PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 94 SEPTEMBER 2010 77–86

Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation

Jesús Giménez, Lluís Màrquez

TALP Research Center, Universitat Politècnica de Catalunya

Abstract

This article describes the Asiya Toolkit for Automatic Machine Translation Evaluation and Meta-evaluation, an open framework offering system and metric developers a text interface to a rich repository of metrics and meta-metrics.

1. Introduction

Evaluation methods are a key ingredient in the development cycle of Machine Translation (MT) systems (see Figure 1). They are used to identify the system weak points (error analysis), to adjust the internal system parameters (system refinement) and to measure the system performance, as compared to other systems or to different versions of the same system (evaluation). Evaluation methods are not a static component. On the contrary, far from being perfect, they evolve in the same manner that MT systems do. Their development cycle is similar: their weak points are analyzed, they are refined, and they are compared to other metrics or to different versions of the same metric so as to measure their effectiveness. For that purpose they rely on additional meta-evaluation methods.

In this article, we present Asiya, an open toolkit aimed at covering the evaluation needs of system and metric developers along the development cycle[1]. In short, Asiya

[1]Asiya was the Israelite wife of the Pharaoh who adopted Moses after her maids found him floating in the Nile river (see `http://en.wikipedia.org/wiki/Asiya`). The Asiya toolkit is the natural evolution/extension of its predecessor, the IQMT Framework (Giménez and Amigó, 2006). Asiya is publicly available at `http://www.lsi.upc.edu/~nlp/Asiya`.

Corresponding author: jgimenez@lsi.upc.edu
Cite as: Jesús Giménez, Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-) Evaluation. The Prague Bulletin of Mathematical Linguistics No. 94, 2010, pp. 77–86.
doi: 10.2478/v10108-010-0022-6.

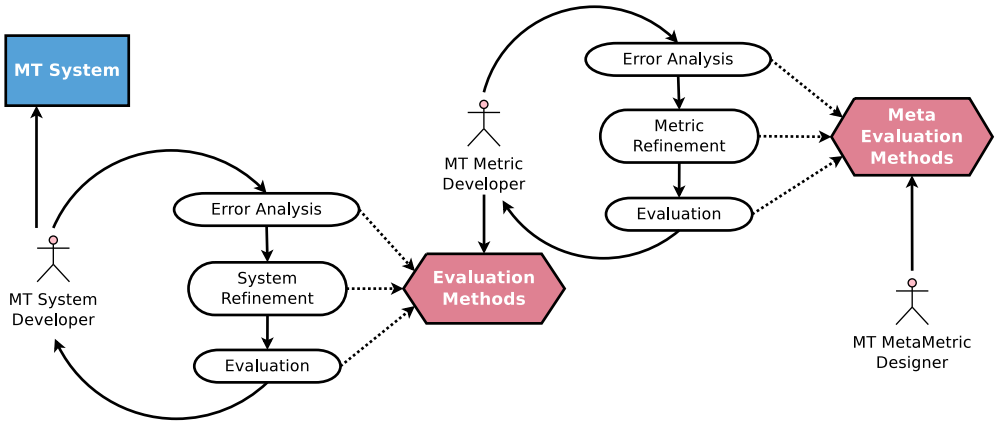

Figure 1. System development cycle in Machine Translation

is a common interface to a compiled collection of evaluation and meta-evaluation methods (i.e., hexagonal boxes in Figure 1). The metric repository incorporates the latest versions of most popular metrics, operating at different linguistic dimensions (lexical, syntactic, and semantic) and based on different similarity assumptions (precision, recall, overlap, edit rate, etc.). Asiya also incorporates schemes for metric combination, i.e., for integrating the scores conferred by different metrics into a single measure of quality. The meta-metric repository includes both measures based on human acceptability (e.g., correlation with human assessments), and human likeness, such as Orange (Lin and Och, 2004a) and King (Amigó et al., 2005).

2. Tool Description

Asiya operates over predefined test suites, i.e., over fixed sets of translation test cases (King and Falkedal, 1990). A test case consists of a source segment, a set of candidate translations and a set of manually-produced reference translations. The utility of a test suite is intimately related to its representativity, which depends on a number of variables (e.g., language pair, translation domain, number and type of references, system typology, etc.). These variables determine the space in which MT systems and evaluation metrics will be allowed to express their capabilities, and, therefore, condition the results of any evaluation and meta-evaluation process conducted upon them.

Asiya requires the user to provide the test suite definition through a configuration file. Different test suites must be placed in different folders with their correspond-

ing configuration files. Preferred input format is the NIST XML as specified in the Metrics MaTr Evaluation Plan (Callison-Burch et al., 2010)[2]. For instance, the sample configuration file in Table 1 defines source material (source.xml), candidate translations (candidates.xml), and reference translations (references.xml). If the source file is not provided, the first reference will be used as source for those metrics which take it into consideration. Candidate and reference files are required.

```
# lines starting with '#' are ignored

src=source.xml
sys=candidates.xml
ref=references.xml

some_metrics=-TERp METEOR-pa CP-STM-6 DP-Or(*) SR-Or(*) DR-Or(*) DR-STM-6
some_systems=system01 system05 system07
some_refs=reference02 reference04
```

Table 1. Sample configuration file ('sample.config')

Asiya may be then called by typing the following on the command line: `Asiya.pl sample.config'. When called without any additional option further than the name of the configuration file, Asiya will read the file and check its validity (i.e., whether the defined files exist and are well-formed). No output will be delivered to the user other than status and error messages. However, several files will be generated. Input XML files are processed and texts are extracted and saved as plain '.txt' files in the original data folder. There will be one source file, and as many candidate and reference files as systems and reference sets are specified in the XML file. The correspondence between text files and document and segment identifiers is kept through simple index files ('.idx').

2.1. Evaluation Options

Evaluation reports are generated using the '-eval' option followed by a comma-separated list of evaluation schemes to apply. Three schemes are currently available:

- **Single** metric scores
- **Ulc** normalized arithmetic mean of metric scores
- **Queen** scores as defined by Amigó et al. (2005)

[2] http://www.nist.gov/itl/iad/mig/metricsmatr10.cfm

Several output formats are available through the '-o' option. Default format is '-o mmatrix' (one system, doc or segment per line, each metric in a different column). By default metrics are sorted according to the order as typed by the user. It is also possible to sort them alphabetically using the '-sorted name' option. Other output formats are '-o smatrix' (one metric per line, each system in a different column) and 'o nist' which saves metric scores into files complying with the NIST output format as specified in the Metrics MaTr Evaluation Plan.

As an additional option, evaluation scores for the reference translations may be also retrieved through the '-include_refs' option. References will be evaluated against all other references in the test suite.

Besides evaluation reports, Asiya generates, for convenience, several intermediate files:

- **Metric scores:** Results of metric executions are stored in the './scores/' folder in the working directory, so as to avoid having to re-evaluate already evaluated translations. It is possible, however, to force metric recomputation by setting the '-remake' flag. Moreover, because each metric generates its reports in its own format, we have designed a specific XML representation format which allows us to access metric scores in a unified manner.
- **Linguistic annotations:** Metrics based on syntactic and semantic similarity may perform automatic linguistic processing of the source, candidate and reference material. When necessary, these will be stored in the original data folder so as to avoid having to repeat the parsing of previously parsed texts.

2.2. Meta-Evaluation Options

Meta-evaluation reports are generated using the '-metaeval' option followed by a comma-separated list of metric combination schemes and a comma-separated list of meta-evaluation criteria to apply. Five criteria are currently available:

- **Pearson** correlation coefficients
- **Spearman** correlation coefficients
- **Kendall** correlation coefficients
- **King** scores (Amigó et al., 2005)
- **Orange** scores (Lin and Och, 2004a)

In order to compute correlation coefficients, human assessments must be provided using the '-assessments' option followed by the name of the file containing them. The assessments file must comply with the NIST CSV format (i.e., comma-separated fields, one assessment per line.

By default, correlation coefficients are accompanied by 95% confidence intervals computed using the Fisher's z-distribution. It is also possible to compute correlation coefficients and confidence intervals applying bootstrap resampling (Koehn, 2004), If the number of samples is reasonably small, as it may be the case when computing correlation with system-level assessments, exhaustive resampling is feasible ('-ci

xbootstrap'). Otherwise, the number of resamplings may be selected using the '-ci bootstrap' and '-n_resamplings' options (1,000 resamplings by default). Also, the degree of statistical may be adjusted using the '-alfa' option. Asiya implements also paired metric bootstrap resampling. All metrics are compared pairwise. The proportion of times each metric outperforms the other, in terms of the selected criterion, is retrieved.

Finally, Asiyaprovides a mechanism to determine optimal metric sets. These may be found using the '-optimize' option followed by a specific evaluation scheme and meta-evaluation criterion (see Section 2.2).

2.3. General Options

Input Format Candidate and reference translations may be represented in a single file or in separate files. Apart from the NIST XML format, previous NIST SGML and plain text formats are also accepted. Input format is specified using the '-i' option followed by any of the formats available ('nist' or 'raw').

Language Pair By default, Asiya assumes the test suite to correspond to an into-English translation task. This behavior may be changed using the '-srclang' (source language) and 'trglang' (target language) options. Metrics based on linguistic analysis, or using dictionaries or paraphrases, require a proper setting of these values. It is also possible to tell Asiya whether text case matters or not. By default, Asiya will assume the text to be case-sensitive. This behavior may be changed using the '-srccase' (source case) '-trgcase' (target case) options.

Pre-defined Sets The set of metrics to used may be specified using the '-metric_set' and/or the '-m' options. The '-metric_set' option must be followed by the name of the set as specified in the config file (see Table 1). The '-m' option must be followed by a comma-separated list of metric names. The effect of these options is cumulative. Analogously, you may tell Asiya to focus on specific system sets ('-system_set' and '-s') and reference sets ('-reference_set' and '-r'). The full list of metric system and reference names defined in the test suite may be listed using the '-metric_names', '-system_names' and '-reference_names' options, respectively[3].

Other Options Another important parameter is the granularity of the results. Setting the granularity allows developers to perform separate analyses of system-level, document-level and segment-level results, both over evaluation and meta-evaluation reports. This parameter may be set using the '-g' option. Default granularity is at the system level. The length and precision of floating point numbers may be adjusted using the '-float_length' (10 by default) and '-float_precision' options (8 by default). Finally, the '-tex' flag produces, when applicable, (meta-)evaluation reports directly in LaTeX format.

[3]The set of available metrics depends on language pair settings.

3. Metric Set

Today, Asiya includes repository of more than 600 metrics. In the following, we provide a brief description. We have grouped metrics according to the linguistic level at which they operate.

- **Lexical Similarity**

  BLEU Eight variants for different n-gram lengths, cumulative and non-cumulative, and smoothed or not, have been considered (Papineni et al., 2001).

  NIST Ten variants for different n-gram lengths, cumulative and non-cumulative, and smoothed or not, have been considered (Doddington, 2002).

  GTM . We included three variants taking different values of the e parameter ($e \in \{1, 2, 3\}$) weighting the importance of the length of matching n-grams (Melamed et al., 2003).

  METEOR Four variants, progressively adding 'exact', 'stem', 'synonym' and 'paraphrase' modules have been considered (Denkowski and Lavie, 2010).

  ROUGE Eight variants, for different n-gram lengths, allowing for skip bigrams or not, weighted or not, have been considered (Lin and Och, 2004b).

  TERp Four variants, with and without paraphrasing support, have been included (Snover et al., 2009).

  $\mathbf{O_l}$ Lexical overlap (Giménez and Màrquez, 2010).

- **Syntactic Similarity**

  Shallow Parsing (SP) Average lexical overlap over parts of speech, and base phrase chunk types, and NIST score over sequences of lemmas, parts of speech, and chunks (Giménez and Màrquez, 2010).

  Dependency Parsing (DP) Head-word chain matching (Liu and Gildea, 2005) over word forms, grammatical categories and relations, and average lexical overlap between tree nodes according to their tree level, category or relation (Giménez and Màrquez, 2010).

  Constituency Parsing (CP) Average lexical overlap over parts of speech and syntactic constituents (Giménez and Màrquez, 2010), and syntactic tree matching (Liu and Gildea, 2005).

- **Semantic Similarity**

  Named Entities (NE) Average lexical overlap between NEs according to their type (Giménez and Màrquez, 2010).

  Semantic Roles (SR) Average lexical overlap between SRs according to their type, and average role overlap, i.e., overlap between semantic roles independently from their lexical realization (Giménez and Màrquez, 2010).

  Discourse Representations (DR) Average lexical and morphosyntactic overlap between DRs according to their type (Giménez and Màrquez, 2010).

metric	bbn-combo	dcu	dcu-combo	google	jhu	lium-systran	rbmt3
BLEU$_s$	0.31	0.27	0.31	0.31	0.27	0.27	0.20
NIST	7.95	7.36	7.91	8.05	7.31	7.33	6.22
GTM$_2$	0.28	0.25	0.29	0.29	0.24	0.26	0.22
ROUGE$_W$	0.35	0.32	0.34	0.34	0.32	0.32	0.29
-TER$_p$	-0.47	-0.50	-0.48	-0.46	-0.51	-0.51	-0.58
METEOR$_{pa}$	0.55	0.53	0.55	0.54	0.52	0.52	0.49
SP-NIST$_p$	6.85	6.40	6.92	7.07	6.24	6.49	5.88
CP-STM$_6$	0.40	0.38	0.39	0.41	0.37	0.38	0.35
DP-HWC$_w$	0.21	0.18	0.20	0.22	0.18	0.18	0.15
DP-HWC$_c$	0.34	0.32	0.32	0.35	0.32	0.32	0.30
DP-HWC$_r$	0.30	0.28	0.28	0.31	0.28	0.28	0.26
DP-O$_r$($\star$)	0.25	0.23	0.24	0.26	0.22	0.23	0.19
NE-O$_e$($\star$)	0.38	0.35	0.40	0.40	0.29	0.38	0.34
SR-O$_r$($\star$)	0.24	0.20	0.23	0.24	0.20	0.20	0.18
DR-O$_r$($\star$)	0.32	0.29	0.31	0.33	0.28	0.29	0.24
DR-O$_{rp}$($\star$)	0.48	0.45	0.47	0.48	0.44	0.45	0.44
DR-STM$_6$	0.45	0.42	0.44	0.46	0.41	0.43	0.39

Table 2. ASIYA-generated evaluation report (system level), WMT09 fr-en

4. A Use Case

In this section, we illustrate some of the Asiya functionalities over a particular test suite. Specifically, we have used the French-English (fr-en) translation task from the 2009 ACL Workshop on Machine Translation, WMT09, (Callison-Burch et al., 2009). There have been three main reasons for selecting this test bed: (i) it is publicly available, (ii) it is reasonably heterogeneous, since it includes system based on different paradigms (statistical vs. rule-based, hybrid, combined), and (iii) it is neutral, since systems are evaluated out-of-domain, i.e., in a domain other than the training domain.

The test suite consists of 111 documents totaling 2525 segments, one reference translation and automatic translations by 21 different systems. Human assessments at the segment level based on different criteria are available for a subset of segments.

First, we use Asiya to evaluate a subset of the participant systems based on a selected set of metrics operating at different linguistic levels. We use the '-tex' flag to generate directly the table in LaTeX format[4]. The output is Table 2.

[4]The command is the following: `Asiya.pl -v -m BLEUs,NIST,GTM-2,ROUGE-W,-TERp,METEOR-pa, SP-pNIST,CP-STM-6,DP-HWC_w-4,DP-HWC_c-4,DP-HWC_r-4,DP-Or(*),NE-Oe(*),SR-Or(*),DR-Or(*), DR-Orp(*),DR-STM-6 -s bbn-combo,dcu,dcu-combo,google,jhu,lium-systran,rbmt3 -eval single -o smatrix -float_precision 2 -g sys -tex Asiya.config'.

metric	ρ	confidence interval
BLEU$_s$	0.90	(0.76, 0.97)
NIST	0.89	(0.66, 0.97)
GTM$_2$	0.89	(0.72, 0.97)
ROUGE$_W$	0.93	(0.80, 0.98)
-TER$_p$	0.86	(0.66, 0.96)
METEOR$_{pa}$	0.91	(0.78, 0.98)
SP-NIST$_p$	0.83	(0.58, 0.94)
CP-STM$_6$	0.93	(0.79, 0.99)
DP-HWC$_w$	0.91	(0.75, 0.98)
DP-HWC$_c$	0.96	(0.88, 0.99)
DP-HWC$_r$	0.94	(0.83, 0.99)
DP-O$_r$ $(\star)$	0.93	(0.81, 0.98)
NE-O$_e$ $(\star)$	0.67	(0.29, 0.87)
SR-O$_r$ $(\star)$	0.93	(0.80, 0.98)
DR-O$_r$ $(\star)$	0.93	(0.77, 0.98)
DR-O$_{rp}$ $(\star)$	0.92	(0.76, 0.98)
DR-STM$_6$	0.93	(0.80, 0.99)

Table 3. ASIYA-generated meta-evaluation report (system level), WMT09 fr-en

Now, let us use Asiya to evaluate a selected set of metrics. Since we count on human assessments we can compute correlation coefficients. For this example we have used the 'rank' assessments. Each assessor was presented with a set of translation outputs to be ranked from best to worst being 1 assigned to the best output, 2 to the second best and so on. The total number of assessments is 2,668. We take the negative rank as a positive measure of quality. With this kind of assessments, segment-level Pearson correlation coefficients would not be very reliable/informative. We can, however, compute Spearman correlation coefficients at the system level. Confidence intervals are computed via bootstrap resampling at a 95% statistical significance[5].

5. Ongoing and Future Steps

Current development of the toolkit goes in two main directions. First, we are augmenting the metric repository. We are incorporating new metrics and we are porting linguistic metrics to other languages. We also plan to design and implement a mech-

[5]The command is the following: `Asiya.pl -v -m BLEUs,NIST,GTM-2,ROUGE-W,-TERp,METEOR-pa, SP-pNIST,CP-STM-6,DP-HWC_w-4,DP-HWC_c-4,DP-HWC_r-4,DP-Or(*),NE-Oe(*),SR-Or(*),DR-Or(*), DR-Orp(*),DR-STM-6 -metaeval single spearman -assessments data/rank.csv -ci bootstrap -n_resamplings 1000 -float_precision 2 -g sys -tex Asiya.config'.

anism so users can easily incorporate their own metrics. Moreover, we are currently implementing measures for confidence estimation (i.e., when the reference translation is not available). Also, in the future, we plan to consider more sophisticated metric combination schemes and alternative meta-evaluation criteria.

The second direction is on the construction of a visual interface for Asiya. We are designing a web application for monitoring the whole development cycle. This application will allow system and metric developers to upload their test suites and perform error analysis, automatic and manual evaluation, and meta-evaluation, using their Internet browser.

Acknowledgements

This work has been partially funded by the Spanish Government (OpenMT-2, TIN-2009-14675-C03) and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement numbers 247762 (FAUST, FP7-ICT-2009-4-247762) and 247914 (MOLTO, FP7-ICT-2009-4-247914).

Bibliography

Amigó, Enrique, Julio Gonzalo, Anselmo Penas, and Felisa Verdejo. QARLA: a Framework for the Evaluation of Automatic Summarization. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 280–289, 2005.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, 2009.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, 2010.

Denkowski, Michael and Alon Lavie. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, July 2010.

Doddington, George. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology*, pages 138–145, 2002.

Giménez, Jesús and Enrique Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 685–690, 2006.

Giménez, Jesús and Lluís Màrquez. Linguistic Features for Automatic MT Evaluation. *To Appear in Machine Translation*, 2010.

King, Margaret and Kirsten Falkedal. Using Test Suites in Evaluation of MT Systems. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING)*, pages 211–216, 1990.

Koehn, Philipp. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, 2004.

Lin, Chin-Yew and Franz Josef Och. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 501–507, 2004a.

Lin, Chin-Yew and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2004b.

Liu, Ding and Daniel Gildea. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32, 2005.

Melamed, I. Dan, Ryan Green, and Joseph P. Turian. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, 2003.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation, RC22176. Technical report, IBM T.J. Watson Research Center, 2001.

Snover, Matthew, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, 2009.

Address for correspondence:
Jesús Giménez
jgimenez@lsi.upc.edu
Universitat Politècnica de Catalunya
C/Jordi Girona, 1-3. Campus Nord. Edifici Omega, despatx S-107
Barcelona, 08028. Spain