A Web Application for the Diagnostic Evaluation of Machine Translation over Specific Linguistic Phenomena

Antonio Toral Sudip Kumar Naskar Joris Vreeke Federico Gaspari Declan Groves
School of Computing
Dublin City University
Ireland
{atoral, snaskar, fgaspari, dgroves}@computing.dcu.ie joris.vreeke@dcu.ie

Abstract

This paper presents a web application and a web service for the diagnostic evaluation of Machine Translation (MT). These web-based tools are built on top of DELiC4MT, an open-source software package that assesses the performance of MT systems over user-defined linguistic phenomena (lexical, morphological, syntactic and semantic). The advantage of the web-based scenario is clear; compared to the standalone tool, the user does not need to carry out any installation, configuration or maintenance of the tool.

1 Automatic Evaluation of Machine Translation beyond Overall Scores

Machine translation (MT) output can be evaluated using different approaches, which can essentially be divided into human and automatic, both of which, however, present a number of shortcomings. Human evaluation tends to be more reliable in a number of ways and can be tailored to a variety of situations, but is rather expensive (both in terms of resources and time) and is difficult to replicate. On the other hand, standard automatic MT evaluation metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) are considerably cheaper and provide faster results, but return rather crude scores that are difficult to interpret for MT users and developers alike. Crucially, current standard automatic MT evaluation metrics also lack any diagnostic value, i.e. they cannot identify specific weaknesses in the MT output. Diagnostic information can be extremely valuable for MT developers and users, e.g. to improve the performance of the system or to decide which output is more suited for particular scenarios.

An interesting alternative to the traditional MT evaluation metrics is to evaluate the performance of MT systems over specific linguistic phenomena. While retaining the main advantage of automatic metrics (low cost), this approach provides more fine-grained linguistically-motivated evaluation. The linguistic phenomena, also referred to as linguistic checkpoints, can be defined in terms of linguistic information at different levels (lexical, morphological, syntactic, semantic, etc.) that appear in the source language. Examples of such linguistic checkpoints, what translation information they can represent, and their relevance for MT are provided in Table 1.

Checkpoint	Relevance for MT
Lexical	Words that can have multiple translations in the target. For example, the preposition "de" in Spanish can be translated into English as "of" or "from" depending on the context.
Syntactic	Syntactic constructs that are difficult to translate. E.g., a checkpoint containing the sequence a noun (*noun1*) followed by the preposition "de", followed by another noun (*noun2*) when translating from Spanish to English. The equivalent English construct would be *noun2*'s *noun1*, the translation thus involving some reordering.
Semantic	Words with multiple meanings, which possibly correspond to different translations in the target language. Polysemous words can be collected from electronic dictionaries such as WordNet (Miller, 1995).

Table 1: Linguistic Checkpoints

Checkpoints can also be built by combining el-

ements from different categories. For example, by combining lexical and syntantic elements, we could define a checkpoint for prepositional phrases (syntactic element) which start with the preposition "de" (lexical element).

Woodpecker (Zhou et al., 2008) is a tool that performs diagnostic evaluation of MT systems over linguistic checkpoints for English–Chinese. Probably due to its limitation to one language pair, its proprietary nature as well as rather restrictive licensing conditions, Woodpecker does not seem to have been widely used in the community, in spite of its ability to support diagnostic evaluation.

DELiC4MT[1] is an open-source software that follows the same approach as Woodpecker. However, DELiC4MT is easily portable to any language pair[2] and provides additional functionality such as filtering of noisy checkpoint instances and support for statistical significance tests. This paper focuses on the usage of this tool through a web application and a web service from the user's perspective. Details regarding its implementation, evaluation, etc. can be found in (Toral et al., 2012; Naskar et al., 2011).

2 Web Services for Language Technology Tools

There exist many freely available language processing tools, some of which are distributed under open-source licenses. In order to use these tools, they need to be downloaded, installed, configured and maintained, which results in high cost both in terms of manual effort and computing resources. The requirement for in-depth technical knowledge severely limits the usability of these tools amongst non-technical users, particularly in our case amongst translators and post-editors.

Web services introduce a new paradigm in the way we use software tools where only providers of the tools are required to have knowledge regarding their installation, configuration and maintenance. This enables wider adoption of the tools and reduces the learning curve for users as the only information needed is basic knowledge of the functionality and input/output parameters (which can be easily included, e.g. as part of an online tutorial). While this paradigm is rather new in the field of computational linguistics, it is quite mature and successful in other fields such as bioinformatics (Oinn et al., 2004; Labarga et al., 2007).

Related work includes two web applications in the area of MT evaluation. iBLEU (Madnani, 2011) organises BLEU scoring information in a visual manner. Berka et al. (2012) perform automatic error detection and classification of MT output.

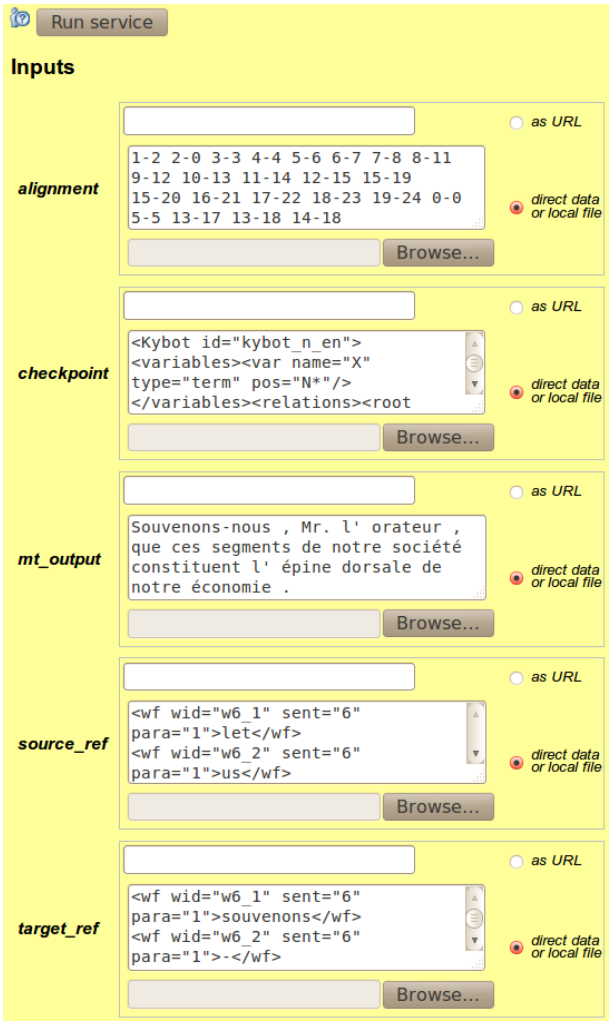

Figure 1: Web interface for the web service.

3 Demo

The demo presented in this paper consists of a web service and a web application built on top of DELiC4MT that allow to assess the performance of MT systems on different linguistic phenomena de-

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

[2] It has already been tested on language pairs involving the following languages: Arabic, Bulgarian, Dutch, English, French, German, Hindi, Italian, Turkish and Welsh.

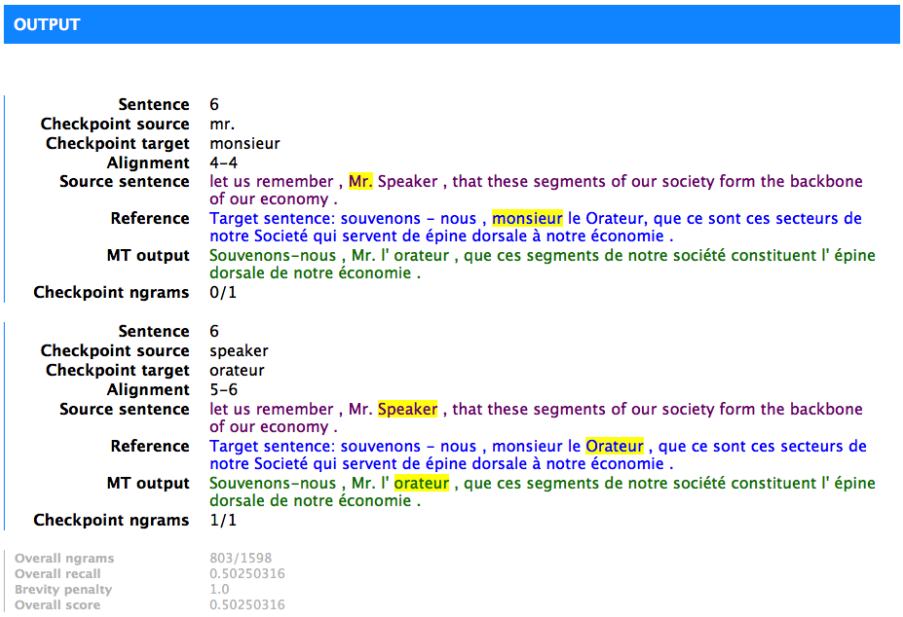

Figure 2: Screenshot of the web application (visualisation of results).

fined by the user. The following subsections detail both parts of the demo.

3.1 Web Service

A SOAP-compliant web service[3] has been built on top of DELiC4MT. It receives the following input parameters (see Figure 1):

1. Word alignment between the source and target sides of the testset, in the GIZA++ (Och and Ney, 2003) output format.
2. Linguistic checkpoint defined as a Kybot[4] (Vossen et al., 2010) profile.
3. Output of the MT system to be evaluated, in plain text, tokenised and one sentence per line.
4. Source and target sides of the testset (or gold standard), in KAF format (Bosma et al., 2009).[5]

The tool then evaluates the performance of the MT system (input parameter 3) on the linguistic phenomenon (parameter 2) by following this procedure:

- Occurrences of the linguistic phenomenon (parameter 2) are identified in the source side of the testset (parameter 4).
- The equivalent tokens of these occurrences in the target side (parameter 5) are found by using word alignment information (parameter 1).
- For each checkpoint instance, the tool checks how many of the n-grams present in the reference of the checkpoint instance are contained in the output produced by the MT system (parameter 3).

3.2 Web Application

The web application builds a graphical interface on top of the web service. It allows the user to visualise the results in a fine-grained manner, the user can see the performance of the MT system for each single occurrence of the linguistic phenomenon.

Sample MT output for the "noun" checkpoint for the English to French language direction is shown in Figure 2. Two occurrences of the checkpoint are shown. The first one regards the source noun "mr." and its translation in the reference "monsieur", identified through word alignments. The alignment (4-4) indicates that both the source and target tokens appear at the fifth position (0-based index) in the sentence. The reference token ("monsieur") is not found in the MT output and thus a score of $0/1$

[3] http://registry.elda.org/services/301

[4] Kybot profiles can be understood as regular expressions over KAF documents, http://kyoto.let.vu.nl/svn/kyoto/trunk/modules/mining_module/

[5] An XML format for text analysis based on representation standards from ISO TC37/SC4.

(0 n-gram matches out of a total of 1 possible n-gram) is assigned to the MT system for this noun instance. Conversely, the score for the second occurrence ("speaker") is $1/1$ since the MT output contains the 1-gram of the reference translation ("orateur").

The recall-based overall score is shown at the bottom of the figure (0.5025). This is calculated by summing up the scores (matching n-grams) for all the occurrences (803) and dividing the result by the total number of possible n-grams (1598).

4 Conclusions

In this paper we have presented a web application and a web service for the diagnostic evaluation of MT output over linguistic phenomena using DELiC4MT. The tool allows users and developers of MT systems to easily receive fine-grained feedback on the performance of their MT systems over linguistic checkpoints of their interest. The application is open-source, freely available and adaptable to any language pair.

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 and PIAP-GA-2012-324414 and through Science Foundation Ireland as part of the CNGL (grant 07/CE/I1142)

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the ACL-05 Workshop*, pages 65–72, University of Michigan, Ann Arbor, Michigan, USA.

Jan Berka, Ondej Bojar, Mark Fishel, Maja Popovi, and Daniel Zeman. 2012. Automatic MT Error Analysis: Hjerson Helping Addicter. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

W. E. Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*, September.

Alberto Labarga, Franck Valentin, Mikael Andersson, and Rodrigo Lopez. 2007. Web services at the european bioinformatics institute. *Nucleic Acids Research*, 35(Web-Server-Issue):6–11.

Nitin Madnani. 2011. iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 213–214, Washington, DC, USA. IEEE Computer Society.

George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, November.

Sudip Kumar Naskar, Antonio Toral, Federico Gaspari, and Andy Way. 2011. A Framework for Diagnostic Evaluation of MT based on Linguistic Checkpoints. In *Proceedings of the 13th Machine Translation Summit*, pages 529–536, Xiamen, China, September.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29:19–51, March.

Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. 2004. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, November.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antonio Toral, Sudip Kumar Naskar, Federico Gaspari, and Declan Groves. 2012. DELiC4MT: A Tool for Diagnostic MT Evaluation over User-defined Linguistic Phenomena. *The Prague Bulletin of Mathematical Linguistics*, pages 121–132.

Piek Vossen, German Rigau, Eneko Agirre, Aitor Soroa, Monica Monachini, and Roberto Bartolini. 2010. KYOTO: an open platform for mining facts. In *Proceedings of the 6th Workshop on Ontologies and Lexical Resources*, pages 1–10, Beijing, China.

Ming Zhou, Bo Wang, Shujie Liu, Mu Li, Dongdong Zhang, and Tiejun Zhao. 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*, COLING '08, pages 1121–1128, Stroudsburg, PA, USA. Association for Computational Linguistics.