer number: <DET> un nombre de la couche

(3b) shows that *n:m* alignments are possible (the two French marker chunks *un nombre* and *de la couche* are absorbed into one following the lexical similarities between *layer* and *couche* and *number* and *nombre*, respectively) given the sub-sentential alignment algorithm of (Gough & Way, 2004b).

By generalising over the marker lexicon, a set of marker templates is produced by replacing the marker word by its relevant tag. From the examples in (3), the generalised templates in (4) are derived:

(4) a. <QUANT> layer has: <QUANT> couche a
b. <DET> layer number: <DET> nombre de la couche

These templates increase the robustness of the system and make the matching process more flexible. Now any marker word can be inserted after the relevant tag if it appears with its translation in the lexicon, so that (say) *the layer number* can now be handled by the generalised template in (4b) and inserting a (or all) translation(s) for *the* in the system's lexicon.

2.2 Word- and Phrase-Based SMT

SMT systems require two large probability tables in order to generate translations of new input:

1. a translation model induced from a large amount of bilingual data;

2. a target language model induced from a(n even) large(r) quantity of separate monolingual text.

Essentially, 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 most likely target word order. 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.

Until quite recently, SMT models of translation were based on the simple word alignment models of (Brown et al., 1990). Nowadays, however, SMT practitioners also get their systems to learn phrasal as well as lexical alignments (e.g. (Koehn et al., 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.

3 Comparing EBMT and Word-Based SMT

(Way and Gough, 2005) obtained a large translation memory from *Sun Microsystems* containing 207,468 English–French sentence pairs, of which 3,939 sentence pairs were randomly extracted as a test set, with the remaining 203,529 sentences used as training data. The average sentence length for the English test set was 13.1 words and 15.2 words for the corresponding French test set. The EBMT system used was their Marker-based system as described in section 2.1 above. In order to create the necessary SMT language and translation models, they used:

- *Giza++* (Och & Ney, 2003);[2]
- the CMU-Cambridge statistical toolkit;[3]
- the ISI ReWrite Decoder.[4]

Translation was performed from English–French and French–English, and the resulting translations were evaluated using a range of automatic metrics: BLEU (Papineni et al., 2002), Precision and Recall

[2] http://www.isi.edu/∼och/Giza++.html
[3] http://mi.eng.cam.ac.uk/∼prc14/toolkit.html
[4] http://www.isi.edu/licensed-sw/rewrite-decoder/

(Turian et al., 2003), and Word- and Sentence Error Rates. In order to see whether the amount of training data affected the (relative) performance of the EBMT and SMT systems, (Way and Gough, 2005) split the training data into three sets, of 50K (1.1M words), 100K (2.4M words) and 203K (4.8M words) sentence pairs (TS1–TS3 in what follows).

3.1 English–French Results

Table 1: Comparing the EBMT system of (Gough & Way, 2004b) with a Word-Based SMT (WB-SMT) system for English–French.

		BLEU	Prec.	Recall	WER	SER
TS1	*WB-SMT*	.2971	.6739	.5912	54.9	90.8
	EBMT	.3318	.6525	.6183	54.3	89.2
TS2	*WB-SMT*	.3375	.6824	.5962	51.1	89.9
	EBMT	.4534	.7355	.6983	44.8	77.5
TS3	*WB-SMT*	.3223	.6513	.5704	53.5	89.1
	EBMT	.4409	.6727	.6877	52.4	65.6

The results obtained by (Gough & Way, 2004b) for English–French for their EBMT system and word-based SMT (WB-SMT) are given in Table 1. Essentially, all the automatic evaluation metrics bar one (Precision) suggest that EBMT can outperform SMT from English–French. Surprisingly, however, apart from SER, all evaluation scores are higher using 100K sentence pairs as training data rather than the full 203K sentences. It is generally assumed that increasing the size of the training data for corpus-based MT systems will improve the quality of the output translations. (Way and Gough, 2005) observe that while this dip in performance may be due to a degree of over–fitting, they 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.

With respect to SER, however, for both SMT and EBMT, the figures improve as more training data is made available. However, the improvement is much more significant for EBMT (20.6%) than for SMT (0.1%). While the WER scores are much the same, indicating that both systems are identifying reasonable target vocabulary that should appear in the output translation, the vast differences in SER using TS3 indicate that a system containing essentially no information about target syntax has very little hope of arranging these target words in the right order. On the contrary, even a system containing some basic knowledge of how phrases fit together such as the Marker-based EBMT system of (Gough & Way, 2004b) will generate translations of far higher quality.

3.2 French–English Results

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

		BLEU	Prec.	Recall	WER	SER
TS1	*WB-SMT*	.3794	.7096	.7355	52.5	86.5
	EBMT	.2571	.5419	.6314	69.7	89.2
TS2	*WB-SMT*	.3924	.7206	.7433	46.2	81.3
	EBMT	.4262	.6731	.7962	55.2	66.2
TS3	*WB-SMT*	.4462	.7035	.7240	46.8	80.8
	EBMT	.4611	.6782	.7441	50.8	51.2

The results obtained by (Way and Gough, 2005) for French–English translations are presented in Table 2. Translating in this language direction is inherently 'easier' than for English–French as far fewer agreement errors and cases of boundary friction are likely. Accordingly, all WB-SMT results in Table 2 are better than for the reverse direction, while for EBMT, improved results are to be seen for BLEU, Recall and SER.

While the majority of metrics obtained for English–French indicate that EBMT outperforms WB-SMT, the results for French–English are by no means as conclusive. Of the 15 tests, WB-SMT outperforms EBMT in nine.

4 Comparing EBMT and Phrase-Based SMT

From the results in the previous sections for French–English and for English–French, (Way and Gough, 2005) observe that EBMT outperforms WB-SMT in the majority of tests. If we are to treat each of the metrics as being equally significant, it can be said that EBMT appears to outperform WB-SMT by a factor of two to one. In fact, the only metric for which EBMT seems to consistently underperform is precision for French–English which, when we examine WER, indicates that the EBMT system's knowledge of word correspondences is incomplete and not as comprehensive as that of the WB-SMT system.

However, it has been apparent for some time now that phrase-based SMT outperforms previous systems using word-based models. The results obtained by (Way and Gough, 2005) for SER also indicate that if phrase-based SMT were used, then improvements in translation quality ought to be seen.

Accordingly, in this section we describe a set of experiments which extends the work of (Way and Gough, 2005) by evaluating the Marker-based EBMT system of (Gough & Way, 2004b) against a phrase-based SMT system built using the following components:

- Giza++, to extract the word-level correspondences;
- The Giza++ word alignments are then refined and used to extract phrasal alignments ((Och & Ney, 2003); or (Koehn et al., 2003) for a more recent implementation);
- Probabilities of the extracted phrases are calculated from relative frequencies;
- The resulting phrase translation table is passed to the Pharaoh phrase-based SMT decoder which along with SRI language modelling toolkit[5] performs translation.

4.1 English–French Results

Table 3: Seeding Pharaoh with Giza++ and EBMT sub-sentential alignments for English–French.

		BLEU	Prec.	Recall	WER	SER
TS3	*GIZA-DATA*	.3753	.6598	.5879	58.5	86.82
	EBMT-DATA	.3643	.6661	.5759	61.33	87.99

We seeded the phrase-based SMT system constructed from the publicly available resources listed above with the word- and phrase-alignments derived via both Giza++ and the Marker-Based EBMT system of (Gough & Way, 2004b). Using the full 203K training set of (Gough & Way, 2004b), and testing on their near 4K test set, the results are given in Table 3. It is clear to see that the Giza++ alignments obtain better scores than the EBMT sub-sentential data. Before one considers the full impact of these results, one should take into account that the size of the EBMT data set (word- and phrase-alignments) is 403,317, while there are over four times as many SMT sub-sentential alignments (1,732,715).

Comparing these results with those in Table 1, we can see that for the same training-test data, the phrase-based SMT system outperforms the WB-SMT system on most metrics, considerably so with respect to BLEU score (.3753 vs. .3223). WER, however, is somewhat worse (.585 vs. .535), and SER remains disappointingly high. Compared to the EBMT system of (Gough & Way, 2004b), the phrase-based SMT system still falls well short with respect to BLEU score (.4409 for EBMT vs. .3573 for SMT), and again, notably for SER (.656 EBMT, .868 SMT).

4.2 French–English Results

Table 4: Seeding Pharaoh with Giza++ and EBMT sub-sentential alignments for French–English.

		BLEU	Prec.	Recall	WER	SER
TS3	*GIZA-DATA*	.4198	.6527	.7100	62.93	82.84
	EBMT-DATA	.3952	.6151	.6643	74.77	86.21

Again, the phrase-based SMT system was seeded with the Giza++ and EBMT alignments, trained on the full 203K training set, and tested on the 4K test set. The results are given in Table 4. As for English–French, the Giza++ alignments obtain better scores than when the EBMT sub-sentential data is used.

Comparing these results with those in Table 2, we see that the phrase-based SMT system actually does worse than WB-SMT, which is an unexpected result[6]. As expected, therefore, the results for phrase-based SMT here are worse still compared to EBMT.

5 Towards Hybridity: Merging SMT and EBMT Alignments

We decided to experiment further by combining parts of the EBMT sub-sentential alignments with parts of the data induced by Giza++. In the following sections, for both English–French and French–English, we seed the Pharaoh phrase-based SMT system with:

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

[6] The Pharaoh system is untuned, so as to provide an easily replicable baseline for other similar research. It is quite possible that with tuning the phrase-based SMT system will outperform the word-based system.

1. the EBMT phrase-alignments with the Giza++ word-alignments;
2. all the EBMT and Giza++ sub-sentential alignments (both words and phrases).

5.1 Giza++ Words and EBMT Phrases

Here we seeded Pharaoh with the word-alignments induced by Giza++ and the EBMT phrasal chunks only (i.e. no Giza++ phrases and no EBMT lexical alignments).

5.1.1 English–French Results

Table 5: Seeding Pharaoh with Giza++ word and EBMT phrasal alignments for English–French.

	BLEU	Prec.	Recall	WER	SER
TS3	.3962	.6773	.5913	59.32	85.43

Using the full 203K training set of (Gough & Way, 2004b), and testing on their near 4K test set, the results are given in Table 5. Comparing these figures to those in Table 3, we can see that all automatic evaluation metrics improve with this hybrid system configuration. Note that the data set size is 430,336, compared to 1.73M for the phrase-based SMT system seeded solely with Giza++ alignments. With respect to the EBMT system *per se* in Table 1, these results remain slightly below those figures (except for precision).

5.1.2 French–English Results

Table 6: Seeding Pharaoh with Giza++ word and EBMT phrasal alignments for French–English.

	BLEU	Prec.	Recall	WER	SER
TS3	.4265	.6424	.6918	68.05	83.40

Running the same experimental set up for the reverse language direction gives the results in Table 6. While recall drops slightly, all the other metrics show a slight increase compared to the performance obtained when Pharaoh is seeded with Giza++ word- and phrase-alignments (cf. Table 4).

5.2 Merging All Data

The following two experiments were carried out by seeding Pharaoh with *all* the EBMT and Giza++ sub-sentential alignments, i.e. both words and phrases.

5.2.1 English–French Results

Table 7: Seeding Pharaoh with all Giza++ and EBMT sub-sentential alignments for English–French.

	BLEU	Prec.	Recall	WER	SER
TS3	.4259	.7026	.6099	54.26	83.63

Inserting all Giza++ and EBMT data into Pharaoh's knowledge sources gives the results in Table 7. These are considerably better than the scores for the 'semi-hybrid' system described in section 5.1.1. This indicates that a phrase-based SMT system is likely to perform better when EBMT word- and phrase-alignments are used in the calculation of the translation and target language probability models. Note, however, that the size of the data set increases to over 2M items. Despite this, compared to the results for the EBMT system of (Gough & Way, 2004b) shown in Table 1, these results for the 'fully hybrid' SMT system still fall somewhat short (except for Precision: .6727 vs. .7026).

5.2.2 French–English Results

Table 8: Seeding Pharaoh with all Giza++ and EBMT sub-sentential alignments for French–English.

	BLEU	Prec.	Recall	WER	SER
TS3	.4888	.6927	.7173	56.37	78.42

Carrying out a similar experiment for the reverse language direction gives the results in Table 8. This time this hybrid SMT system does outperform the EBMT system of (Gough & Way, 2004b), with respect to BLEU score (.4888 vs .4611) and Precision (.6927 vs. 6782), but the EBMT system still wins out where Recall, WER and SER are concerned. Regarding this latter, it seems that the correlation between low SER and high BLEU score is not as important as is claimed in (Way and Gough, 2005).

6 Conclusions

(Way and Gough, 2005) carried out a number of experiments designed to test their large-scale Marker-Based EBMT system described in (Gough & Way, 2004b) against a WB-SMT system constructed from publicly available tools. While the results were a little mixed, the EBMT system won out overall.

Nonetheless, WB-SMT has long been abandoned in favour of phrase-based models. We extended the work of (Way and Gough, 2005) by performing a range of experiments using the Pharaoh phrase-based decoder. Our main observations are as follows:

- Seeding Pharaoh with word- and phrase-alignments induced via Giza++ generates better results than if EBMT sub-sentential data is used.

- Seeding Pharaoh with a 'hybrid' dataset of Giza++ word alignments and EBMT phrases improves over the baseline phrase-based SMT system primed solely with Giza++ data. This would appear to indicate that the quality of the EBMT phrases is better than the SMT phrases, and that SMT practitioners should use EBMT phrasal data in the calculating of their language and translation models, if available.

- Seeding Pharaoh with *all* data induced by Giza++ and the EBMT system leads to the best-performing hybrid SMT system: for English–French, as well as EBMT phrasal data, EBMT word alignments also contribute positively, but the EBMT system *per se* still wins out (except for Precision); for French–English, however, our hybrid Example-Based SMT system outperforms the EBMT system of (Gough & Way, 2004b) (cf. Table 9).

Table 9: Comparing the hybrid phrase-based SMT system using both the full Giza++ and full EBMT data against the EBMT system of (Gough & Way, 2004b) for the full training set (TS3).

		BLEU	**Prec.**	**Recall**	**WER**	**SER**
EN-FR	*HYBRID*	.2971	.6739	.5912	54.9	90.8
	EBMT	.3318	.6525	.6183	54.3	89.2
FR-EN	*HYBRID*	.2971	.6739	.5912	54.9	90.8
	EBMT	.3318	.6525	.6183	54.3	89.2

A number of avenues of further work remain open to us. We would like to extend our investigations into hybrid example-based statistical approaches to machine translation by experiment with seeding the Marker-Based system of (Gough & Way, 2004b) with the SMT data, and combinations thereof with the EBMT sub-sentential alignments, to investigate the effect on translation quality. Given our findings here, we are optimistic that 'hybrid statistical EBMT' will outperform the baseline EBMT system, and that our findings will prompt EBMT practitioners to augment their data resources with SMT alignments, something which to our knowledge is currently not done. In addition, we intend to continue this line of research on different and larger data sets, and for other language pairs.

7 Final Remarks

Finally, as (Way and Gough, 2005) observe, it is difficult to explain why to this day SMT practitioners have not made full use of the large body of existing work on EBMT, from (Nagao, 1984) to (Carl & Way, 2003) and beyond, which has contributed greatly to the field of corpus-based MT.

From its very inception EBMT has made use of a range of sub-sentential data – both phrasal and lexical – to perform translations whereas, until quite recently, SMT models of translation were based on the relatively simple word alignment models of (Brown et al., 1990). With the advent of phrase-based SMT systems the line between EBMT and SMT has become significantly blurred, yet we are still unaware of any papers on SMT which acknowledge their debt to EBMT or which describe their approach as 'example–based'.

Despite it becoming increasingly difficulty to distinguish between EBMT and (phrase–based) SMT models of translation, some differences still exist. Rather than using models of syntax in a *post hoc* fashion, as is the case with most SMT systems, an EBMT model of translation builds in syntax *at its core*. Given this, a phrase–based SMT system is more likely to 'learn' chunks that an EBMT system would not, as the system learns *n*-gram sequences rather than syntactically-motivated phrases *per se*. Furthermore, our research here has demonstrated quite clearly that if available, merging SMT and EBMT data improves the quality of the resulting hybrid SMT system, as phrases extracted by both methods that are more likely to function as syntactic units (and therefore be more beneficial during the translation process) are given a higher statistical significance. Conversely, the probabilities of those 'less useful' SMT *n*-grams that are not also gener-

ated by the EBMT system are reduced. Essentially, the EBMT data helps the SMT system to make the best use of phrase alignments during translation.

Moreover, we see the fact that it is becoming increasingly difficult to describe the differences between EBMT and SMT as a good thing, and that as here, this convergence can lead to hybrid systems capable of outperforming leading EBMT systems as well as state-of-the-art phrase-based SMT.

We hope that the research presented here, together with that begun by (Way and Gough, 2005), will lead to new areas of collaboration between both sets of researchers, to the clear benefit of the MT research community and the wider public.

Acknowledgements

We would like to thank Nano Gough for supplying us with our EBMT training data. Thanks also to three anonymous reviewers for their insightful comments. The work presented in this paper is partly supported by an IRCSET[7] PhD Fellowship Award.

References

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Fred Jelinek, Robert Mercer, and Paul Roossin. 1990. A statistical approach to machine translation *Computational Linguistics* **16**:79–85.

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

Michael Carl and Andy Way (eds). 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer, Dordrecht, The Netherlands.

Nano Gough and Andy Way. 2004. Example-Based Controlled Translation. In *Proceedings of the Ninth EAMT Workshop*, Valetta, Malta, pp.73–81.

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

[7]http://www.ircset.ie

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

Mary Hearne and Andy Way. 2003. Seeing the Wood for the Trees: Data-Oriented Translation. In *MT Summit IX*, New Orleans, LA., pp.165–172.

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

Philipp Koehn, Franz Och, and Dan Marcu. 2003. Statistical Phrase-Based Translation. *Human Language Technology Conference, (HLT-NAACL)*, Edmonton, Canada, pp.48–54.

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

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

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* **29**:19–51.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA., pp.311–318.

Joseph Turian, Luke Shen and Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *MT Summit IX*, New Orleans, LA., pp.386–393.

Hideo Watanabe, Sadao Kurohashi and Eiji Aramaki. 2003. Finding Translation Patterns from Paired Source and Target Dependency Structures. In M. Carl & A. Way (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.397–420.

Andy Way and Nano Gough. 2005. Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering* [in press].