A Graphical Interface for MT Evaluation and Error Analysis

Meritxell Gonzàlez and **Jesús Giménez** and **Lluís Màrquez**
TALP Research Center
Universitat Politècnica de Catalunya
{mgonzalez,jgimenez,lluism}@lsi.upc.edu

Abstract

Error analysis in machine translation is a necessary step in order to investigate the strengths and weaknesses of the MT systems under development and allow fair comparisons among them. This work presents an application that shows how a set of heterogeneous automatic metrics can be used to evaluate a test bed of automatic translations. To do so, we have set up an online graphical interface for the ASIYA toolkit, a rich repository of evaluation measures working at different linguistic levels. The current implementation of the interface shows constituency and dependency trees as well as shallow syntactic and semantic annotations, and word alignments. The intelligent visualization of the linguistic structures used by the metrics, as well as a set of navigational functionalities, may lead towards advanced methods for automatic error analysis.

1 Introduction

Evaluation methods are a key ingredient in the development cycle of machine translation (MT) systems. As illustrated in Figure 1, they are used to identify and analyze the system weak points (error analysis), to introduce new improvements and adjust the internal system parameters (system refinement), and to measure the system performance in comparison to other systems or previous versions of the same system (evaluation).

We focus here on the processes involved in the error analysis stage in which MT developers need to understand the output of their systems and to assess the improvements introduced.

Automatic detection and classification of the errors produced by MT systems is a challenging problem. The cause of such errors may depend not only on the translation paradigm adopted, but also on the language pairs, the availability of enough linguistic resources and the performance of the linguistic processors, among others. Several past research works studied and defined fine-grained typologies of translation errors according to various criteria (Vilar et al., 2006; Popović et al., 2006; Kirchhoff et al., 2007), which helped manual annotation and human analysis of the systems during the MT development cycle. Recently, the task has received increasing attention towards the automatic detection, classification and analysis of these errors, and new tools have been made available to the community. Examples of such tools are AMEANA (Kholy and Habash, 2011), which focuses on morphologically rich languages, and Hjerson (Popović, 2011), which addresses automatic error classification at lexical level.

In this work we present an online graphical interface to access ASIYA, an existing software designed to evaluate automatic translations using an heterogeneous set of metrics and meta-metrics. The primary goal of the online interface is to allow MT developers to upload their test beds, obtain a large set of metric scores and then, detect and analyze the errors of their systems using just their Internet browsers. Additionally, the graphical interface of the toolkit may help developers to better understand the strengths and weaknesses of the existing evaluation measures and to support the development of further improvements or even totally new evaluation metrics. This information can be gathered both from the experi-

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 139–144,
Jeju, Republic of Korea, 8-14 July 2012.

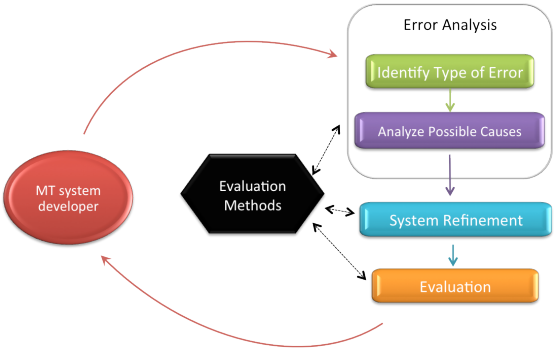

Figure 1: MT systems development cycle

ence of ASIYA's developers and also from the statistics given through the interface to the ASIYA's users.

In the following, Section 2 gives a general overview of the ASIYA toolkit. Section 3 describes the variety of information gathered during the evaluation process, and Section 4 provides details on the graphical interface developed to display this information. Finally, Section 5 overviews recent work related to MT error analysis, and Section 6 concludes and reports some ongoing and future work.

2 The ASIYA Toolkit

ASIYA is an open toolkit designed to assist developers of both MT systems and evaluation measures by offering a rich set of metrics and meta-metrics for assessing MT quality (Giménez and Màrquez, 2010a). Although automatic MT evaluation is still far from manual evaluation, it is indeed necessary to avoid the bottleneck introduced by a fully manual evaluation in the system development cycle. Recently, there has been empirical and theoretical justification that a combination of several metrics scoring different aspects of translation quality should correlate better with humans than just a single automatic metric (Amigó et al., 2011; Giménez and Màrquez, 2010b).

ASIYA offers more than 500 metric variants for MT evaluation, including the latest versions of the most popular measures. These metrics rely on different similarity principles (such as precision, recall and overlap) and operate at different linguistic layers (from lexical to syntactic and semantic). A general classification based on the similarity type is given below along with a brief summary of the information they use and the names of a few examples[1].

Lexical similarity: n-gram similarity and edit distance based on word forms (e.g., PER, TER, WER, BLEU, NIST, GTM, METEOR).

Syntactic similarity: based on part-of-speech tags, base phrase chunks, and dependency and constituency trees (e.g., SP-Overlap-POS, SP-Overlap-Chunk, DP-HWCM, CP-STM).

Semantic similarity: based on named entities, semantic roles and discourse representation (e.g., NE-Overlap, SR-Overlap, DRS-Overlap).

Such heterogeneous set of metrics allow the user to analyze diverse aspects of translation quality at *system*, *document* and *sentence* levels. As discussed in (Giménez and Màrquez, 2008), the widely used lexical-based measures should be considered carefully at sentence level, as they tend to penalize translations using different lexical selection. The combination with complex metrics, more focused on adequacy aspects of the translation (e.g., taking into account also semantic information), should help reducing this problem.

3 The Metric-dependent Information

ASIYA operates over a fixed set of translation test cases, i.e., a source text, a set of candidate translations and a set of manually produced reference translations. To run ASIYA the user must provide a test case and select the preferred set of metrics (it may depend on the evaluation purpose). Then, ASIYA outputs complete tables of score values for all the possible combination of metrics, systems, documents and segments. This kind of results is valuable for rapid evaluation and ranking of translations and systems. However, it is unfriendly for MT developers that need to manually analyze and compare specific aspects of their systems.

During the evaluation process, ASIYA generates a number of intermediate analysis containing partial work outs of the evaluation measures. These data constitute a priceless source for analysis purposes since a close examination of their content allows for analyzing the particular characteristics that

[1] A more detailed description of the metric set and its implementation can be found in (Giménez and Màrquez, 2010b).

Reference	The remote **control of** the Wii **helps** to diagnose **an infantile ocular** disease .	O_l score
Candidate 1	The Wii Remote to **help** diagnose **childhood eye** disease .	$\frac{7}{17} = 0.41$
Candidate 2	The control of the Wii helps to diagnose an ocular infantile disease .	$\frac{13}{14} = 0.93$

Table 1: The reference sentence, two candidate translation examples and the O_l scores calculation

differentiate the score values obtained by each candidate translation.

Next, we review the type of information used by each family of measures according to their classification, and how this information can be used for MT error analysis purposes.

Lexical information. There are several variants under this family. For instance, *lexical overlap* (O_l) is an F-measure based metric, which computes similarity roughly using the Jaccard coefficient. First, the sets of all lexical items that are found in the reference and the candidate sentences are considered. Then, O_l is computed as the cardinality of their intersection divided by the cardinality of their union. The example in Table 1 shows the counts used to calculate O_l between the reference and two candidate translations (boldface and underline indicate non-matched items in candidate 1 and 2, respectively). Similarly, metrics in another category measure the edit distance of a translation, i.e., the number of word insertions, deletions and substitutions that are needed to convert a candidate translation into a reference. From the algorithms used to calculate these metrics, these words can be identified in the set of sentences and marked for further processing. On another front, metrics as BLEU or NIST compute a weighted average of matching n-grams. An interesting information that can be obtained from these metrics are the weights assigned to each individual matching n-gram. Variations of all of these measures include looking at stems, synonyms and paraphrases, instead of the actual words in the sentences. This information can be obtained from the implementation of the metrics and presented to the user through the graphical interface.

Syntactic information. ASIYA considers three levels of syntactic information: shallow, constituent and dependency parsing. The shallow parsing annotations, that are obtained from the linguistic processors, consist of word level part-of-speech, lemmas and chunk Begin-Inside-Outside labels. Useful figures such as the matching rate of a given (sub)category of items are the base of a group of metrics (i.e., the ratio of prepositions between a reference and a candidate). In addition, dependency and constituency parse trees allow for capturing other aspects of the translations. For instance, DP-HCWM is a specific subset of the dependency measures that consists of retrieving and matching all the head-word chains (or the ones of a given length) from the dependency trees. Similarly, CP-STM, a subset of the constituency parsing family of measures, consists of computing the lexical overlap according to the phrase constituent of a given type. Then, for error analysis purposes, parse trees combine the grammatical relations and the grammatical categories of the words in the sentence and display the information they contain. Figure 2 and 3 show, respectively, several annotation levels of the sentences in the example and the constituency trees.

Semantic information. ASIYA distinguishes also three levels of semantic information: named entities, semantic roles and discourse representations. The former are post-processed similarly to the lexical annotations discussed above; and the semantic predicate-argument trees are post-processed and displayed in a similar manner to the syntactic trees. Instead, the purpose of the discourse representation analysis is to evaluate candidate translations at document level. In the nested discourse structures we could identify the lexical choices for each discourse sub-type. Presenting this information to the user remains as an important part of the future work.

4 The Graphical Interface

This section presents the web application that makes possible a graphical visualization and interactive access to ASIYA. The purpose of the interface is twofold. First, it has been designed to facilitate the use of the ASIYA toolkit for rapid evaluation of test beds. And second, we aim at aiding the analysis of the errors produced by the MT systems by creating

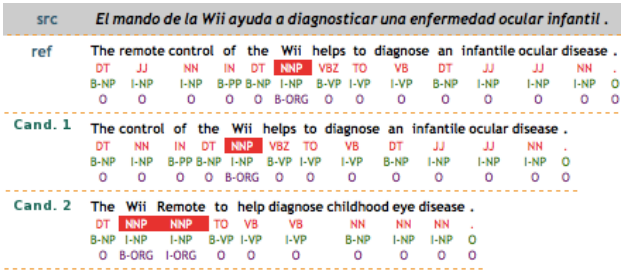

Figure 2: PoS, chunk and named entity annotations on the source, reference and two translation hypotheses

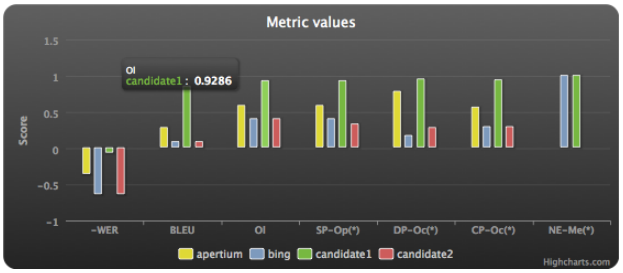

Figure 4: The bar charts plot to compare the metric scores for several systems

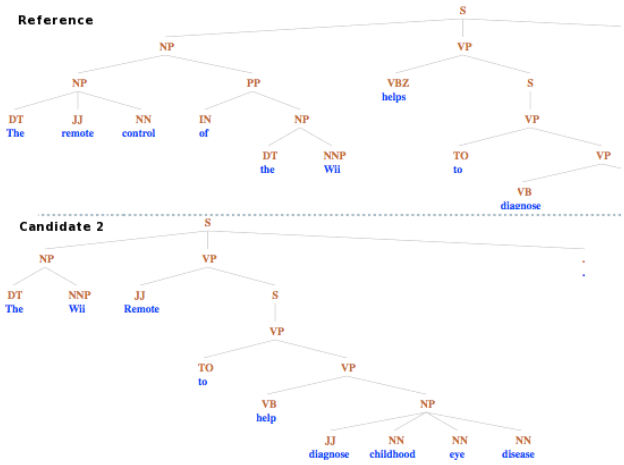

Figure 3: Constituency trees for the reference and second translation candidate

a significant visualization of the information related to the evaluation metrics.

The online interface consists of a simple web form to supply the data required to run ASIYA, and then, it offers several views that display the results in friendly and flexible ways such as interactive score tables, graphical parsing trees in SVG format and interactive sentences holding the linguistic annotations captured during the computation of the metrics, as described in Section 3.

4.1 Online MT evaluation

ASIYA allows to compute scores at three granularity levels: *system* (entire test corpus), *document* and *sentence* (or *segment*). The online application obtains the measures for all the metrics and levels and generates an interactive table of scores displaying the values for all the measures. Table organization can swap among the three levels of granularity, and it can also be transposed with respect to system and metric information (transposing rows and columns). When the metric basis table is shown, the user can select one or more metric columns in order to re-rank the rows accordingly. Moreover, the source, reference and candidate translation are displayed along with metric scores. The combination of all these functionalities makes it easy to know which are the highest/lowest-scored sentences in a test set.

We have also integrated a graphical library[2] to generate real-time interactive plots to show the metric scores graphically. The current version of the interface shows interactive bar charts, where different metrics and systems can be combined in the same plot. An example is shown in Figure 4.

4.2 Graphically-aided Error Analysis and Diagnosis

Human analysis is crucial in the development cycle because humans have the capability to spot errors and analyze them subjectively, in relation to the underlying system that is being examined and the scores obtained. Our purpose, as mentioned previously, is to generate a graphical representation of the information related to the source and the translations, enabling a visual analysis of the errors. We have focused on the linguistic measures at the syntactic and semantic level, since they are more robust than lexical metrics when comparing systems based on different paradigms. On the one hand, one of the views of the interface allows a user to navigate and inspect the segments of the test set. This view highlights the elements in the sentences that match a

[2]http://www.highcharts.com/

given criteria based on the various linguistic annotations aforementioned (e.g., PoS prepositions). The interface integrates also the mechanisms to upload word-by-word alignments between the source and any of the candidates. The alignments are also visualized along with the rest of the annotations, and they can be also used to calculate artificial annotations projected from the source in such test beds for which there is no linguistic processors available. On the other hand, the web application includes a library for SVG graph generation in order to create the dependency and the constituent trees dynamically (as shown in Figure 3).

4.3 Accessing the Demo

The online interface is fully functional and accessible at `http://nlp.lsi.upc.edu/asiya/`. Although the ASIYA toolkit is not difficult to install, some specific technical skills are still needed in order to set up all its capabilities (i.e., external components and resources such as linguistic processors and dictionaries). Instead, the online application requires only an up to date browser. The website includes a tarball with sample input data and a video recording, which demonstrates the main functionalities of the interface and how to use it.

The current web-based interface allows the user to upload up to five candidate translation files, five reference files and one source file (maximum size of 200K each, which is enough for test bed of about 1K sentences). Alternatively, the command based version of ASIYA can be used to intensively evaluate a large set of data.

5 Related Work

In the literature, we can find detailed typologies of the errors produced by MT systems (Vilar et al., 2006; Farrús et al., 2011; Kirchhoff et al., 2007) and graphical interfaces for human classification and annotation of these errors, such as BLAST (Stymne, 2011). They represent a framework to study the performance of MT systems and develop further refinements. However, they are defined for a specific pair of languages or domain and they are difficult to generalize. For instance, the study described in (Kirchhoff et al., 2007) focus on measures relying on the characterization of the input documents (source, genre, style, dialect). In contrast, Farrús et al. (2011) classify the errors that arise during Spanish-Catalan translation at several levels: orthographic, morphological, lexical, semantic and syntactic errors.

Works towards the automatic identification and classification of errors have been conducted very recently. Examples of these are (Fishel et al., 2011), which focus on the detection and classification of common lexical errors and misplaced words using a specialized alignment algorithm; and (Popović and Ney, 2011), which addresses the classification of inflectional errors, word reordering, missing words, extra words and incorrect lexical choices using a combination of WER, PER, RPER and HPER scores. The AMEANA tool (Kholy and Habash, 2011) uses alignments to produce detailed morphological error diagnosis and generates statistics at different linguistic levels. To the best of our knowledge, the existing approaches to automatic error classification are centered on the lexical, morphological and shallow syntactic aspects of the translation, i.e., word deletion, insertion and substitution, wrong inflections, wrong lexical choice and part-of-speech. In contrast, we introduce additional linguistic information, such as dependency and constituent parsing trees, discourse structures and semantic roles. Also, there exist very few tools devoted to visualize the errors produced by the MT systems. Here, instead of dealing with the automatic classification of errors, we deal with the automatic selection and visualization of the information used by the evaluation measures.

6 Conclusions and Future Work

The main goal of the ASIYA toolkit is to cover the evaluation needs of researchers during the development cycle of their systems. ASIYA generates a number of linguistic analyses over both the candidate and the reference translations. However, the current command-line interface returns the results only in text mode and does not allow for fully exploiting this linguistic information. We present a graphical interface showing a visual representation of such data for monitoring the MT development cycle. We believe that it would be very helpful for carrying out tasks such as error analysis, system comparison and graphical representations.

The application described here is the first release of a web interface to access ASIYA online. So far, it includes the mechanisms to analyze 4 out of 10 categories of metrics: shallow parsing, dependency parsing, constituent parsing and named entities. Nonetheless, we aim at developing the system until we cover all the metric categories currently available in ASIYA.

Regarding the analysis of the sentences, we have conducted a small experiment to show the ability of the interface to use word level alignments between the source and the target sentences. In the near future, we will include the mechanisms to upload also phrase level alignments. This functionality will also give the chance to develop a new family of evaluation metrics based on these alignments.

Regarding the interactive aspects of the interface, the grammatical graphs are dynamically generated in SVG format, which proffers a wide range of interactive functionalities. However their interactivity is still limited. Further development towards improved interaction would provide a more advanced manipulation of the content, e.g., selection, expansion and collapse of branches.

Concerning the usability of the interface, we will add an alternative form for text input, which will require users to input the source, reference and candidate translation directly without formatting them in files, saving a lot of effort when users need to analyze the translation results of one single sentence.

Finally, in order to improve error analysis capabilities, we will endow the application with a search engine able to filter the results according to varied user defined criteria. The main goal is to provide the mechanisms to select a case set where, for instance, all the sentences are scored above (or below) a threshold for a given metric (or a subset of them).

Acknowledgments

This research has been partially funded by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03) and the European Community's Seventh Framework Programme under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009- 4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4- 247914).

References

Enrique Amigó, Julio Gonzalo, Jesús Giménez, and Felisa Verdejo. 2011. Corroborating text evaluation results with heterogeneous measures. In *Proc. of the EMNLP, Edinburgh, UK*, pages 455–466.

Mireia Farrús, Marta R. Costa-Jussà, José B. Mariño, Marc Poch, Adolfo Hernández, Carlos Henríquez, and José A. Fonollosa. 2011. Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan—Spanish Language Pair. *LREC*, 45(2):181–208.

Mark Fishel, Ondřej Bojar, Daniel Zeman, and Jan Berka. 2011. Automatic Translation Error Analysis. In *Proc. of the 14th TSD*, volume LNAI 3658. Springer Verlag.

Jesús Giménez and Lluís Màrquez. 2008. Towards Heterogeneous Automatic MT Error Analysis. In *Proc. of LREC, Marrakech, Morocco*.

Jesús Giménez and Lluís Màrquez. 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4):77–86.

Ahmed El Kholy and Nizar Habash. 2011. Automatic Error Analysis for Morphologically Rich Languages. In *Proc. of the MT Summit XIII, Xiamen, China*, pages 225–232.

Katrin Kirchhoff, Owen Rambow, Nizar Habash, and Mona Diab. 2007. Semi-Automatic Error Analysis for Large-Scale Statistical Machine Translation Systems. In *Proc. of the MT Summit XI*, Copenhagen, Denmark.

Maja Popović and Hermann Ney. 2011. Towards Automatic Error Analysis of Machine Translation Output. *Computational Linguistics*, 37(4):657–688.

Maja Popović, Hermann Ney, Adrià de Gispert, José B. Mariño, Deepa Gupta, Marcello Federico, Patrik Lambert, and Rafael Banchs. 2006. Morpho-Syntactic Information for Automatic Error Analysis of Statistical Machine Translation Output. In *Proc. of the SMT Workshop*, pages 1–6, New York City, USA. ACL.

Maja Popović. 2011. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–68.

Sara Stymne. 2011. Blast: a Tool for Error Analysis of Machine Translation Output. In *Proc. of the 49th ACL, HLT, Systems Demonstrations*, pages 56–61.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In *Proc. of the LREC*, pages 697–702, Genoa, Italy.