Meta-Evaluation of a Diagnostic Quality Metric for Machine Translation

Sudip Kumar Naskar* **Antonio Toral** **Federico Gaspari** **Declan Groves**
Computer and System Sciences
Visva-Bharati University
India
sudip.naskar@visva-bharati.ac.in

School of Computing
Dublin City University
Ireland
{atoral, fgaspari, dgroves}@computing.dcu.ie

Abstract

Diagnostic evaluation of machine translation (MT) is an approach to evaluation that provides finer-grained information compared to state-of-the-art automatic metrics. This paper evaluates DELiC4MT, a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. We present the results obtained using this diagnostic metric when evaluating three MT systems that translate from English to French, with a comparison against both human judgements and a set of representative automatic evaluation metrics. In addition, as the diagnostic metric relies on word alignments, the paper compares the margin of error in diagnostic evaluation when using automatic word alignments as opposed to gold standard manual alignments. We observed that this diagnostic metric is capable of accurately reflecting translation quality, can be used reliably with automatic word alignments and, in general, correlates well with automatic metrics and, more importantly, with human judgements.

1 Introduction

The study presented in this paper addresses the topic of diagnostic evaluation of machine translation (MT), which is receiving increasing attention due to its potentially crucial but still largely unexplored role in the development and subsequent deployment of MT systems. Diagnostic evaluation might be particularly useful to complement the overall system-level scores provided by automatic MT evaluation metrics. On the one hand, these automatic metrics represent cost-effective, objective and easily replicable measures, on the other, they provide only global indications that are normally too coarse to explain the performance of an MT system. An associated issue is that diagnostic evaluation needs to be as fine-grained as possible to be really useful in targeting specific weaknesses detected in MT output, for the system developers to be able to take corrective actions accordingly, and the users to asses the actual impact of the system's weaknesses.

This paper evaluates a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. Focusing on English to French translation as a case study, the use of alternative automatic word alignments is investigated and compared against gold standard manual alignment to discuss how these different approaches impact on the results of diagnostic MT evaluation. The paper also presents a comparative evaluation of three MT systems judged according to standard automatic MT evaluation metrics, the diagnostic evaluation metric over a range of linguistic checkpoints, and human judgements. Additionally, we investigate how these different types of MT evaluation correlate to each other.

The paper is structured as follows. Section 2 presents previous work in diagnostic evaluation of MT, discussing the methodologies and tools that exist in this area, focusing in particular on the features of the diagnostic metric used in this study. Section 3 describes the datasets that were used for the experiments and Section 4 details the experi-

*Work done while at CNGL, School of Computing, Dublin City University.

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 135–142.

mental setup. The results of the investigation are presented and analysed in Section 5, and finally some conclusions are drawn and possible avenues for future work are outlined in Section 6.

2 Related Work

Recognising that the ability to automatically identify and evaluate specific MT errors with diagnostic relevance is of paramount importance, Popović et al. (2006) propose a framework for the automatic classification of MT errors based on morpho-syntactic features. They show that linguistically-sensitive measures provide useful feedback to alleviate the problems encountered by MT. In a similar vein, Popović and Burchardt (2011) present a method for automatic error classification and compare its use with results obtained from human evaluation. They show good correlation between their automatic measures and human judgements across various error classes for different MT output.

Popović (2011) describes a tool for automatic classification of MT errors, which are grouped into five classes (morphological, lexical, reordering, omissions and unnecessary additions). The tool needs full-form reference translation(s) and hypotheses with their corresponding base forms. Additional information at the word level (such as PoS tags) can be used for a more delicate analysis. The tool computes the number of errors for each class at the document and sentence levels.

Max et al. (2010) propose an approach to contrastive diagnostic MT evaluation based on comparing the ability of different systems (or implementations of the same system) to correctly translate source-language words. Their contrastive lexical evaluation method does not rely on the direct comparison of the system's hypotheses with the reference translations, but for each source-language word it identifies which of the MT systems under consideration provide the correct output matching the reference. Their study is devoted to English–French and they point out the crucial role played by the quality of the alignment, suggesting that inaccuracies in the automatic alignment are bound to impair the reliability of this approach for lexical diagnostic evaluation.

Fishel et al. (2012) provide an overview of the field of diagnostic evaluation of MT, presenting a collection of freely available translation error-annotation corpora for various language pairs and comparing the performance of two state-of-the-art tools on automatic error analysis of MT.

Zhou et al. (2008) describe a tool for diagnostic MT evaluation called Woodpecker,[1] which is based on linguistic checkpoints. These are particularly interesting (or problematic) linguistic phenomena for MT processing identified by the user or developer who conducts the evaluation, e.g. ambiguous words, challenging collocations or PoS-n-gram constructs, etc. One needs to define a linguistic taxonomy which describes the phenomena to be captured in the diagnostic evaluation, deciding which elements of the source language one wants to investigate. This scheme is extremely flexible, and can be formulated at different levels of specificity, whereby the granularity of the checkpoints included depends on the objectives of the diagnostic evaluation.

While the notion of linguistic checkpoints is very useful within the context of diagnostic MT evaluation, Woodpecker has some limitations. First of all, language-dependent data for English–Chinese (the language pair covered in the study presented in (Zhou et al., 2008)) is hardcoded in the software, which therefore cannot be easily adapted to other language pairs. In addition, the licence with which Woodpecker is distributed (MSR-LA)[2] is quite restrictive, in that e.g. researchers cannot publicly release their own adaptations of the tool.

DELiC4MT[3](Toral et al., 2012) is a free open-source tool for diagnostic evaluation which offers similar functionality to Woodpecker. We chose to carry out experiments with DELiC4MT due to its language-independent nature. This recall-based diagnostic evaluation metric essentially works like other n-gram-based automatic MT evaluation metrics (i.e. counting n-gram matches between the MT output and the reference translations), except that it focuses on specific segments of the reference identified through linguistic constructs found in the source (i.e. linguistic checkpoints) and word alignment.

[1] http://research.microsoft.com/en-us/downloads/ad240799-a9a7-4a14-a556-d6a7c7919b4a/

[2] https://research.microsoft.com/en-us/projects/pex/msr-la.txt

[3] http://www.computing.dcu.ie/~atoral/delic4mt/

The final recall score produced by DELiC4MT is computed as in equation 1, where *R* is the set of references (*r*) of all the checkpoints (*c*) in *C*. A length-based penalty is introduced to penalise longer candidate translations (otherwise longer translations would have a better chance of returning higher scores) as in equation 2, where *length*(*C*) is the average candidate translation length and *length*(*R*) is the average reference translation length.

$$R(C) = \frac{\sum_{r \in R} \sum_{n-gram \in r} match(n-gram)}{\sum_{r \in R} \sum_{n-gram \in r} count(n-gram)} * penalty \quad (1)$$

$$penalty = \begin{cases} \frac{length(R)}{length(C)} & \text{if } length(C) > length(R) \\ 1 & otherwise \end{cases} \quad (2)$$

3 Datasets

The initial key decisions that had to be made to set up the experiment concerned the languages to be focused on as well as the domain and specific dataset to be selected for the investigation.

We decided to work on English–French, as human-aligned datasets are readily available for this language pair. We investigated a number of options in terms of manually annotated aligned English–French data to serve as gold standard, and considered, for example, using Biblical texts made available as part of the Blinker Annotation Project (Melamed, 1998). However, the syntax and vocabulary of this dataset presented some specific features which were not in line with actual uses envisaged for diagnostic evaluation in research or industrial settings.

The dataset that was chosen for our experiment was initially created for the shared task on word alignment held as part of the HLT/NAACL 2003 Workshop on Building and Using Parallel Texts (Mihalcea and Pedersen, 2003). The dataset used for this study consists of 447 English–French word-aligned sentence pairs drawn from the Canadian Hansard Corpus, consisting of parliamentary debates (Och and Ney, 2000), for a total of 7,020 tokens in English and 7,761 in French. It should be noted that we did not differentiate between 'sure' and 'probable' word alignments in this dataset and treat them as having the same weight.

Choosing a bilingual dataset from the domain of parliamentary speeches allowed us to conduct a fair and direct comparison with a closely related baseline English–French MT system built using the Europarl corpus[4] (Koehn, 2005).

4 Experimental Setup

4.1 MT Systems

We experimented with three MT systems: Google Translate[5], Systran[6] and a baseline Moses[7] system. Among the three MT systems, Google Translate and Moses are statistical MT systems while Systran is predominantly a rule-based system. The Moses system used for our experiments was trained on 3.6 million English–French sentence pairs taken from Europarl, the News Commentary corpus and a randomly selected section of the UN corpus. The system was tuned on a held-out development set consisting of 1,025 sentence pairs and used a 5-gram language model built using the SRILM toolkit (Stolcke, 2002).

4.2 Word Alignment

The diagnostic evaluation was carried out using both gold standard human alignments and three sets of automatic alignments. Thus, in total we carried out experiments on 4 different sets of word alignments. The idea behind this study was primarily to show whether the different possible alignments had an impact on the effectiveness of the diagnostic MT evaluation metric, also in comparison with gold-standard manual alignment and human evaluation.

We used GIZA++[8] (Och and Ney, 2003) to derive the automatic alignments between the source and target sides of the testset. We extracted three sets of alignments using the union, intersection and grow-diag-final heuristics, as implemented by the Moses training scripts. Since the testset is far too small to be accurately word-aligned using a statistical word-aligner and would suffer from data sparseness, additional parallel training data from the Europarl corpus was used. The additional training data was first tokenised, filtered (using source-target length ratio) and lower-cased. The testset was also subjected to tokenisation and lower-casing. The testset was then appended with

[4] http://www.statmt.org/europarl/
[5] http://translate.google.com
[6] http://www.systran.co.uk/
[7] http://www.statmt.org/moses/
[8] http://code.google.com/p/giza-pp/

the additional training data and word-aligned using GIZA++. Finally, from the word-alignment file only the word alignments for the sentences that correspond to the testset were extracted.

4.3 Linguistic Checkpoints

Regarding the linguistic phenomena, we considered a basic set of PoS-based checkpoints: adjectives (a), nouns (n), verbs (v), adverbs (r), determiners (dt), miscellaneous (misc), and pronouns (pro). The 'misc' checkpoint contains a variety of other PoS tags (CC, IN, RP and TO) (Santorini, 1990). We used Treetagger[9] (Schmid, 1994) to PoS-tag both sides of the testset.

It should be noted that the evaluation framework can potentially focus on more complex user-defined linguistic phenomena. In fact, it can be applied to a wide range of composite linguistic structures of interest to the MT developer or user for evaluation purposes. The metric can handle, e.g., combinations of literal words or lemmas with PoS tags. Evaluation on named entities and dependency structures is also supported by this diagnostic MT evaluation metric.

4.4 Human Judgements

In order to verify the results of the diagnostic evaluation, we carried out human evaluations on the output of the 3 different MT systems. These were done by 2 evaluators, both native French speakers and experienced in translation evaluation. They were asked to assign fluency and adequacy scores to the translations based on a discrete 5-point scale (LDC, 2005). In addition, they were asked to evaluate translation quality in terms of 5 PoS-based checkpoints (a, n, v, r and dt), again using a 5-point scale, with 1 representing instances where there were severe errors in the translation of all instances of the checkpoint and 5 indicating that all instances were translated perfectly. The evaluators were also asked to give a does-not-apply ('NA') score to sentences that did not contain the linguistic phenomenon under consideration.

5 Results

5.1 Diagnostic Evaluation

Table 1 shows the diagnostic evaluation results obtained on the gold standard word alignment. Tables 2, 3 and 4 present results obtained on the grow-diag-final, union and intersection alignments respectively. Each of these tables shows checkpoint-specific scores across systems. Table 1 shows in addition the number of checkpoint-specific instances (#Inst) extracted from the source side of the testset.

Checkpoint-specific statistically significant improvements are reported in these tables as superscripts. For representation purposes, we use *a*, *b*, and *c* for Google, Moses and Systran, respectively. For example, the Google score $0.4993^{b,c}$ for the adjective checkpoint in Table 1 means that the improvement provided by Google for this checkpoint is statistically significant over both Moses and Systran.

In addition to the checkpoint-specific scores, each of these tables provides an arithmetic mean (avg) and a weighted mean (w-avg, weighted by the number of instances for each checkpoint). The weighted average is considered as the system-level score for diagnostic evaluation. Tables 2, 3 and 4 also show the ratios with manual alignment (m-ratio). For example, Table 2 shows that the weighted means obtained by Google, Moses and Systran on grow-diag-final alignments are respectively 0.7337, 0.7132 and 0.7126 times those obtained on manual alignments.

	#Inst	Systems		
		Google	Moses	Systran
a	426	$0.4993^{b,c}$	0.4345	0.4369
n	1,649	$0.5420^{b,c}$	0.5025	0.5013
v	1,296	0.4037^{c}	0.3974^{c}	0.3603
r	348	0.4462	0.4198	0.4352
dt	824	$0.5968^{b,c}$	0.5479	0.5718
misc	1,079	$0.5788^{b,c}$	0.5376	0.5367
pro	428	$0.5740^{b,c}$	0.5049	0.5415
avg	6,050	0.5201	0.4778	0.4834
w-avg		0.5201	0.4831	0.4815

Table 1: Diagnostic evaluation results on manual alignments

As the scores in Table 1 suggest, Google clearly outperforms the other systems on all of the phenomena, and most of these improvements are statistically significant. The Moses baseline system performs slightly better than Systran according to the weighted averages. While some of the phenomena (e.g., nouns, verbs) are better handled by

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

the Moses baseline system the scores in Table 1 also show that Systran performs quite better than this baseline system for adverbs, determiners and pronouns. This trend can be observed across Tables 2, 3 and 4 as well.

	Systems		
	Google	Moses	Systran
a	$0.3056^{b,c}$	0.2591	0.2440
n	$0.3374^{b,c}$	0.2958	0.2896
v	0.2583^{c}	0.2483^{c}	0.2272
r	0.3266^{b}	0.3061	0.3016
dt	$0.5117^{b,c}$	0.4621	0.4853
misc	$0.5199^{b,c}$	0.4698	0.4676
pro	0.4465^{b}	0.3976	0.4450
avg	0.3866	0.3484	0.3515
w-avg	0.3816	0.3445	0.3431
m-ratio	0.7337	0.7132	0.7126

Table 2: Diagnostic evaluation results on grow-diag-final alignments

	Systems		
	Google	Moses	Systran
a	$0.2748^{b,c}$	0.2281	0.2195
n	$0.3108^{b,c}$	0.2690	0.2650
v	0.2423^{c}	0.2305^{c}	0.2113
r	0.3191^{b}	0.3016	0.2937
dt	$0.4787^{b,c}$	0.4324	0.4552
misc	$0.4916^{b,c}$	0.4453	0.4447
pro	0.4281^{b}	0.3865	0.4272
avg	0.3636	0.3276	0.3309
w-avg	0.3575	0.3218	0.3214
m-ratio	0.6873	0.6661	0.6674

Table 3: Diagnostic evaluation results on union alignments

5.2 Automatic Metrics

We also evaluated the performances of the MT systems using a set of state-of-the-art automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). Table 5 presents the system-level evaluation results for the different types of metrics considered (automatic, diagnostic and human judgements). For diagnostic evaluation it reports the weighted averages (see w-avg in Tables 1, 2, 3 and 4). According to BLEU, NIST and METEOR, Google

	Systems		
	Google	Moses	Systran
a	$0.5126^{b,c}$	0.4365	0.4365
n	$0.5494^{b,c}$	0.5042	0.4989
v	0.4074^{c}	0.4261^{c}	0.3496
r	0.5768	0.5431	0.5603
dt	$0.6529^{b,c}$	0.5926	0.6248
misc	$0.7195^{b,c}$	0.6628	0.6542
pro	$0.6331^{b,c}$	0.5493	0.6030
avg	0.5788	0.5307	0.5325
w-avg	0.5683	0.5285	0.5183
m-ratio	1.0926	1.0940	1.0764

Table 4: Diagnostic evaluation results on intersection alignments

is the best system, followed by Moses and Systran, while TER ranks Systran over Moses. Diagnostic evaluation on gold standard alignment also yields the same ranking as BLEU, NIST and METEOR. More importantly, for this work, the use of any automatically derived word alignments (i.e., grow-diag-final, union or intersection) in diagnostic evaluation replicates the same ranking obtained with gold standard alignments.

Method		Systems		
		Google	Moses	Systran
Diagnostic	manual	0.5201	0.4831	0.4815
	gdf	0.3816	0.3445	0.3431
	union	0.3575	0.3218	0.3214
	intersection	0.5683	0.5285	0.5183
Automatic	BLEU	0.2012	0.1621	0.1471
	NIST	5.11	4.54	4.44
	METEOR	0.5033	0.4390	0.4258
	TER	0.6508	0.7059	0.6980
Human	Evaluator 1	3.7864	3.3658	3.4497
	Evaluator 2	4.2417	3.9989	4.0503

Table 5: System level evaluation results

5.3 Human Judgements

Tables 6 and 7 present the results of human evaluation of the MT systems. The mean of adequacy (adq) and fluency (fln) is considered as the overall human judgement score. According to both evaluators, at the system level, Google is the best system, followed by Systran and Moses. As far as fine-grained checkpoint-specific

human judgements are concerned, both evaluators agree that Google handles all of the linguistic phenomena better than the other two systems, as is also revealed by the diagnostic evaluation. According to evaluator 1, Moses translates nouns better than Systran does, while Systran does well on adjectives, verbs, adverbs and determiners. Diagnostic evaluation matches almost perfectly with fine-grained checkpoint-specific human judgements obtained from evaluator 1, except for the translation of verbs for Moses and Systran. But, according to evaluator 2, Moses only translates determiners better than Systran, while Systran does better on the rest of the checkpoints.

	Google	Moses	Systran
Adq	3.9284	3.4765	3.5906
Fln	3.6443	3.2550	3.3087
Avg (Adq, Fln)	3.7864	3.3658	3.4497
noun	4.4758	4.0435	4.0097
verb	4.2138	3.9430	4.1900
adverb	4.6171	4.3237	4.4138
adjective	4.2296	3.8163	4.0302
determiner	4.7754	4.4727	4.7578

Table 6: Human judgements of evaluator 1

	Google	Moses	Systran
Adq	4.6578	4.6577	4.6711
Fln	3.8255	3.3400	3.4295
Avg (Adq, Fln)	4.2417	3.9989	4.0503
noun	4.4734	4.2302	4.3318
verb	4.4143	4.3868	4.4043
adverb	4.6507	4.4855	4.4908
adjective	4.6324	4.4542	4.5210
determiner	4.3605	4.0937	4.0047

Table 7: Human judgements of evaluator 2

Table 8 presents Pearson's correlations for checkpoint-specific evaluation across systems. It shows that checkpoint-specific diagnostic evaluation using either manual or automatic alignments correlates well with checkpoint-specific human judgements in general. However, the correlation is very poor in the case of verbs, as both human evaluators preferred Systran over Moses, while diagnostic evaluation (even with gold standard alignments) ranked Moses over Systran for this checkpoint. We manually inspected the outputs of Moses and Systran for those sentences for which diagnostic evaluation contradicted human evaluation for the verb checkpoint. We found that in some of the cases the problem was due to the failure of DELiC4MT to consider synonyms. Most of the existing automatic evaluation metrics (except METEOR) also suffer from this problem. Availability of mutiple reference translations can circumvent this problem and DELiC4MT also supports evaluation with mutiple references. It should be also noted that the scoring of DELiC4MT being n-gram based, the metric might be slightly biased toward SMT systems (Callison-Burch et al, 2006).

The checkpoint-specific inter-annotator agreements (Fleiss' Kappa) between the two annotators were 0.32 (adjectives), 0.13 (adverbs), 0.12 (determiners), 0.24 (nouns) and 0.29 (verbs). This somewhat low agreement may be due to the fact that although the evaluators are experienced in translation evaluation in terms of adequacy and fluency, they never performed diagnostic evaluation of this sort. It can be noticed from Tables 6 and 7 that evaluator 2 consistently gives higher scores for adequacy and fluency than evaluator 1 across systems; but these scores still correlate perfectly (cf. Table 9). A limitation of the current study regards the low number of human annotators, as having more of them might probably result in more stable results.

Finally, Table 9 presents the Pearson's correlation coefficients between the system-level scores across systems. As it can be seen from this table, system-level diagnostic evaluation scores obtained on automatically derived word alignments correlate very highly with those obtained on the gold standard alignment. In fact, diagnostic evaluation using grow-diag-final and union alignments (as opposed to using manual alignments) was found to correlate better with human judgements, while the use of intersection alignments produced better correlations with the majority of the automatic MT evaluation metrics. This indicates that using automatic word alignments is sufficient for carrying out diagnostic evaluation. Diagnostic evaluation correlates well with all automatic evaluation scores (including TER, which being an error metric shows strong negative or inverse association) as well as human judgements, indicating that this type of evaluation is accurate at predicting true system quality.

Pearson's Correlation	Noun	Verb	Adv	Adj	Det
Evaluator 1 – Evaluator 2	0.880	0.959	0.962	0.987	0.327
Evaluator 1 – Diagnostic (manual)	0.999	-0.305	0.951	0.872	0.885
Evaluator 2 – Diagnostic (manual)	0.898	-0.021	0.830	0.940	0.729
Evaluator 1 – Diagnostic (gdf)	0.999	-0.123	0.890	0.710	0.873
Evaluator 2 – Diagnostic (gdf)	0.853	0.163	0.980	0.815	0.746
Diagnostic (manual) – Diagnostic (gdf)	0.996	0.983	0.704	0.964	1.000
Diagnostic (manual) – Diagnostic (union)	0.999	0.968	0.704	0.984	1.000
Diagnostic (manual) – Diagnostic (intersection)	0.998	0.932	0.996	0.999	0.999

Table 8: Pearson's correlation for checkpoint-specific evaluation across systems

		Diagnostic				Human	
		manual	gdf	union	intersection	evaluator 1	evaluator 2
Diagnostic	manual	1.000	1.000	1.000	0.988	0.975	0.972
	gdf	1.000	1.000	1.000	0.987	0.976	0.973
	union	1.000	1.000	1.000	0.983	0.980	0.978
	intersection	0.988	0.987	0.983	1.000	0.927	0.922
Automatic	BLEU	0.972	0.971	0.966	0.997	0.895	0.890
	NIST	0.995	0.994	0.992	0.998	0.947	0.942
	METEOR	0.992	0.992	0.989	0.999	0.940	0.935
	TER	-0.986	-0.986	-0.990	-0.947	-0.998	-0.998
Human	Evaluator 1	0.975	0.976	0.980	0.927	1.000	1.000
	Evaluator 2	0.972	0.973	0.978	0.922	1.000	1.000

Table 9: Pearson's correlation between the system level scores

6 Conclusions and Future Work

This paper has evaluated a diagnostic metric that assesses the performance of MT systems on user-defined linguistic phenomena. This has been done by means of a case study for the English–French language direction.

As this metric is dependent on word alignments, one of the objectives was to find the margin of error in diagnostic evaluation using automatic word alignments as opposed to using gold standard manual alignments. In order to determine that, we carried out diagnostic evaluation using manual alignments as well as a set of commonly used automatic alignments (grow-diag-final, union and intersection). In addition, we also calculated the correlation with several state-of-the-art automatic MT evaluation metrics as well as with human judgements.

From the experimental results we found that automatically-derived word alignments can be considered as effective as gold standard alignments when carrying out diagnostic evaluation. We also observed that diagnostic evaluation can accurately capture translation quality and, in general, correlates well both with system-level automatic evaluation metrics and with human judgements.

As an extension to this work, we would like to explore the impact of different automatic aligners on the results of diagnostic evaluation of MT. Also, the low correlation with human judgements obtained for verbs requires a deeper analysis of this linguistic phenomenon and how it is treated by the diagnostic metric, which we plan to explore in further detail.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreements FP7-ICT-4-248531 (CoSyne) and PIAP-GA-2012-324414 (Abu-MaTran) and through Science Foundation Ireland as part of the CNGL (grant 07/CE/I1142).

References

Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluation the Role of Bleu in Machine Translation Research. In *Proceedings of EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics*.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145.

Fishel, M., Bojar, O., and Popović, M. (2012). Terra: a collection of translation error-annotated corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* , pages 7–14.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86.

LDC (2005). Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report. Revision 1.5.

Max, A., Crego, J. M., and Yvon, F. (2010). Contrastive lexical evaluation of machine translation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Melamed, I. D. (1998). Manual annotation of translational equivalence: The blinker project. *CoRR*, cmp-lg/9805005.

Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10.

Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2* , pages 1086–1090.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* , pages 311–318.

Popović, M. (2011). Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Popović, M. and Burchardt, A. (2011). From human to automatic error classification for machine translation output. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*.

Popović, M., Ney, H., de Gispert, A., Mariño, J. B., Gupta, D., Federico, M., Lambert, P., and Banchs, R. (2006). Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 1–6.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report, Department of Computer and Information Science, University of Pennsylvania. (3rd revision, 2nd printing).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Stolcke, A. (2002). SRILM–an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286.

Toral, A., Naskar, S. K., Gaspari, F., and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. In *The Prague Bulletin of Mathematical Linguistics*, 98:121–132.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D., and Zhao, T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1.* , pages 1121–1128.