The NIST Metrics for MAchine TRanslation 2010 Challenge (MetricsMaTr10)

Evaluation Plan

Version: 1-3 Date: April 19, 2010

Questions related to information in this document should be addressed to: mt_poc@nist.gov.

MetricsMaTr website: http://www.nist.gov/itl/iad/mig/matr.cfm

1 Introduction

The NIST 2010 Metrics for Machine Translation Challenge (MetricsMaTr10) is the second event in an ongoing series of evaluations of machine translation (MT) metrics.

NIST has been conducting formal evaluations of machine translation technology since 2002, and while the evaluations have been successful, there was, and remains, a need for a better understanding of exactly how useful the state-of-the-art technology is, and how to best interpret the scores reported during evaluation.

This need exists primarily due to the shortcomings with the current methods employed for the evaluation of Machine Translation technology:

- Automatic metrics have not yet been proved able to consistently predict the usefulness, adequacy, and reliability of MT technologies.
- Automatic metrics have not demonstrated that they are as meaningful in target languages other than English.
- Human assessments are expensive, slow, subjective, and are difficult to standardize. Furthermore, they only pertain to the translations evaluated, and are of no use even to updated translations from the same system.
- Both automatic metrics and human assessments need more insights into what properties of the translation should be evaluated, as well as insights into how to evaluate those properties.
- Some MT technology approaches evaluated incorporate algorithms that optimize scores on MT metric(s). These optimizations fail in the same respects that the metrics fail.

These problems, and the need to overcome them through the development of improved automatic (and even semi-automatic) metrics, have been a point of discussion at past NIST MT evaluations. Without more appropriate metrics to address these shortcomings, the impact of formative and summative MT technology evaluations remains limited.

This situation led to the Metrics for Machine Translation (MetricsMaTr) challenge, an evaluation series focused entirely on MT metrology. In MetricsMaTr, innovative MT metrics, rather than MT technology itself, are the subject of evaluation. NIST successfully implemented the first instance of MetricsMaTr in 2008.[1] 32 new metrics of various approaches were evaluated. The results of MetricsMaTr08 are available online.[2] As a result of the first round of MetricsMaTr, the landscape of metrics available to open evaluations such as NIST OpenMT has begun to become richer and more diverse. A special issue of the journal *Machine Translation* contains an overview paper describing MetricsMaTr08[3], as well as several papers describing new metrics that were submitted to MetricsMaTr08.

[1] http://www. http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008

[2] http://www.itl.nist.gov/iad/mig/tests/metricsmatr/2008/results

[3] Przybocki, M., Peterson, K., Bronsart, S., & Sanders, G. (2010). *Machine Translation.* DOI 10.1007/s10590-009-9065-6.

For its second evaluation, the MetricsMaTr challenge joins efforts with the Fifth Workshop on Statistical Machine Translation (WMT)[4] as the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMaTr10[5]. The WMT series has included an MT metric evaluation component (the "shared evaluation task") since 2008. Joining this component of WMT with MetricsMaTr in 2010 will allow metric developers to focus their efforts on one rather than two similar MT metric evaluations. Also, the joint effort will allow testing submitted metrics on a substantially larger amount of MT output data – the NIST MetricsMaTr data as well as the WMT data.

The workshop that concludes the evaluation will be a two-day event held as a satellite workshop of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010) in Uppsala, Sweden, July 15-16, 2010. The goal of the MetricsMaTr part of the workshop is to inform other MT technology evaluation campaigns and conferences with regard to improved MT metrology.

MetricsMaTr does not exclude modifications and extensions of existing metrics. However, there is a strong emphasis on the development of clearly innovative, even revolutionary, metrics that have the potential to once more initiate a substantial paradigm shift in the field of MT metrology, much as the introduction of BLEU[6] did in 2001.

Participants are encouraged to be adventurous and creative in their metric development and to avoid being overly influenced by techniques that have already been attempted.[7]

The MetricsMaTr challenge is designed to appeal to a wide and varied audience including researchers of MT technology and metrology, acquisition programs such as MFLTS, and commercial vendors. We welcome submissions from a wide range of disciplines including computer science, statistics, mathematics, linguistics, and psychology. NIST encourages submissions from participants not currently active in the field of MT.

2 Evaluation Procedures

Metric developers are required to develop software that implements a scoring algorithm which assesses machine translation quality. The scoring software is to be packaged and submitted to NIST for evaluation. The submitted package should identify system requirements, minimum versions of installed tools, and include a short description of interpretation of the scores (e.g., accuracy vs. error). A small data set will be provided as an installation check set; participants will be required to score this set using their metric(s), and submit the scores along with their metric(s). NIST will install the metric(s) and check those scores against scores on the same set generated by the instance of the metric(s) installed at NIST, to ensure that the metric was installed and is working as intended.

NIST will then score output from a variety of MT systems (the evaluation sets described in sections 3.1.2 and 3.2.2) with the submitted metrics, and will determine how well the scores produced by each metric correlate with carefully created human assessments of the same MT output.

3 Data

Data is an important component of metric evaluation. MetricsMaTr10 will make use of two evaluation test sets, and several development test sets. Explained in detail below, the evaluation test sets will be partitioned into the NIST MetricsMaTr test data that remains blind, and the WMT evaluation data that will be released post-evaluation. Development data will be provided that is similar to each of the evaluation test sets. Metrics submitted to MetricsMaTr10 will be evaluated separately for the NIST MetricsMaTr and WMT data sets, as well as over all the data.

---

[4] http://www.statmt.org

[5] http://www.statmt.org/wmt10

[6] Papineni, K., S. Roukos, T. Ward, & W.J. Zhu (2002). "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002)*.

[7] Section 6 describes some practical limitations and guidelines for metric development.

3.1 NIST MetricsMaTr Data Sets

The NIST MetricsMaTr data sets for 2010 remain unchanged from the MetricsMaTr08 data sets. **It is possible that they will be updated with more data and/or more or different human assessments before the evaluation**.

3.1.1 Development Data

The MetricsMaTr development (DEV) data set will be available to metric developers upon registering for participation in the evaluation[8] and submitting a data license agreement to the Linguistic Data Consortium (LDC).[9] (See section 10 for the target availability date of the DEV set.) The DEV data consists of:

- Several versions of system translations
- Up to four independently created reference translations
- Segment level human assessments of *adequacy*
    - Document and system level scores created as described in section 5
- Segment level human judgments of *preference*
- (Source data available on request, may require a separate license agreement)

The DEV data comes from the NIST OpenMT06[10] evaluation and from DARPA TRANSTAC training dialogs. The documents from OpenMT06 were selected by examining the document level BLEU scores across several systems. NIST hand-selected each document to provide varying levels of performance, as determined by theOpenMT06 official evaluation metric, BLEU. See Table 1 for DEV corpus information. For the TRANSTAC dialogs, the offline training data was readily available and was included to provide some sampling of an alternative data style.

Table 1: MetricsMaTr DEV Statistics

Source of Data	MT06	TRANSTAC
Genre	newswire	training dialogs
Number of documents/scenario	25	1*
Total number of segments	249	17
Source Language	Arabic	Iraqi Arabic
Number of system translations	8	5

*The single document for the TRANSTAC training dialog contains 17 segments of unrelated text.

System translations: Translations from multiple systems are included in the DEV set. Five systems are included from the January 2007 TRANSTAC offline evaluation. Eight anonymized systems were selected from the OpenMT06 evaluation, selected to cover a variety of MT algorithmic approaches (statistical MT, rule-based MT, hybrids …).

Reference translations: OpenMT06 reference translations are provided. The Linguistic Data Consortium supplied four independently created translations for each document. Each translation provider was given the same set of translation guidelines[11]. References for the TRANSTAC dialog data were created by Appen, using specific guidelines for transcription and translation.

[8] http://www.nist.gov/itl/iad/mig/NISTMetricsMaTr10Registration.pdf
[9] http://www.nist.gov/itl/iad/mig/NISTMetricsMaTr10LDCAgreement_v1-1.pdf
[10] The 2006 NIST Open MT evaluation is documented here: http://www.nist.gov/speech/tests/mt/2006/doc. Rights to the evaluation test set are handled by the Linguistic Data Consortium; see http://www.ldc.upenn.edu.
[11] http://projects.ldc.upenn.edu/translation/MT08

Human assessments of adequacy: Assessments of adequacy were performed using an application developed at NIST, "TAP-ET" for "Translation Adequacy and Preference Evaluation Tool".[12] Each translation segment was assessed by two judges. After independently and completely assessing the entire DEV set, the judges reviewed their individual assessments together and settled on a single final score.

Human Assessments of preference: Assessments of preference were also performed using the NIST TAP-ET application. Every translation segment was compared to every other corresponding translation by two judges. After completely assessing the entire DEV for preferences, the judges reviewed their individual preferences together and settled on a single final preference.

Source Data: The source data can be made available if required by the developed metric. Contact mt_poc@nist.gov to discuss the implications of signing the required license agreement.

3.1.2 Evaluation Data

The evaluation (EVAL) data set will not be distributed to participants. The data set includes data from OpenMT08, GALE, and TRANSTAC evaluations. The EVAL set is similar to, but more expansive than, the DEV set, and includes different systems and data translated from different source languages not represented in the DEV set, allowing for analysis on data with properties on which the metrics could not have been specifically tuned. To the extent possible, the evaluation data will be categorized in ways that might assist in interpretation of the results (e.g., by Interagency Language Roundtable (ILR) level or genre).

The same adequacy and preference judgments as described for the DEV set above were created for the EVAL set and serve as the primary human assessments for MetricsMaTr10. Multiple independent judgments on a segment were not adjudicated into a single score, but transformed into a single score computationally (e.g., by averaging across judgments, for adequacy). Several subsets of the data are annotated with additional pre-existing, typically program-specific, assessments (e.g., HTER for the GALE portion of the data); these will be included in the analyses as well.

3.2 WMT Data Sets

3.2.1 Development Data

Previous years' WMT data (system translations, human reference translations, source data, and human assessments) are publicly available for purposes of metric development.[13]

3.2.2 Evaluation Data

The WMT10 evaluation data set will consist of WMT10 submitted system translations, human reference translations, and the corresponding WMT10 human assessments[14] for all WMT10 language pairs:

- English-German, German-English
- English-French, French-English
- English-Spanish, Spanish-English
- English-Czech, Czech-English

[12] Przybocki, M., K. Peterson, & S. Bronsart (2008). "Translation Adequacy and Preference Evaluation Tool (TAP-ET)." *Proceedings of the Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.
[13] http://www.statmt.org/wmt08/results.html, http://www.statmt.org/wmt09/results.html
[14] See http://statmt.org/wmt10/translation-task.html for a brief description of WMT10's human assessments.

4 Evaluation Tracks

Metrics-MATR will analyze the submitted metrics in two tracks, differing by the number of reference translations available.

4.1 Single Reference track

There is a cost associated with creating reference translations for use in evaluation. If metrics were determined to be great predictors of MT quality based on only one manual reference translation instead of multiples, evaluation data sets could grow in size. This in turn would provide better grounds for determining statistical differences.

For the Single Reference track, NIST will analyze the metric performance when limiting all of the evaluation data to one pre-selected reference translation per test segment, generated by the same reference provider.

4.2 Multiple References track

Most will agree that often there is not a single "best" or "perfect" translation of every given source sentence. For some language pairs, a reasonable translation may not be possible. There are other issues as well, such as multiple acceptable ways to handle idioms, name variants, and synonymy.

Previous experiments have found that metrics that make use of more than one independently created reference translation tend to asymptote for metric stability as measured against human judgments of quality around the use of four references. For much of the MetricsMaTr data, four reference translations will be available.

NIST will analyze submitted metric performance separately for metrics that are designed to use more than one reference.

5 Evaluating the Metrics

5.1 Correlation with Human Judgments

For this evaluation, the reference against which the metrics are measured is not a correct translation, but rather the agreed-upon subjective grade of adequacy (or preference) assigned to each translation. The method for grading the submitted metrics is defined to be how well the produced scores correlate with these grades, and how well they can provide insight as to the quality of the MT translations.

Segment level: The basic assessments were performed at the segment level; segments were presented in order for each document. Each segment received two independent judgments, and those two judgments were then adjudicated into one score.

Document level: NIST created document level assessment scores by averaging the set of segment scores, weighted by segment length.

System level: NIST created system level assessment scores by averaging the set of segment scores, weighted by segment length.

5.2 Correlation Measures

NIST will calculate the correlation coefficients between human assessments and automatic metrics for:

- Pearson's r
- Kendall's tau
- Spearman's rho

6 Metric Properties

There are many desirable properties for metrics employed for evaluation. In this section, we discuss a few general guidelines that should be considered during metric development. This list is known to be incomplete, and other evaluations may have different guidelines. Section 6.1 describes properties required for practical implementation and testing of metrics, and section 6.2 describes the characteristics or capabilities of metrics that are so clearly missing.

6.1 Practical Implementation Properties

The following set of properties describes characteristics that all automatic metrics will possess in order to be deemed useful for evaluation purposes.

- **Automaticity**: Metrics that are "automatic", that is, metrics that do not require human intervention outside the creation of the reference translations, are useful to evaluate systems over large test sets. Large test sets lead to greater power of statistical tests and allow for evaluation over greater populations. Automatic metrics can also be used in training by certain MT technology approaches.
- **Repeatability (Reliability)**: It is extremely desirable that metrics produce the exact same score each time they are used to evaluate the same set of data.
- **Portability**: Metric software should be universally usable. Metrics should not require support of antiquated software, or unusual operating systems. NIST expects that a knowledgeable system administrator will be able to install and compile all components of the developed metric within approximately four hours.
  The metric might make use of internet resources. They should be failsafe in case the internet is unavailable.
  For MetricsMaTr, the metric software will need to run on at least one of the following operating systems:
  - Windows XP
  - MAC OS X
  - Linux CENT OS 5 (or newer)
- **Speed**: Metric software should be relatively quick to run. If a metric requires more than five hours to score the complete DEV, the developer should contact NIST to discuss other options before the evaluation.
- **Limited Annotation of Reference Data**: The evaluation infrastructure will include up to four independently created reference translations for each translated segment.[15] Reference translations are created following standard translation guidelines and do not include additional mark-up for items such as proper names and alternations.
  Some algorithms may require additional mark-up, but in the implementation of the MetricsMaTr's evaluation, the references are not released. NIST suggests that in cases where the submitted metrics might benefit from additional mark-up of the references, the possibility for success be demonstrated by an automatic mark-up process. If promise is shown, NIST may invest in updating the references for future evaluations.

6.2 Metrology Objectives

The following properties are strongly sought after behaviors and capabilities that are missing from many existing automatic MT metrics.

[15] http://projects.ldc.upenn.edu/translation/MT08

- **Correlation with Human Assessments of MT Quality**: Currently, the slow, tedious, and subjective process of humans comparing system to reference translations is one of the most accepted ways of determining which systems are better than others. Thus correlations with human assessments are the primary metric to be used in this evaluation.
- **Ability to Differentiate Between Systems of Varying Quality**: To the extent possible, metrics should be able to differentiate quality between two different systems. That is, the reported scores should be fine-grained enough to rank even systems that are fairly close in quality.
- **Intuitive Interpretation**: A complaint levied against current automatic MT metrics is that the reported score is difficult to relate to quality. This makes it difficult to demonstrate how meaningful MT improvements are. To the extent possible, it is desirable that the reported score be directly related to quality and be intuitive even to persons without specific technical background in machine translation.
- **Applicability to Multiple Target Languages**: Metrics that work on a wide variety of target languages will be of most benefit.
- **Stability against Optimization**: In the framework of this evaluation, the system translations that are evaluated were not optimized for the metrics being developed. There is a chance that results on this blind evaluation data set may differ from results on translations that were optimized for the particular metric. The goal is to get away from gaming and metric tuning.

7 Metric descriptions

All metric developers participating in MetricsMaTr10 must submit an informal metric description to NIST via e-mail to mt_poc@nist.gov. These descriptions will be made available to the workshop participants, but they will not be published in the ACL 2010 proceedings. Additionally, developers are encouraged to submit a short (4-page) ACL paper outlining their metric(s) for the workshop, following the ACL guidelines for short papers;[16] the relevant dates for such submissions are as given on the joint workshop website.[17].

8 File Formats

This section describes the file format for both the input files that the metrics will be required to read, and the output files that the metrics should produce.

8.1 Metric Input Files (System Translations and Reference Translations)

Input files will be in an XML format that NIST uses for the Open MT evaluations. NIST has defined a set of XML tags that are used to format MT source, translation, and reference files for evaluation. Each set of translations for a single system will be identified in separate files. See Appendix A for detailed file format information.

8.2 Metric Output Files

Analysis of the submitted metrics will take place on various levels. MetricsMaTr prefers metrics be designed to output system, document, and segment level scores, but it may be the case that a metric is not designed to do so. In such cases, please alert NIST before the evaluation so we can prepare accordingly.

Metric developers must output scores in the format described below. This will allow for plug-in comparisons for the various correlation tests, and it will significantly reduce the possibility of human-introduced errors in a reformatting process.

One running of the software on a single translation file should produce at least three files:

[16] http://www.acl2010.org/papers.html
[17] http://www.statmt.org/wmt10

1. <systemname>-sys.scr *# System level scores*
2. <systemname>-doc.scr *# Document level scores*
3. <systemname>-seg.scr *# Segment level scores*

Contents of these three files are described below. See the current MT XML DTD[18] for definitions and Appendix A for descriptions and samples of the elements referenced here.

8.2.1 System Scores

The evaluated metric should have the capability for assigning a single overall "score" for a system. To assist in analysis, we are requiring the metric to output a system level score file "**<systemname>-sys.scr**" for each input "tstset" evaluated. The output should be a single tab separated record:

```
<TEST_ID>   <SYSTEM_ID>   <SYSTEM LEVEL SCORE>   <OPTIONAL>
```

Where:

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
SYSTEM LEVEL SCORE is the overall system level score.

Followed by optionally included items, each separated by a tab (confidence scores, statistics ...).

8.2.2 Document Scores

The evaluated metric should have the capability for assigning a score to each **document** translated by the system. Note that for some data types (e.g., transcripts of dialogs), *document* is the term we will use to refer to a single grouped exchange, scenario, or discussion. To assist in analysis, we are requiring the metric to output a document level score file "**<systemname>-doc.scr**" for each input "tstset" evaluated. The output should be a single tab separated record for each document:

```
<TEST_ID>   <SYSTEM_ID>   <DOCUMENT_ID>   <DOCUMENT LEVEL SCORE>   <OPTIONAL>
```

Where:

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
DOCUMENT_ID is the document identified by the **docid** attribute in the translation file.
DOCUMENT LEVEL SCORE is the overall document score.

Followed by optionally included items, each separated by a tab (confidence scores, statistics ...).

8.2.3 Segment Scores

The evaluated metric should have the capability for assigning a score to each **segment** translated by the system. To assist in analysis, we are requiring the metric to output a system level score file "**<systemname>-seg.scr**" for each input "tstset" evaluated. The output should be a single tab separated record for each segment:

```
<TEST_ID> <SYSTEM_ID> <DOCUMENT_ID> <SEGMENT_ID> <SEGMENT SCORE> <OPTIONAL>
```

Where:

[18] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd

TEST_ID is the particular test set identified by the **setid** attribute in the translation file.
SYSTEM_ID is the system identified by the **sysid** attribute in the translation file.
DOCUMENT_ID is the document identified by the **docid** attribute in the translation file.
SEGMENT_ID is the segment identified by the **id** attribute of the **seg** tag.
SEGMENT SCORE is the score for the particular segment.

Followed by optionally included items, each separated by a tab (confidence scores, statistics ...).

9 Workshop

The report-out session for this evaluation will be at the ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and MetricsMaTr10 in Uppsala, Sweden. While attendance is not mandatory, NIST encourages all MetricsMaTr10 participants to send a representative to the workshop to discuss their work. A section of the workshop will be dedicated to MetricsMaTr. NIST will provide an evaluation overview, an overview of the metrics submitted, and report on the correlations with human assessments as described in section 5. Select metrics will be presented in more detail by their developers.

10 Schedule

The following table outlines the important dates for this evaluation cycle. **Please see the joint workshop website[19] regarding dates for voluntary submissions of metric descriptions as short papers to ACL 2010.**

January 11 2010	MetricsMaTr08 development data set re-release for MetricsMaTr10
March 26 2010	Metric submission commitment due at NIST
March 26 - May 14 2010	Metric submission period; metrics must be installed and operational at NIST by May 14, 5pm EDT
June 18 2010	Informal metric descriptions due at NIST (mandatory)
July 15 - 16 2010	Joint Fifth Workshop on Statistical Machine Translation and MetricsMaTr (at ACL 2010 in Uppsala, Sweden)
September 16 2010	Official public release of results

[19] http://www.statmt.org/wmt10

Appendix A: NIST MT XML Data Format

The translation data to be scored by the submitted metrics will be in XML format, conforming to the standards set forth in this appendix. **Metrics must be capable of handling this data format; they cannot be accepted otherwise.**

I. System Translation File Format

A single system translation file will contain all the system translations of an identified data set (**setid**). The file format is defined by the current MT XML DTD,[20] and will begin with the following three lines (numbered for identification):

```
1. <?xml version="1.0" encoding="UTF-8"?>
2. <!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
3. <mteval>
```

Line 1: XML header, definition statement

Line 2: DTD identifier

Line 3: MTEVAL tag identifies "the beginning of a test set.

A translation file contains one or more "tstset" elements, immediately after the root "mteval" element. Each "tstset" element has the following attributes:

- "setid": The dataset.
- "srclang": The source language.
- "trglang": The target language.
- "sysid": A name identifying site and system.

Each "tstset" element contains one or more "doc" elements, which have the following attributes:

- "docid": The document.
- "genre": The data genre.

Each "doc" element contains several segments ("seg" elements). Each segment has a single attribute, "id", which must be enclosed using double quotes or single quotes.
Note that there are other possible tags that may be present in the system translation files. See the MT XML DTD for a complete description of possible tags.

Sample XML system translation file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
<mteval>
        <tstset setid="sample_set" srclang="Arabic" trglang="English" sysid="NIST">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
                        <seg id="2">ENGLISH SYSTEM TRANSLATION #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ENGLISH SYSTEM TRANSLATION #1</seg>
                        ...
                </doc>
                ...
        </tstset>
        ...
</mteval>
```

[20] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd

II. Reference Translation File Format

A single reference translation file will contain all the reference translations available for an identified data set (**setid**). Some of the system translations will have only one reference translation, while others will have more. The file format is defined by the current MT XML DTD,[21] and will begin with the following three lines (numbered for identification):

```
1. <?xml version="1.0" encoding="UTF-8"?>
2. <!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
3. <mteval>
```

Line 1: XML header, definition statement

Line 2: DTD identifier

Line 3: MTEVAL tag identifies the beginning of a test set.

A reference file contains one or more "refset" elements, immediately after the root "mteval" element. Each "refset" element has the following attributes:

- "setid": The dataset.
- "srclang": The source language.
- "trglang": The target language.
- "refid": The current reference.

Each "refset" element contains one or more "doc" elements, which have the following attributes:

- "docid": The document.
- "genre": The data genre.

Each "doc" element contains several segments ("seg" elements). Each segment has a single attribute, "id", which must be enclosed using double quotes or single quotes.

Sample XML reference translation file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
<mteval>
        <refset setid="sample_set" srclang="Arabic" trglang="English" refid="reference01">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
                        <seg id="2">ENGLISH REFERENCE TRANSLATION #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ENGLISH REFERENCE TRANSLATION #1</seg>
                        ...
                </doc>
                ...
        </refset>
        ...
</mteval>
```

[21] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd

III. Source File Format

Not all metrics will make use of the source data.

A single source file will contain the source data for an identified data set (**setid**). The file format is defined by the current MT XML DTD,[22] and will begin with the following three lines (numbered for identification):

```
1. <?xml version="1.0" encoding="UTF-8"?>
2. <!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
3. <mteval>
```

Line 1: XML header, definition statement

Line 2: DTD identifier

Line 3: MTEVAL tag identifies the beginning of a test set.

A source file contains one single “srcset” element, immediately after the root “mteval” element. The “srcset” element has the following attributes:

- “setid”: The dataset.
- “srclang”: The source language.

The “srcset” element contains one or more “doc” elements, which have the following attributes:

- “docid”: The document.
- “genre”: The data genre.

Each “doc” element contains several segments (“seg” elements). Each segment has a single attribute, “id”, which must be enclosed using double quotes or single quotes.

Sample XML source file:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd">
<mteval>
        <srcset setid="sample_set" srclang="Arabic">
                <doc docid="sample_document_1" genre="nw">
                        <seg id="1">ARABIC SENTENCE #1</seg>
                        <seg id=”2”>ARABIC SENTENCE #2</seg>
                        ...
                </doc>
                <doc docid="sample_document_2" genre="nw">
                        <seg id="1">ARABIC SENTENCE #1</seg>
                        ...
                </doc>
                ...
        </srcset>
</mteval>
```

[22] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.5.dtd