Automatic Evaluation of Generation and Parsing for Machine Translation with Automatically Acquired Transfer Rules

Yvette Graham **Deirdre Hogan** **Josef van Genabith**
National Centre for Language Technology
School of Computing, Dublin City University
Dublin 9, Ireland
{ygraham,dhogan,josef}@computing.dcu.ie

Abstract

This paper presents a new method of evaluation for generation and parsing components of transfer-based MT systems where the transfer rules have been automatically acquired from parsed sentence-aligned bitext corpora. The method provides a means of quantifying the upper bound imposed on the MT system by the quality of the parsing and generation technologies for the target language. We include experiments to calculate this upper bound for both handcrafted and automatically induced parsing and generation technologies currently in use by transfer-based MT systems.

1 Introduction

Automatic methods of evaluation for MT include BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), GTM (Turian et al., 2003), TER (Snover et al., 2006) and dependency-based evaluation (Owczarzak et al., 2007). Each of these evaluation methods gives an overall result for the entire MT system, based on a comparison of the sentence output by the MT system with a reference sentence. Unlike other approaches to MT, such as Statistical Machine Translation, transfer-based MT involves three main components: parsing, transfer and generation and each of these contributes to the errors produced by the MT system. Transfer-based MT systems rely heavily on the quality of the parsing and generation components. In order to understand fully the overall results of such a system, the quality of the parsing and generation components should also be tested in isolation to the MT system. However, previous work in the area of transfer-based MT, for example (Furuse and Hitoshi, 1992; Meyers et al., 1998; Menezes and Richardson, 2001; Riezler and Maxwell, 2006), have relied solely on mainstream MT evaluation methods and have not included any breakdown of results for the parsing and generation components of the system.

Existing methods of evaluating parsing and generation technologies as stand-alone systems, however, are insufficient for evaluating how well such technologies will perform as part of a transfer-based MT system, as they do not take into account the fact that the MT system relies on the degree to which the parsing and generation technologies perform together. In addition, they give no indication of how well the generation and parsing technologies will perform when tested on MT data which, in the case of statistical parsers and generators, can be of a very different domain to the domain of the parser and generator training data. Finally, current methods for evaluating sentence generators (Langkilde-Geary, 2002; Callaway, 2003; Nakanishi et al., 2005; Cahill and van Genabith, 2006), rely on gold standard structures for creating input to the generator. The standard of the inputs to the generator is therefore unrealistically high and generator results will not adequately reflect how well the generator might perform in an MT setting.

This paper presents a new methodology for testing sentence generators which gives a more realistic evaluation of how well the generator might fare

as a component of an MT system. In addition, the methodology provides a combined evaluation of parsing and generation components of an MT system and therefore evaluates how well they work together. The new evaluation technique also gives a means of quantifying the upper bound imposed on the performance of a transfer-based MT system by the target language parsing and generation technologies components.

The paper is structured as follows: Section 2 describes in detail why existing methods for parsing and generation evaluation are not adequate with respect to transfer-based MT evaluation. Section 3 proposes a new method of evaluation for parsing and generation for transfer-based MT, where transfer rules have been automatically acquired from sentence-aligned automatically parsed bitext corpora. Section 4 details experiments in which our method is used to evaluate two different parsing and generation technologies for English. Section 5 discusses results and, finally, section 6 gives some conclusions of the work presented in this paper.

2 Existing Generation and Parsing Evaluation Methods

There is a considerable body of work on sentence realisation from abstract linguistic representation where evaluation has been carried out on standalone generators and independent of any MT system. However, an oft-cited future application for such generators is as the generation component of an MT system. Recently there has been an increasing amount of work in the area of robust, broad coverage sentence generation, tested on newswire text, for example (Langkilde-Geary, 2002; Callaway, 2003; Nakanishi et al., 2005; Cahill and van Genabith, 2006). Abstract semantic/syntactic inputs to these generators were automatically constructed from sections of the hand-crafted Penn Treebank. Sentences were then generated for these inputs and compared to the original sentences, using automatic string comparison metrics such as NIST and BLEU. Testing on previously unseen sections of the Penn Treebank demonstrates to what degree a generator has achieved broad coverage and high accuracy (according to BLEU and NIST scores). However, if we wish to take into account how the generator might fare as a component of a machine translation system, this evaluation methodology is unrealistic in two main respects. Firstly, there is the problem of domain adaptation (well documented for parsing, see for example (McClosky et al., 2006; Foster et al., 2007)). Domain adaptation is particularly relevant for the systems of (Langkilde-Geary, 2002; Nakanishi et al., 2005; Cahill and van Genabith, 2006) which are trained on sections of the Wall Street Journal Penn treebank. It is highly likely that were these generators tested in an MT setting, the testing domain would change[1] and, as in statistical parsing, testing on a domain which differs to the domain of the training data would lead to a deterioration in generation results. In addition, the inputs to the generators are constructed from gold standard (hence near perfect) trees, whereas, as a component of an MT system, the inputs to a generator would be constructed in an entirely automatic process and therefore would presumably be of a lower quality.

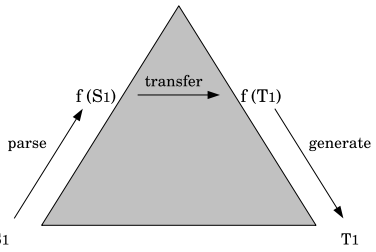

Figure 1: Translation of Source Language Sentence to Target Language

Similar and related issues arise concerning the evaluation of (statistical) parsers which will be used inside MT systems. Parsing technologies are evaluated by a comparison with a gold standard. These gold standards are unlikely to be from the same domain as the MT application testing domain and therefore results would be different were an evaluation done on the new domain. Since parsing evaluation methods do not take generation into account, good results do not ensure that the representation produced by the parser performs well as input to the generator. For example, in statistical parsing and generation, an inconsistency between the structures used for generator training and the parser pro-

[1]For example, the domain of the EUROPARL MT data set, often used in MT, is European parliamentary proceedings.

duced structures could result in the generator underperforming as part of the MT system, which may not be apparent when parsing and generation are evaluated in isolation from one another.

Our method of evaluating parsing and generation for transfer-based MT does not require a gold standard[2] and therefore avoids all the problems related to evaluating on a different domain. It can be easily applied to a new test set and therefore can be used to evaluate the technologies for the MT test set, providing an upper bound for the system for these sentences.

3 Parsing/Generation Evaluation Method for Transfer-based MT

3.1 Transfer-based Machine Translation with Automatically Acquired Transfer Rules

Transfer-based MT with automatically acquired transfer rules is composed of two phases, training and translation. The first phase, training, involves parsing a sentence-aligned bilingual corpus, so that the desired abstract representation of each of the training sentence pairs is obtained (see Figure 2). The parsing technology of the source language is employed to parse the source language sentences, as is the target language technology to parse the target language sentences of the bilingual corpus. Transfer rules/mappings can be automatically induced from the abstract representations of parsed sentence pairs of the corpus. These transfer rules can then be used in phase two to translate from source to target abstract representation for an unseen source language sentence.

In phase two of transfer-based MT (see Figure 1) an unseen raw text sentence is parsed using the source language parsing technologies to get an approximation of the abstract representation for the sentence. The transfer rules that have been automatically induced in phase one are then applied to this representation to produce a target language representation. This target language structure is then given to a generator, so that a target language sentence can be produced.

[2]Note that although the evaluation method does not require a hand-crafted gold standard test set, both parsing and generation technologies require hand-crafted data for training.

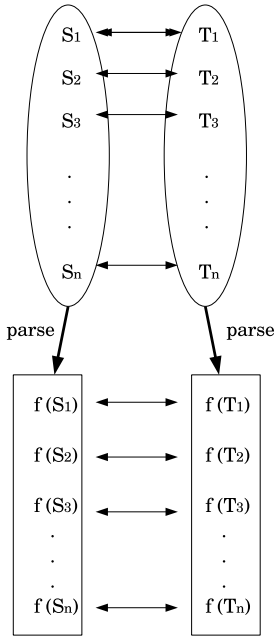

Figure 2: Parsing Bilingual Corpus, $S_n \leftrightarrow T_n$: aligned sentence pair *n* consisting of SL sentence S_n and TL sentence T_n, $f(S_n)$: abstract representation of S_n, $f(T_n)$: abstract representation of T_n

3.2 Parsing/Generation Evaluation Method

The method we propose evaluates the quality of the parsing and generation technologies for the target language of a transfer-based MT system with automatically acquired transfer rules. It involves three simple steps that are carried out automatically on each of the reference translations in the test set:

- Parse
- Generate
- Compare

Each target language sentence in the test section of the bilingual corpus is firstly parsed to the abstract representation. This abstract representation is then given to the generator. The generator produces a string for the abstract representation. The target language sentence is then compared with the generator-produced string. Any string comparison method can be used to compare the two strings, e.g BLEU or NIST.

When this method of evaluation is used to evaluate the target language technologies of a transfer-

based MT system with automatically acquired transfer rules, it provides an upper bound to the results that could be achieved by an MT system which relies on those target language technologies. Recall from Section 3.1 (and as illustrated in Figure 2) that transfer rules are induced by learning the mapping from parser-produced source language to parser-produced target language representations. Given a source sentence, the better the source language parser and the better the transfer component, the closer the abstract target representation (input to the target language generator) is to the output of the target parser. In fact, the 'best' input such a system can theoretically provide to the generator approximates to the output of the target language parser (for target language sentences in the bilingual test/development sections of the bitext corpus). Thus, by parsing, then generating, target language sentences, and comparing the strings output in this process to the original target language strings, we are in effect finding an upper-bound to an MT-system which would rely on these technologies[3].

Having a straightforward means such as this of producing an upper bound for a transfer-based MT system with automatically acquired transfer rules is highly advantageous because it gives a very early indication of whether or not the task of transfer MT is promising using a set of target language parsing and generation technologies. This evaluation method can be used before the transfer system is actually implemented, so that in a case where a very low result is achieved by the target language parsing/generation technologies, it could be discovered that these technologies need improvement before the MT system is implemented.

[3]It is of course possible to artificially set this upper bound to its absolute value to claim that the parsing and generation components could achieve perfect results by making the parser and generator collude with each other. For example, by including the actual surface form and word order in the *abstract* representation, the generation component could without question reproduce exactly the same surface form that was input to the parser. However, this would not be the intended use of our evaluation method, the abstract structure that is given to the generator in our evaluation method should only contain information that is possible to be produced by the transfer component. Such information, like the actual surface form and word order in the target language would not be possible to have in the target language structure if it adheres to the transfer-based architecture that we describe in Section 3.1.

4 Experimental Results

We conducted two sets of experiments to evaluate the quality of hand-crafted LFG parsing and generation technologies (Riezler et al., 2002) and the treebank-induced LFG parsing (Cahill et al., 2004) and generation (Cahill and van Genabith, 2006; Hogan et al., 2007) technologies using our new method. Both the automatically-induced technologies and the hand-crafted technologies are currently being used as part of transfer-based MT systems. Overall evaluation results for one of these MT systems are available and we include these results in Section 4.2 (Riezler and Maxwell, 2006). We evaluate the parsing/generation technologies on the English language components of three different MT bitext test sets: the 1755 English sentences of length 5 to 15 of the Europarl test set used in Koehn et al. (Koehn et al., 2003), that were also used to evaluate the MT system; the first 500 English sentences of all lengths used in (Koehn et al., 2003); and 766 English sentences of the Homecentre corpus of all lengths. We also include the results achieved by the automatically induced resources using existing methods of generation evaluation on Section 23 of the Wall Street Journal so that a comparison can be made between these and the results of the new method on the same test set.

We give results of each experiment for two different types of evaluation: *entire test set* and *in-coverage only*. The *in-coverage only* method follows that typically applied when reporting generation results (Langkilde-Geary, 2002; Callaway, 2003; Nakanishi et al., 2005; Cahill and van Genabith, 2006; Hogan et al., 2007): we give the BLEU and NIST scores for the sentences for which output was produced and we report coverage to indicate the percentage of sentences for which output was produced. The *entire test set* evaluation gives BLEU and NIST scores for the entire test set, regardless of whether any output was produced for some of the sentences. Where no output was generated for a sentence, we include this empty string in the output test set, to be compared with the reference sentence for this particular segment. When comparing two different systems the *entire test set* evaluation gives a more realistic evaluation of how two systems compare. It is not possible to give a meaningful com-

parison of results across systems that do not have full coverage using *in-coverage only* evaluation as results are not necessarily for the same set of sentences.

4.1 Evaluation of Automatically Induced Resources

Using the methodology described in detail in Section 3, we evaluated the history-based statistical generator of Hogan et al. (2007). This generator, an extension of the work presented in (Cahill and van Genabith, 2006), generates sentences from LFG f-structures and achieves state-of-the-art results when tested on input generated from Penn Treebank gold standard trees. In order to generate f-structure inputs in a completely automated fashion for evaluation, we used the Charniak and Johnson (2005) reranking parser to parse the original test sentences into Penn Treebank style trees [4]. The f-structure annotation algorithm of Cahill et al. (2004) was then applied to the parser-generated trees to create a set of f-structures for testing.

Section 23 (2416 sentences)		
Input	NIST	BLEU
From Gold-standard Trees	13.29	0.6680
From Parser Trees	13.01	0.6511

Table 1: *Entire Test Set* Results on Section 23 of the Penn Treebank for the generator of (Hogan et al., 2007) on input automatically generated from gold standard and from parser generated trees.

Section 23 (2416 sentences)			
Input	NIST	BLEU	Coverage
from Gold Trees	13.31	0.6693	99.88%
from Parser Trees	13.02	0.6515	99.96%

Table 2: *In-Coverage only* Results on Section 23 of the Penn Treebank for the generator of (Hogan et al., 2007) on input automatically generated from gold standard and from parser generated trees.

Table 1 gives the *entire test set* BLEU and NIST scores for Section 23 of the Penn Treebank for sentences generated from gold-standard generated f-structures compared with results for the same sentences generated from parser-generated f-structures. As anticipated, generation accuracy declines when f-structure inputs are generated from parser trees rather than from gold standard trees. Both the decrease in BLEU score from 0.6680 to 0.6511 and NIST score from 13.29 to 13.01 are statistically significant ($p < 0.0001$), where significance is calculated using a bootstrap resampling method with a resampling rate of 1000.[5] Table 2 gives the same results using conventional *in-coverage only* evaluation. There is little difference between *entire test set* and *in-coverage only* scores because coverage is so high.

	NIST	BLEU
WSJ Section 23	13.01	0.6511
Europarl (5-15)	11.72	0.6968
Europarl (all)	10.24	0.5716
Homecentre	10.06	0.6640

Table 3: *Entire Test Set* Parsing/Generation Results for Treebank-induced technologies of Cahill et al. (2004) and Cahill et al. (2006)

We evaluated the parsing and generation resources further on the English sentences of the three MT test sets described Section 4. The NIST and BLEU results of the parsing/generation evaluation are given in Tables 3 and 4. Table 3 shows the *entire test set* evaluation scores for each of the English MT test sets, as well as results for Section 23 of the Wall Street Journal. Table 4 shows the *in-coverage only* NIST and BLEU evaluation results for the same test sets [6].

4.2 Evaluation of Hand-crafted Resources

The English hand-crafted grammar of Riezler et al. (2002) with XLE (Kaplan et al., 2002) (software for parsing to and generating from LFG f-structures) was previously used as the target language analysis and generation technology of a transfer-based

[4] Note that the parser generated phrase structure trees contain less information than Penn Treebank trees in that they do not contain empty nodes and trace information, nor do they give any Penn-II functional annotation tags on nodes.

[5] Scripts for running the bootstrapping method are available at projectile.is.cs.cmu.edu/research/public/tools/bootStrap/tutorial.htm.

[6] We repeat the inclusion of the WSJ results in Tables 3 and 4 for ease of comparison.

	NIST	BLEU	Coverage
WSJ Section 23	13.02	0.6515	99.96%
Europarl (5-15)	11.72	0.6968	100%
Europarl (all)	10.24	0.5716	100%
Homecentre	10.06	0.6640	100%

Table 4: *In-coverage only* Parsing/Generation Results for Treebank-induced technologies of Cahill et al. (2004) and Cahill et al. (2006)

MT system (Riezler and Maxwell, 2006). This MT system uses SMT phrasal alignments to hypothesise candidate transfer rules from source language f-structures to target language f-structures. In addition, the system models and trains statistical components using SMT techniques. Even though the system takes advantage of these statistical methods to hypothesise candidate transfer rules and to choose the best target language translation, it is based on the classic transfer-based MT architecture described in Section 3.1 and therefore our method of evaluation can be used to quantify the upper bound that the target language parsing and generation technologies impose on the system.

We evaluated the English parsing and generation technologies employed by this MT system using our evaluation method on the English sentences of the three MT test sets[7]. The NIST result of the MT system as well as the NIST and BLEU results of the parsing/generation evaluation are given in Tables 5 and 6. Table 5 shows the NIST results for German-to-English translation of the MT system of Riezler et al. (2006) on the Europarl parallel data test set used in Koehn et al.(2003) [8]. The table also shows the *entire test set* NIST and BLEU evaluation scores for the target language (English) parsing and generation technologies on the English sentences of this test set as well as corresponding scores for the two other test sets. Table 6 shows the *in-coverage only* NIST and BLEU evaluation results for the same three test sets.

5 Discussion

The results of the experiments on the WSJ Section 23, given in Tables 1 and 2, show that there is a significant difference between generation scores when sentences are generated from hand-crafted gold structures compared to when they are generated from automatically produced structures. In addition, the results in Tables 3, 4, 5 and 6 show that the upper bound imposed by the parsing and generation technologies changes dramatically with each test set. The results in Table 3 show the effect of the change of test domain on the Penn Treebank trained parsing and generation technologies. These technologies achieve a high NIST and BLEU score when tested on Penn Treebank data (NIST: 13.01, BLEU: 0.6511). The most accurate estimate of the effect of change of domain from the WSJ to the Europarl data can be seen by comparing these results to the results achieved on 500 Europarl sentences of all lengths, as these two test sets are most similar with respect to sentence length out of all of the test sets used. Here, the NIST score falls from 13.01 for the WSJ text to 10.24 for the Europarl text and BLEU from 0.6511 to 0.5716. When the Europarl test set is restricted to sentences of length 5 to 15 the treebank-trained technologies achieve their highest scores outside the training domain (NIST:11.72, BLEU:0.6968) and this can be accredited to the elimination of long sentences. The Homecentre corpus evaluation showed a slightly lower result (NIST:10.06, BLEU:0.6640) than the Europarl (5-15) (NIST: 11.72, BLEU:0.6968) even though this corpus consists of mainly short sentences. This could be due to the domain of language of the corpus being a printer instruction manual, which contains many imperative sentences which are infrequent in the newspaper domain of the training text.

The results shown in Table 5 for the hand-crafted XLE resources show a dramatic difference between the NIST score achieved by the MT system (NIST:5.62) and the upper bound imposed by

[7]For occasions when there is more than one possible surface form realisation of a given f-structure, the XLE generator produces a packed representation of these sentences. By setting an XLE parameter to either *shortest* or *longest*, a single sentence is selected. For the experiments detailed here the parameter was set to *shortest*, so that in the case of more than one possible surface form, the shortest sentence was selected. Experiments were also carried out with the parameter set to *longest*, but since results were slightly lower when the longest sentence is selected we include the results achieved by XLE with the parameter set to *shortest*. Alternatively, a language model could be employed for the task of selecting the sentence in such cases, which could possibly improve results.

[8]No BLEU score is included here for the MT system, as none is reported in (Riezler and Maxwell, 2006)

		NIST	BLEU
MT System	Europarl (5-15)	5.62	
Parsing/ Generation Evaluation	Europarl (5-15)	12.08	0.7431
	Europarl (all)	6.33	0.4785
	Homecentre	10.75	0.7523

Table 5: MT system Results of Riezler and Maxwell (2006) and *entire test set* Parsing/Generation Results for hand-crafted technologies of Riezler et al. (2002)

	NIST	BLEU	Coverage
Europarl (5-15)	12.26	0.7800	95%
Europarl (all)	12.1	0.7591	80%
Homecentre	10.81	0.7931	98%

Table 6: *In-Coverage only* Results for hand-crafted technologies of Riezler et al. (2002)

the English parsing and generation technologies (NIST:12.08). As was also seen with the automatically induced technologies, the test set used greatly effects the hand-crafted technologies. Sentence length restriction to 5-15 for the Europarl data has a more dramatic effect on the hand-crafted technologies than the automatically induced technologies, increasing the NIST score from 6.33 to 12.08 and BLEU from 0.4785 to 0.7431. Tables 6 show the *in-coverage only* results for the hand-crafted technologies.

The *entire test set* results for the hand-crafted technologies compared with the treebank-induced technologies on the Europarl data sets show that the hand-crafted technologies achieve a better result for short sentences of length 5 to 15 (NIST:12.08, BLEU:0.7431) than the treebank-induced technologies (NIST:11.72, BLEU:0.6968), whereas the treebank-induced technologies do better when tested on all sentence lengths (NIST:10.24, BLEU:0.5716) than the hand-crafted technologies (NIST:6.33, BLEU:0.4785). The high scores achieved by the hand-crafted technologies on the Homecentre corpus (NIST:10.75, BLEU:0.7523) also indicate they are perhaps better at absorbing the effects of domain variation than the treebank-induced technologies (NIST:10.06, BLEU:0.6640).

6 Conclusion

We presented a method for the evaluation of target-language parsing and generation components of transfer-based MT systems with automatically acquired transfer rules. Transfer-based MT systems rely heavily on the quality of the parsing and generation technologies employed and it is therefore highly advantageous to have a simple, inexpensive and effective way to evaluate these system components to provide a realistic result in relation to the task of MT. Results of parsing and generation technologies using existing evaluation methods are usually for a different domain to that of the MT test set and therefore, such results do not provide a realistic evaluation of how the technologies will perform on a new MT domain. Unlike existing methods of evaluation for parsing and generation, the proposed method can easily be applied to a new domain without requiring expensive gold standards. It also provides a means of quantifying the upper bound imposed by the parsing and generation technologies of the target language given a particular test set. We have shown how this upper bound changes dramatically from one test set to the next depending on domain variation and sentence length. The proposed evaluation method is based on the idea that the target language parsing and generation components of a transfer-based MT system with automatically acquired transfer rules should not be evaluated in isolation from one another. Transfer-based MT not only needs high quality parsing technologies and generation technologies but it also needs the generation technologies to work well when given parser output as their input. Evaluating the technologies in isolation does not take this into account and neither rewards a system that can achieve this nor punishes one that cannot.

7 Acknowledgements

The work presented in this paper was partly funded by a Science Foundation Ireland PhD studentship P07077-60101. We would also like to thank our reviewers and Mary Hearne for their helpful comments.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on "Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization" ACL 2005*, pages 65-73, Ann Arbor, Michigan.

Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-based LFG approximations. In *Proceedings of the 42nd ACL*.

Aoife Cahill and Josef van Genabith. 2006. Robust PCFG-Based Generation using Automatically Acquired LFG Approximations. In *Proceedings of COLING-ACL 2006*, pages 1033-1040, Sydney, Australia.

Charles B. Callaway. 2003. Evaluating Coverage for Large Symbolic NLG Grammars. In *Proceedings of the 18th IJCAI*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best Parsing and MaxEnt Discriminative Reranking. In *Proceedings of ACL 2005*.

George Doddington 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proceedings of HLT 2002*.

Jennifer Foster, Joachim Wagner, Djam Seddah and Josef van Genabith. 2007. Adapting WSJ-Trained Parsers to the British National Corpus using In-Domain Self-Training. In *Proceedings of IWPT 2007*.

Osamu Furuse and IIda Hitoshi. 1992. An Example-Based Method for Transfer-Driven Machine Translation. In *Proceedings of TMI 1992*, pages 139-150, Montral, Canada.

Deirdre Hogan, Conor Cafferkey, Aoife Cahill and Josef van Genabith . 2007. Exploiting Multi-Word Units in History-Based Probabilistic Generation. In *Proceedings of the EMNLP 2007*.

Ronald M. Kaplan, Tracy H. King and John T. Maxwell. 2002. Adapting existing grammars: the XLE experience. In *Proceedings of COLING 2002*, Taipei, Taiwan.

Philip Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL 2003*, pages 48-54, Edmonton, Alberta.

Irene Langkilde-Geary 2002. An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator. In *INLG 2002*.

David McClosky, Eugene Charniak and Mark Johnson 2006. Reranking and Self-Training for Parser Adaptation. In *Proceedings of COLING-ACL*.

Arul Menezes and Stefan D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of Workshop on Data-Driven MT*, Toulouse, France.

Adam Meyers, Roman Yangarber, Ralph Grishman, Catherine Macleod and Antonio Moreno-Sandoval. 1998. Deriving Transfer Rules from Dominence-Preserving Alignments. In *Proceedings of COLING-ACL 1998*, Montreal, Canada.

Hiroko Nakanishi, Yusuke Miyao and Jun'ichi Tsujii 2005. Probabilistic Models for Disambiguation of an HPSG-based Chart Generator. In *Proceedings of IWPT*.

Karolina Owczarzak, Josef van Genabith and Andy Way. 2007. Dependency-based Automatic Evaluation for Machine Translation. To appear in *Proceedings of HLT-NAACL 2007 Workshop on Syntax and Structure in Statistical Translation (SSST)*,

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. A Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311-318, Philadelphia.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using Lexical Functional Grammar and discriminitive estimation techniques . (grammar version 2005) In *Proceedings of ACL 2002)*, Philadelphia, July 2002.

Stefan Riezler and John Maxwell. 2006. Grammatical Machine Translation. In *proceedings of HLT-ACL*, pages 248-255, New York.

Mathew Snover, Bonnie Dorr, Richard Scwartz, John Makhoul and Linnea Micciula. 2006. A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of AMTA 2006*, pages 223-231, Boston, Massachusetts.

Joseph P. Turian, Luke Shen and Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit 2003*, pages 386-393, New Orleans, Luisiana.