A Diagnostic Evaluation Approach Targeting MT Systems for Indian Languages

Renu Balyan[#1], *Sudip Kumar Naskar*[†], *Antonio Toral*[†], *Niladri Chatterjee*[‡]
(#) Centre for Development of Advanced Computing, Noida, India
(†) CNGL, School of Computing, Dublin City University, Dublin, Ireland
(‡) Department of Mathematics, Indian Institute of Technology, Delhi, India
renubalyan@cdac.in, {snaskar,atoral}@computing.dcu.ie,
niladri.iitd@gmail.com

ABSTRACT

This paper addresses diagnostic evaluation of machine translation (MT) systems for Indian languages, English to Hindi translation in particular. Evaluation of MT output is an important but difficult task. The difficulty arises primarily from some inherent characteristics of the language pairs, which range from simple word-level discrepancies to more difficult structural variations for Hindi from English, such as reduplication of words, free word order etc. The proposed scheme is based on identification of linguistic units (often referred to as checkpoints). We use the diagnostic evaluation tool DELiC4MT to analyze the contribution of various PoS classes for different categories. We further suggest some additional checkpoints based on named entities, ambiguous words, word order and inflections that are relevant for the evaluation of Hindi. The evaluation of these checkpoints provides a detailed analysis and helps in monitoring how an MT system handles these linguistic phenomena as well. This also provides valuable feedback to MT developers as to where the system is performing poorly and how the output can possibly be improved. The effectiveness of the approach was tested on 5 English to Hindi MT systems and it was observed that the system-level DELiC4MT scores correlate well with the scores produced by the most commonly used automatic evaluation metrics (BLEU, NIST, METEOR and TER) while providing finer-grained information.

KEYWORDS: diagnostic evaluation, automatic evaluation metrics, linguistic checkpoints

1 Work done while at CNGL, School of Computing, DCU

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 61–72,
COLING 2012, Mumbai, December 2012.

1 Introduction

Evaluation of MT systems has received a lot of attention over the last decade or so, yet no generally ideal automatic metric could be designed so far. The problem becomes even more pronounced when the source and target languages are distant, (e.g. they belong to different language families). The MT community is very much in need of a suitable evaluation methodology for evaluating translation quality. This is particularly true with respect to Indian languages. In the last 15 years or so, MT into Indian languages (especially Hindi) has gained tremendous research interest in India and elsewhere. Many English to Hindi and Indian Languages to Indian Languages MT systems have been designed, for example AnglaBharati (Sinha et al., 1995), Anusaaraka[2] (Chaudhury et al., 2010), Anuvadaksh[3], Google[4], Sampark[5], MaTra[6] (Ananthakrishnan et al., 2006), to name just a few.

However, the issue of evaluating the output of these MT systems has remained rather unexplored. The state-of-the-art methods for automatic MT evaluation are represented by BLEU (Papineni et al., 2002) and closely related NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) and TER (Snover et al., 2006). These metrics have been widely accepted as benchmarks for MT system evaluation. However, the research community is also aware of the deficiencies of these metrics (Callison-Burch et al., 2006). Globally, these automatic MT evaluation metrics (BLEU, NIST, TER, METEOR, etc.) are being studied with great interest for different language pairs. But their direct applicability to Hindi, or other Indian languages for that matter, needs proper investigation. Indian languages are characteristically different from English and other related European languages for which these metrics are mostly used.

There have been some efforts in this direction for Indian languages (Chatterjee and Balyan, 2011; Gupta et al., 2010; Ananthakrishnan et al., 2007; Chatterjee et al., 2007; Moona et al., 2004). Barring these few exceptions, the subject has not been studied deeply. Most of these approaches, however, either cover human evaluation, or consider modification of existing automatic metrics (like BLEU and METEOR) to make them more suitable for Indian languages. None of these works has been targeted towards diagnostic evaluation (Zhou et al., 2008; Naskar et al., 2011; Popović, 2011), which not only provides quantitative analysis, but also qualitative feedback of the machine translated text. It also provides feedback and detailed analysis of how an MT system performs for different linguistic features like verbs, nouns, compounds etc.

Our final aim is to come up with an approach for diagnostic evaluation of MT that can be adapted to Indian languages. In the present work the experiments have been carried out with the DELiC4MT (Toral et al., 2012) toolkit as it is language independent. The experiments have been carried out to adapt the tool for Hindi, which can be later extended to evaluation of other Indian languages as well. To the best of our knowledge

2 http://anusaaraka.iiit.ac.in/

3 http://tdil-dc.in

4 http://translate.google.com/

5 http:// sampark.iiit.ac.in/sampark/web/index.php/content

6 http://www.cdacmumbai.in/matra/

this is a pioneering work in the direction of diagnostic evaluation with respect to Indian languages.

The rest of the paper is organized as follows. Related work on diagnostic evaluation is discussed in Section 2. Section 3 gives a brief overview of the diagnostic evaluation tool, DELiC4MT, which has been used for this study. In Section 4, the various linguistic checkpoints considered for the study of English and Hindi have been discussed. Section 5 discusses the experimental setup and compares the results obtained on the English-Hindi test set using DELiC4MT and automatic evaluation metrics. This is followed by conclusions and avenues for future work.

2 Related work

Although diagnostic evaluation of MT has been occasionally addressed in the literature in the last few years, no widely accepted solution seems to have emerged till date. A framework proposed by Vilar et al. (2006) analyzes the errors manually. The scheme covers five top-level classes: missing words, incorrect words, unknown words, word order and punctuation errors. Farrús et al. (2010) classified errors at orthographic, morphological, lexical, semantic, and syntactic level. Some automatic methods for error analysis using base forms and PoS tags have been proposed in (Popović et al., 2006; Popović and Ney, 2011). The proposed methods have been used for estimation of inflectional and reordering errors. Popović and Burchardt (2011) present a method for automatic error classification. Popović (2011) describes a tool that classifies errors into five categories based on the hierarchy proposed by Vilar et al. (2006). Popović (2012) describes RGBF, a tool for automatic evaluation of MT output based on n-gram precision and recall. Fishel et al. (2012) quantifies translation quality based on the frequencies of different error categories. Xiong et al. (2010) used a classifier trained with a set of linguistic features to automatically detect incorrect segments in MT output.

EAGLES (1996) distinguishes a type of evaluation whose purpose is to discover the reason(s) why a system did not produce the results it was expected to. Working on these lines Zhou et al. (2008) proposed diagnostic evaluation of linguistic checkpoints. Naskar et al. (2011) proposed a framework for diagnostic MT evaluation which offers similar functionality as proposed in (Zhou et al., 2008) but is language independent.

3 DELiC4MT: A Diagnostic MT Evaluation Tool

DELiC4MT[7] (Diagnostic Evaluation using Linguistic Checkpoints for Machine Translation) is an open source tool for diagnostic evaluation of MT. It allows evaluation of MT systems over linguistic features. The various steps involved for diagnostic evaluation using DELiC4MT are: text analysis and KAF conversion, word alignment extraction, defining kybots and evaluation. The tool makes extensive use of already available NLP tools and representation standards. The evaluation pipeline proceeds as follows.

7 http://www.computing.dcu.ie/~atoral/delic4mt(under the GPL-v3 license).

- The source and target sides of the gold standard (test set) are processed by respective PoS taggers (Treetagger[8] for English and a shallow parser for Hindi) and converted into KYOTO Annotation Format (KAF) (Bosma et al., 2009) to represent textual analysis.
- The test set is word aligned using GIZA++ [9](Och and Ney, 2003), and identifiers of the aligned tokens are stored.
- Kybot[10] (Vossen et al., 2010) profiles specifying the linguistic checkpoints to be extracted are run on the KAF text and the matching terms are extracted.
- The evaluation module takes kybot output, KAF text, word alignments and the output of an MT system (plain text, no word alignment is performed on it) as inputs. It calculates the performance of the MT system over the linguistic checkpoint(s) considered.

The details of the tool regarding KAF files and kybot profiles can be found in Toral et al. (2012).

4 Linguistic Checkpoints

A linguistic checkpoint is a linguistically-motivated unit e.g., it can be an ambiguous word, a verb-particle construction, a noun-noun compound, a PoS n-gram etc. The level of detail and the specific linguistic phenomena included in the taxonomy can vary depending on what the users want to investigate as part of the diagnostic evaluation. However, the taxonomy of automatic diagnostic evaluation should be widely accepted. The categories that are out of scope for current NLP tools to recognize have been ignored in this study. In light of the above consideration, we adopted the taxonomy introduced by Lata et al. (2012), Baskaran et al. (2008) and the IIIT Tagset[11] (Bharati et al., 2006) for Hindi. The taxonomy includes typical checkpoints at word level. Some examples of the representative checkpoints at different levels for English and Hindi languages have been presented in the following subsection.

4.1 English to Hindi Checkpoints

The implementation of the English to Hindi checkpoint taxonomy can take into account various checkpoints at word and phrase level. However, only 8 word level categories have been considered for this study. The taxonomy is shown in Table 1. In practice, any tag used by parsers (e.g. NP, VP, PP, etc.) can be added as a new category easily; though currently these have not been implemented in the system. The system currently works for word level PoS-based checkpoints only. However we also propose to use phrase-level and other checkpoints related to named entities (NE) and ambiguous words for English, which are currently not implemented in the system.

8 http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/

9 http://code.google.com/p/giza-pp/

10 http://kyoto.let.vu.nl/svn/kyoto/trunk/modules/mining_module/

11 http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

The NE checkpoint is important as we found that the existing English to Hindi MT systems do not handle NEs properly. Typically, they provide literal translations of the words, leading to poor translation quality.

Checkpoint	Level	Category
PoS based	Word	Noun, Verb, Modal, Pronoun, Adverb, Possessive Pronoun, Adjective, Preposition
	Phrase	Noun Phrases, Prepositional Phrases, Verb Phrases, Noun Compounds, Verb Particle Constructions
Named Entity		
Ambiguous word		

TABLE 1 – Linguistic checkpoints for English to Hindi translation

4.2 Hindi to English Checkpoints

Using only PoS-based linguistic checkpoints might not be as helpful in evaluating the translation quality as compared to using checkpoints that deal with inflections, word order etc. For Hindi, the proposed linguistic checkpoints belong to the following categories: PoS-based, inflectional, NE, ambiguous word, word order, re-duplicated words and extra, missing or incorrect postpositions. The measures suggested by (Popović and Ney, 2011) could be used for determining the inflectional errors for nouns, adjectives and verbs and the word order problems. The approach suggested by (Popović and Ney, 2011) for missing, extra or incorrect words could be applied for postposition related problems.

5 Evaluation

5.1 Experimental setup

The test set considered for this study consists of 1,000 sentences from the tourism domain. DELiC4MT has been used for diagnostic evaluation of five English to Hindi MT systems: Google Translate (MT1), Bing Translator[12] (MT2), Free-translations[13] (MT3), MaTra2 (MT4) and Anusaaraka (MT5). GIZA++ was used for word alignment. Since the test set is very small, an additional parallel corpus comprising of 25,000 sentences from the same domain was used to avoid data sparseness during word alignment. The test set was appended to the additional corpus and the word alignments were generated. Finally the word alignments for the test set sentences were extracted. Treetagger was used to PoS-tag the English dataset, while the Hindi dataset was PoS-tagged using the PoS tagger developed by IIIT, Hyderabad[14]. For linguistic checkpoints we have considered linguistic units at word level only. Simple PoS-based checkpoints (noun,

12 http://www.bing.com/translator/

13 http://www.free-translator.com/

14 http://sivareddy.in/downloads

verb, adjective, etc.) have been considered at the word level. The experimental results are discussed in 5.2.

5.2 Results and Discussion

The checkpoint-specific diagnostic evaluation scores for word level checkpoints across the MT systems using DELiC4MT are shown in Table 2. In addition to the diagnostic evaluation scores the table also shows the number of instances obtained for the checkpoints. Checkpoint-specific best scores are shown in bold. Checkpoint-specific statistically significant improvements are also reported in Table 2 and these are shown as superscripts. For representation purposes, we use *a, b, c, d* and *e* for MT1, MT2, MT3, MT4 and MT5 respectively. For example, the MT2 score $0.3568^{d,e}$ for noun checkpoint in Table 2 indicates that the improvement provided by MT2 for this checkpoint is statistically significant over MT4 (d) and MT5 (e).

Checkpoint	Instances	MT1(a)	MT2(b)	MT3(c)	MT4(d)	MT5(e)
Noun	4538	**0.3792**^{b,d,e}	0.3568^{d,e}	0.3776^{b,d,e}	0.2552	0.2925^{d}
Pronoun	276	**0.5539**^{d}	0.5000	**0.5539**^{d}	0.4059	0.5490^{d}
Possessive Pronoun	184	**0.3464**^{d,e}	0.3333^{d,e}	**0.3464**^{d,e}	0.0196	0.1699^{d}
Adjective	1859	**0.3785**^{b,d,e}	0.3574^{d,e}	0.3772^{b,d,e}	0.2061	0.2699^{d}
Adverb	663	**0.4347**^{d}	0.4288^{d}	0.4327^{d}	0.2402	0.4103^{d}
verb	2580	**0.2656**^{d,e}	0.2584^{d}	**0.2656**^{d,e}	0.1789	0.2402^{d}
Preposition	2667	**0.6655**^{d,e}	0.6555^{d,e}	0.6646^{d,e}	0.5434	0.6217^{d}
Modal	128	0.3913	0.3696	0.3913	0.3478	**0.4239**
Total / Average Scores	12895	**0.4269**	0.4075	0.4262	0.2746	0.3722
System Scores (Weighted)		**0.4218**	0.4055	0.4208	0.2925	0.3579

TABLE 2 – DELiC4MT scores for word-level checkpoints for MT systems

The following observations were made for evaluation of word-level checkpoints:

- MT1 outperforms all the MT systems in all the categories except modals.
- MT3 performs almost at par with MT1.
- Among the word-level checkpoints verbs seem to be the most problematic checkpoint for all the systems except for MT4 and MT5 which perform the worst for possessive pronouns category.
- MT5 performs best for the modals category in comparison to the rest of systems.

- All the systems perform poorly on possessive pronouns compared to pronouns in general.
- All the systems perform best on prepositions followed by the pronouns category.
- MT1, MT2 and MT3 systems perform better for adverbs as compared to modals, whereas MT4 and MT5 perform just in the reverse manner for these categories.

The performance of all the MT systems was also evaluated using automatic evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006). The automatic evaluation metric scores for all the systems are shown in Table 3.

	MT1	MT2	MT3	MT4	MT5
BLEU	**9.72**	7.41	9.69	2.17	4.67
NIST	**4.15**	3.83	4.14	2.14	2.98
TER	**81.15**	83.07	81.30	88.81	88.96
METEOR	**0.274**	0.257	0.273	0.165	0.206

TABLE 3 – Automatic Evaluation Metric scores for MT systems

According to all the automatic evaluation metrics MT1 performs best followed by MT3, MT2, MT5 and MT4 (the only exception being MT4 ranked higher than MT5 by TER) as is also found by DELiC4MT. However, the point to be noted here is that with automatic evaluation metrics we do not get any additional information about the systems' performance other than the system-level scores but DELiC4MT does provide that information.

Table 4 shows the Pearson correlation coefficients between the scores obtained from DELiC4MT and the automatic evaluation metrics. It can be seen from table 4 that DELiC4MT scores have high correlation with the automatic measures. DELiC4MT scores have the highest correlation with NIST followed by METEOR, BLEU and TER. This entails that, in addition to evaluating on linguistic checkpoints, DELiC4MT can also measure performance of MT systems at system-level. It provides system-level scores for all the MT systems in accordance with other automatic evaluation metrics.

	BLEU	NIST	TER	METEOR	DELiC4MT
BLEU	1.000	.988**	-.954*	.989**	.974**
NIST		1.000	-.935*	.999**	.996**
TER			1.000	-.948*	-.901*
METEOR				1.000	.992**
DELiC4MT					1.000
** Correlation is significant at the 0.01 level (2-tailed).					
* Correlation is significant at the 0.05 level (2-tailed).					

TABLE 4 – Pearson correlation coefficients between DELiC4MT scores and automatic evaluation metrics

Concluding Remarks and Future Work

The paper presents a study on diagnostic evaluation of MT for Indian languages. The main objective of the work was to assess the applicability of the diagnostic evaluation tool DELiC4MT, for Indian languages in general, and Hindi in particular. The linguistic checkpoints considered for this study were PoS-based (word level only). In total 8 word level checkpoints were considered for the study. The paper has presented a detailed analysis of the results obtained for 5 English to Hindi MT systems using DELiC4MT. The translations obtained from these MT systems were also evaluated using some of the most commonly used automatic evaluation metrics. As far as the MT systems are concerned, Google proved to be the best among the 5 systems according to both automatic evaluation metrics and diagnostic evaluation metrics. It was also observed that the system-level DELiC4MT scores correlate well with all other automatic evaluation metric scores with Pearson correlation coefficients above 0.9 for all cases. We have also proposed the use of additional phrase level checkpoints and also checkpoints that can provide feedback related to NEs and ambiguous words.

This work is just a first step towards the development of evaluation measures for Indian languages based on linguistic units which provide feedback on specific translation problems. The work offers a number of possibilities for future work, both for improving the existing measures by adding more sophisticated and meaningful linguistic checkpoints and also exploring the use of other existing toolkits for handling translation errors based on inflections, word order etc. The authors plan to carry out further work in this direction.

Acknowledgements

This work has been funded in part by the European Commission through the CoSyne project (FP7-ICT-4-248531) and Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. The authors would also like to thank Vineet Chaitanya, IIIT-Hyderabad, Sasi Kumar, CDAC and Siva Reddy for providing support.

References

Ananthakrishnan, R., Bhattacharyya, P., Sasikumar, M. and Shah, R. (2007). Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU. *Proceeding of 5th International Conference on Natural Language Processing (ICON-07)*. Hyderabad, India.

Ananthakrishnan, R., Kavitha, M., Hegde, J., Shekhar, C., Shah, R., Bade, S. and Sasikumar, M. (2006). MaTra: A practical approach to fully-automatic indicative English-Hindi machine translation. *Symposium on Modeling and Shallow Parsing of Indian Languages(MSPIL'06)*. IIT Bombay, India.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.* Ann Arbor, Michigan. 65-72.

Baskaran, S., Bali, K., Choudhury, M., Bhattacharya, T., Bhattacharyya, P., Jha, G. N., Rajendran, S., Saravanan, K., Sobha, L. and Subbarao, K. V. (2008). A Common Parts-of-Speech Tagset Framework for Indian Languages. *Proceedings of LREC 2008*. Marrakech, Morocco. 1331-1337.

Bharati, A., Sharma, D. M., Bai, L. and Sangal, R. (2006). AnnCorra : *Annotating Corpora Guidelines for PoS and Chunk Annotation for Indian Languages*. LTRC - Technical Report-31.

Bosma, W. E., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M. and Aliprandi, C. (2009). Kaf: a generic semantic annotation format. *Proceedings of the GL2009 Workshop on Semantic Annotation.*

Callison-Burch, C., Osborne, M. and Koehn, P. (2006). Re-evaluating the Role of Bleu in Machine Translation Research. *Proceedings of the European Chapter of the ACL 2006.*

Chatterjee, N. and Balyan, R. (2011). Towards Development of a Suitable Evaluation Metric for English to Hindi Machine Translation. *International Journal of Translation*, Vol. 23, No. 1, Jan-Jun 2011. 07-26.

Chatterjee, N., Johnson, A. and Krishna, M. (2007). Some improvements over the BLEU metric for measuring the translation quality for Hindi. *Proceedings of the International Conference on Computing: Theory and Applications - ICCTA'07*. Kolkata, India. 485- 490.

Chaudhury, S., Rao, A. and Sharma, D. M. (2010). Anusaaraka: An Expert system based MT System. *Proceedings of IEEE conference on Natural language processing and knowledge management* (IEEE-NLPKE 2010). Beijing, China.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proceedings of the Human Language Technology Conference (HLT)*. San Diego, CA. 128–132.

Farr´us, M., Costa-juss`a, M. R., Mari¨no, J. B. and Fonollosa, J. A. R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. *Proceedings of EAMT*. Saint Rapha¨el, France. 52–57.

Fishel, M., Sennrich, R., Popović, M. and Bojar, O. (2012). TerrorCat: a Translation Error Categorization-based MT Quality Metric. *Proceedings of the Seventh Workshop on Statistical Machine Translation.* Montréal, Canada. 64-70

Gupta, A., Venkatapathy, S. and Sangal, R. (2010). METEOR-Hindi : Automatic MT Evaluation Metric for Hindi as a Target Language. *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, Macmillan Publishers, India.

Lata, S., Chandra, S., Verma, P. and Arora, S. (2012). Standardization of POS Tag Set for Indian Languages based on XML Internationalization best practices guidelines. *Proceedings of LREC-(WILDRE) First Workshop on Indian Language Data: Resources and Evaluation*. Istanbul, Turkey. 01-17.

Lavie, A. and Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second ACL Workshop on Statistical Machine Translation*. Prague, Czech Republic. 228-231.

Moona, R. S., Sangal, R. and Sharma, D. M. (2004). MTeval: A Evaluation methodolgy for Machine Translation system. *Proceedings of SIMPLE'04*. Kharagpur, India. 15-19.

Naskar, S. K., Toral, A., Gaspari, F. and Ways, A. (2011). A framework for Diagnostic Evaluation of MT based on Linguistic Checkpoints. *Proceedings of the 13th Machine Translation Summit.* Xiamen,China. 529-536.

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. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of 40th Annual Meeting of the ACL*. Philadelphia, PA, USA. 311-318.

Popović, M. (2012). rgbF: An Open Source Tool for n-gram Based Automatic Evaluation of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*. 98: 99–108. doi: 10.2478/v10108-012-0012-y.

Popović, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*. Volume 37 Issue 4, (pp. 657-688). MIT Press Cambridge, MA, USA.

Popović, M. and A. Burchardt. (2011). From human to automatic error classification for machine translation output. *Proceedings of EAMT 2011*. Leuven, Belgium. 265-272.

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., Ney, H., Gispert, A. D., 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. *StatMT '06: Proceedings of the Workshop on Statistical Machine Translation.* New York. 1-6.

Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R. and Jain, A. (1995). AnglaBharti: A multilingual machine aided translation project on translation from

English to Hindi. *Proceedings of IEEE International Conference Systems, Man and Cybernetics.* IEEE Press, Vancouver,British Columbia, Canada. 1609-1614.

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

The EAGLES MT Evaluation Working Group. (1996). *EAGLES Evaluation of Natural Language Processing Systems.* Final Report. EAGLES Document EAG-EWG-PR.2, ISBN 87-90708-00-8. Center for Sprogteknologi, Copenhagen.

Toral, A., Naskar, S. K., Gaspari, F. and Groves, D. (2012). DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics.* 98:121–131. doi: 10.2478/v10108-012-0014-9.

Vilar, D., Xu, J., Fernando L. D'Haro, and Ney, H. (2006). Error analysis of statistical machine translation output. *Proceedings of Fifth International Conference on Language Resources and Evaluation* (LREC-2006). Genoa, Italy. 697-702.

Vossen, P., Rigau, G., Agirre, E., Soroa, A., Monachini, M. and Bartolini, R. (2010). Kyoto: an open platform for mining facts. *Proceedings of the 6th Workshop on Ontologies and Lexical Resources*. Beijing, China. 01–10.

Xiong, D., M. Zhang, and Li, H. (2010). Error detection for statistical machine translation using linguistic features. *Proceedings of ACL 2010*. Uppsala, Sweden. 604-611.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang, D. and Zhao, T. (2008). Diagnostic Evaluation of Machine Translation Systems using Automatically Constructed Linguistic Checkpoints. *Proceedings of 22nd International Conference on Computational Linguistics* (COLING 2008). Manchester. 1121-1128.