VERTa: Exploring a Multidimensional Linguistically-Motivated Metric

Elisabet Comelles and **Irene Castellón**
Grial Research Group
Universitat de Barcelona
Barcelona, Spain
{elicomelles, icastellon}@ub.edu

Jordi Atserias
Fundació Barcelona Media
Barcelona, Spain
jordi@yahoo-inc.com

Victoria Arranz and **Olivier Hamon**
ELDA/ELRA
Paris, France
{arranz, hamon}@elda.org

Abstract

This paper describes the first steps in the design and implementation of VERTa, a metric which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. A description of the modules developed up to now is provided, as well as the results of some preliminary experiments conducted in order to modify and improve the metric. No formal evaluation has been performed so far because we are in our first stages, but for the sake of comparison we report some results obtained when comparing our current metric performance with IBM's BLEU.

1 Introduction

Evaluation of MT systems is crucial in their development and improvement. However, human evaluation is expensive and complex. As a consequence, in the last decades several automatic metrics have been developed in order to assess MT output in a simple and less expensive way. From these automatic metrics, the string-based IBM's BLEU (Papineni et al. 2002) is one of the most popular and widely-spread because it is fast and easy to use. However, researchers such as Callison-Burch et al. (2006) and Lavie and Dekowski (2009) have criticized its performance and highlighted its weaknesses in relation to translation quality and its tendency to favour statistically-based MT systems. As a consequence, in response to BLEU weaknesses several linguistically-motivated metrics have arisen. Some of them are based on lexical information, such as METEOR (Banerjee and Lavie 2005); others rely on the use of syntax, either using constituent (Liu and Hildea 2005) or dependency analysis (Owczarzack et al. 2007a and 2007b; He et al. 2010); and others use semantic information, such as Named Entities and semantic roles (Giménez and Márquez 2007 and 2008a). All these metrics work at a certain linguistic level, but little research (Giménez 2008b; Specia and Giménez 2010) has been focused on the use and combination of a wide variety of linguistic information. Therefore, our proposal is a linguistically-motivated metric which aims at using and combining varied linguistic knowledge at different levels in order to cover the key features that must be considered when dealing with MT evaluation from a linguistic point of view. Our hypothesis is that the use and combination of linguistic features at different levels will help us to provide a wider and more accurate coverage than those metrics working at a specific linguistic level.

This paper describes the first stages in the design and the on-going development of the VERTa metric. We provide a description of the modules developed up to now and we report the results obtained in some preliminary experiments

which will help us to see whether we are in the right direction and to discuss the use of certain linguistic knowledge used for the time being. Finally, we draw some conclusions and point out some items which must be under consideration for further development and improvement.

2 Methodology

When approaching MT evaluation from a linguistic point of view, there are different linguistic phenomena which should be taken into account. This can help in the design of our metric and can play an important role when evaluating MT output. In order to define such phenomena, we have considered linguistic issues that we had come across during some work on language data analysis carried out. After such study, we concluded that these phenomena could be classified into lexical, phrase and clause level and that they affected both syntax and semantics. Therefore, the linguistic knowledge that we intend to use is organised in different layers:

- **Lexical information:** We use word-forms and lemmas in order to check lexical units similarity and we also take into account lexical semantic relations such as synonymy, hyperonymy and hyponymy, in other words, semantically-related lexical items.
- **Morphological information:** The information at this level is basically based on lemmas, semantically-related units and the use of Part of Speech tags as the main features in order to cover issues related to inflectional morphology and morphosyntax.
- **Dependency information:** We take into account the dependency relations between the constituents of a sentence. By means of this information we try to solve issues on different word order between the hypothesis and the reference translation. In order to allow a broad coverage, the dependency module is based on the lexical information obtained in the lexical level (see section 2.3)
- **Sentence semantics:** We intend to deal with semantics at sentence level, focusing on semantic arguments.

The use of this varied range of linguistic information allows us to evaluate both adequacy and fluency, thus trying to get closer to human evaluation scores. Given the stage of our work, in this paper we only focus on adequacy for the time being.

In order to combine the above described linguistic features, we have decided to develop one similarity metric per each type of information: lexical similarity metric, morphological similarity metric, dependency similarity and semantic similarity metric respectively. Moreover, we have also added an n-gram similarity module so as to account for similarity between chunks. Each metric works first individually and the final score is the Fmean of the weighted combination of the Precision and Recall of each metric in order to get the results which best correlate with human assessment.

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc) as shown below.

$$P = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(h))}{|\nabla(h)|}$$

$$R = \frac{\sum_{\partial \in D} W_{\partial} * nmatch_{\partial}(\nabla(r))}{|\nabla(r)|}$$

Where r is the reference, h is the hypothesis and ∇ is a function that given a segment will return the elements of each level (e.g. words at lexical level and triples at dependency level). D is the set of different functions to project the level element into the features associated to each level, such as word-form, lemma or partial-lemma at lexical level. $nmatch_{\partial}()$ is a function that returns the number of matches according to the feature ∂ (i.e. the number of lexical matches at the lexical level or the number of dependency triples that match at the dependency level). Finally, W is the set of weights]0 1] associated to each of the different features in a particular level in order to combine the different kinds of matches considered in that level.

Thus far, the metrics implemented are the lexical and morphological similarity metrics, the n-gram similarity metric and part of the dependency metric. As regards the semantic similarity metric, it

has not been explored so far, but we intend to do it in the future. The metric is based on precision and recall and the traditional F-measure is applied in order to get the final score for each pair of segments. In the case of using multiple reference translations, the VERTa metric compares each hypothesis string with the corresponding string of each reference translation and the metric chooses the best score as the final score for that segment.

VERTa works at segment level, comparing the different items of the hypothesis and reference segments from left to right. It must be highlighted that a segment can be composed of one or more sentences. Thus, it could be the case that one segment of the hypothesis contains just one sentence whereas the same segment in the reference has been translated by means of two different sentences, which still belong to the same segment. In order to deal with this issue, segments are split into sentences and the linguistic tools (see sections 2.2 and 2.3 for further details) used in each stage are applied to each sentence separately. Afterwards the metric is applied at segment level; that is to say, we look for the similarity of all items inside the hypothesis segment in relation to all items in the reference segment, regardless of the number of sentences in each segment.

We describe each module in detail in the following sections.

2.1 Lexical Similarity Module

The lexical similarity metric compares lexical items from the hypothesis segment with those in the reference segment. In order to identify these matches we use the following linguistic features: word-forms, lemmas, synonyms, hyperonyms, hyponyms and partial lemmas (lemmas that share the first 4 letters). The approach followed in this module is inspired by METEOR in the sense that the metric relies on lexical items and lexical semantic relations. However, while the most recent version of METEOR (Denkowsi & Lavie, 2011) deals with semantics by means of synonymy and paraphrase tables, our metric uses not only synonymy but tries to exploit other lexical semantic relations such as hyperonymy and hyponymy and avoids the use of paraphrase tables which have to be built up for each language and domain. Moreover, we also use the information provided by lemmas, whereas METEOR relies on stemming. In addition, we also apply a system of weights (W) on the different matches established depending on their importance in terms of semantics, whereas METEOR considers all matches equal, regardless of their difference in terms of meaning.

From the linguistic features that we use, lemmas are obtained by means of WordNet (Feullbaum, 1998). Also the metric relies on some lexical semantic relations such as synonymy, direct hiperonymy and direct hyponymy. These semantic relations are also identified using Wordnet 3.0; however, in order to establish semantic relations we do not use any disambiguation tool, we rely directly on lemmas. As mentioned later in the Experiments section, we thought the use of hyperonymy and hyponymy was a useful strategy to gain more lexical coverage. First we tried to use different levels of hyperonymy and hyponymy but we realised that they introduced noise in the metric, so we decided to restrict their use at immediate levels. However, as shown later, the use of such semantic relations must be reconsidered as they do not always help.

Once established the different linguistic features used by the lexical similarity metric we focus, now, on its mechanism. The metric finds matches between the hypothesis and the reference segment by using the linguistic features explained above in the order established in Table 1.

	W	Match	Examples	
			HYP	REF
1	1	Word-forms	*east*	*east*
2	1	Synonyms	*believed*	*considered*
3	.9	Direct-hypern.	*barrel*	*keg*
4	.9	Direct-hypon.	*keg*	*barrel*
5	.8	Lemma	*is_BE*	*are_BE*
6	.7	Partial-lemma	*danger*	*dangerous*

Table 1. Lexical matches and examples

2.2 Morphological Similarity Module

The morphological similarity metric combines lexical and morphological information. This metric is based on the matches set in the lexical similarity metric, except for the partial-match, in combination with the Part of Speech (POS) tags

from the annotated corpus[1]. By means of this combination, we apply a restriction in terms of fluency because we avoid issues such as stating that *invites* and *invite* are positive matches regarding morphology, and somehow we compensate the broader coverage that we have in the lexical module. Therefore, when assessing MT output in terms of fluency this metric will receive a higher weight, whereas when evaluating adequacy, the weight given to this module will be reduced. This module will be particularly useful when evaluating MT output of languages with a rich inflectional morphology, such as Spanish or Catalan.

Following the approach used in the lexical similarity metric, the morphological similarity metric establishes matches between items in the hypothesis and the reference sentence and a set of weights (W) is applied. However, instead of comparing single lexical items as in the previous module, in this module we compare pairs of features in the order established in Table 2.

	W	Match	Examples	
			HYP	REF
1	1	(Word-form, POS)	(he, PRP)	(he, PRP)
2	1	(Synonym, POS)	(VIEW, NNS)	(OPINON, NNS)
3	.9	(Hypern., POS)	(PUBLICATION, NN)	(MAGAZINE, NN)
4	.9	(Hypon., POS)	(MAGAZINE, NN)	(PUBLICATION, NN)
5	.8	(LEMMA, POS)	can_(CAN, MD)	Could_(CAN, MD)

Table 2. Morphological pairs of matches and examples.

2.3 Dependency Similarity Module – Work in progress

Once covered the lexical and morphological sections, we are now working on the dependency similarity metric which will help us to deal with syntactic structures at a deeper level. By means of this module we will be able to capture the relations between sentence constituents regardless of their position inside the sentence, which will be really helpful when comparing a hypothesis and a reference segment with a different word order of their constituents, as illustrated in the following example:

Example 1:
HYP: ***After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman*** *Haniya said....*
REF: *Haniya said,* ***after a meeting on Monday evening with the head of Egyptian Intelligence General Omar Suleiman****...*

In this example, the adjunct realised by the PP *After a meeting Monday night with the head of Egyptian intelligence chief Omar Suleiman* occupies different positions in the hypothesis and reference strings. In the hypothesis it is located at the beginning of the sentence, preceding the subject *Haniya*, whereas in the reference, it is placed after the verb. By means of dependencies, we can state that although located differently inside the sentence both subject and adjunct depend on the verb as shown in Table 3.

HYPOTHESIS	REFERENCE
nsubj(Haniya, said)	nsubj(Haniya, said)
prep_after(meeting, said)	prep_after(meeting, said)

Table 3. Matching of triples

Therefore, the use of dependencies helps us to establish similarities between equivalent sentences which contain the same constituents but in different positions.

This dependency similarity metric works at sentence level and follows the approach used by Owczarzack et al. (2007a and 2007b) and He et al. (2010) with some linguistic additions in order to adapt it to our metric combination.

Both hypothesis and reference strings are annotated with dependency relations by means of the Stanford parser (de Marneffe et al. 2006). The reason why this parser is used is because after conducting an evaluation (Comelles et al. 2010) where the performance of several dependency parsers was assessed (Stanford, DeSR, MALT, Minipar, RASP) this proved to be the best in terms of linguistic quality. Moreover, the output file provided by this parser contains dependency relations by means of flat triples with the form **Label(Head, Mod)**. These triples are ideal in order

[1] The corpus has been annotated with POS tags using the Stanford Parser (de Marneffe et al. 2006).

to compare the dependency relations in the hypothesis and reference segments.

The dependency similarity metric also relies first on the matches established at lexical level − word-form, synonymy, hyperonymy, hyponymy and lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form. Then, and inspired by He et al. (2010) and Owczarzak et al. (2007a and 2007b), four different types of dependency matches have been designed. Next, we describe the matches and provide examples for each of them:

- Complete (MC): Type of match used when the triples are identical, this means that the label, the head and the modifier match.

  Label1(Head1,Mod1) = Label1(Head2,Mod2)
  Example 2:
  HYP: advmod(difficult, more)
  REF: advmod (difficult, more)

- Partial (MP): Three different types of partial matches are established:
  - Partial_no_mod (MP_no_mod): The label and the head match but the modifier does not match
    - Label1 = Label2
    - Head1 = Head2

    Example 3:
    HYP:**conj_and(difficult**, dangerous)
    REF: **conj_and(difficult**, serious)

  - Partial_no_head (MP_no_head): The label and the modifier match but the head does not match.
    - Label1 = Label2
    - Mod1 = Mod2

    Example 4:
    HYP: prep_between(mentioned, Lebanon)
    REF: prep_between(crisis, Lebanon)

  - Partial_no_label (MP_no_label): The head and the modifier match but the label does not match.
    - Head1 = Head2
    - Mod1 = Mod2

    Example 5:
    HYP: predet(**parties**, **all**)
    REF: det(**parties**,**all**)

Each type of match is given a weight which ranges from the highest to the lowest weight in the following order:

- Complete (1)
- Partial_no_mod (.8)
- Partial_no_head (.7)
- Partial_no_label (.7)

In addition, we have also planned to add some extra-rules in order to capture the similarity between certain structures which are semantically equal but syntactically different. These extra-rules will be applied at phrase and sentence level. An example of these rules at phrase level affects modifiers inside the noun phrase and the latter the passive-active voice alternation. We plan to cover the similarity between an adjective premodifiying a noun and an of-prepositional phrase postmodifying it, as exemplified below.

Example 6:
HYP: ...*between the* ***ministries of interior***...
REF: ...*between the two* ***interior ministries***...

HYP_prep_of(ministries, interior) = REF_amod(ministries, interior)

Although their labels differ, this couple of triples must be considered as an exact match due to their semantic similarity. Otherwise we would penalise a couple of structures which are equal from a semantic point of view. At a clause level, an example of these rules could be the treatment of the active-passive alternation. As shown below, although syntactically different, both structures share the same meaning.

Example 7:
HYP: *After meeting* ***the Moroccan news agency published*** *a joint statement*...
REF: *A joint statement* ***published*** *(...)* ***by the Moroccan news agency***...
HYP_nsubj(published, agency) = REF_agent(published, agency)

Similar to the pair of dependencies dealing with modifiers, *nsubj* and *agent* labels must be considered identical and thus, the previous couple of triples must be scored as an exact match.

Unfortunately, this set of rules has not been implemented yet in the dependency metric. Therefore, results shown in the Experiments section only refer to the use of the different matches.

2.4 N-gram Similarity Module

The n-gram similarity module is aimed at matching chunks[2] in the hypothesis and reference segments. Chunks length goes from bigrams to sentence length. The use of this module allows us to combine both linguistic and statistical approaches and enables us to deal with word order inside the sentence by means of a more simple approach than the parsing of constituents. The n-gram similarity module uses the matches obtained at lexical level in order to align chunks. Thus, we do not only match n-grams relying on the word-form but also taking into account synonymy, hyponymy/hyperonymy and lemmas, as shown in example 8, where the chunks [*the situation in the area*] and [*the situation in the region*] match, although *area* and *region* do not share the same word-form but a relation of synonymy.

Example 8:
HYP: … the situation in the *area*…
REF: … the situation in the *region*…

2.5 Metrics Combination

As mentioned at the beginning of the section, the modules implemented so far are combined in order to cover linguistic features at all levels depending on the type of evaluation. Therefore, if the evaluation is focused on adequacy, those modules more related to semantics will have a higher weight, whereas if evaluating fluency those related to morphology, morphosyntax and constituent word order will be more important. Moreover, metrics should also be combined depending on the type of language evaluated. If a language with a rich inflectional morphology such as Spanish is assessed, the morphology module should be given a higher weight; whereas if the language evaluated does not show such a rich inflectional morphology (i.e. English) the weight of the morphology module should be lower. As a consequence, a set of weights has been established which can be changed manually regarding the type of evaluation. So far weights have been set according to the linguistic characteristics of the language under analysis and the type of evaluation. In a near future we intend to work on the tuning of weights in order to improve the metric performance. The experiments described in the next section are all focused on evaluating adequacy, as a consequence, the lexical and dependency metrics receive higher weights than the morphology and n-gram similarity metrics. For these experiments weights have been set as follows:

- Lexical Module: 0.444
- Morphology Module: 0.111
- N-gram Module: 0.111
- Dependency Module: 0.333

3. Experiments

In this section we report a couple of preliminary experiments at segment and system level to check whether we were in the right direction. These experiments should not be regarded as a formal evaluation, but just as a set of preliminary tests which should give us information on the adequacy of the linguistic features used. They must provide us with material to discuss, reconsider and improve the on-going development of the metric. The experiments were aimed at checking (i) the influence of adding the dependency module and (ii) the influence of hyperonyms and hyponyms. For these experiments we used data provided in the MetricsMaTr 2010 shared-task[3]. From the data provided by the organization we used 100 segments of the NIST Open-MT06 data, the MT output from 8 different MT systems (a total of 28,000 words approximately) and 4 reference translations. The human judgments used were based on adequacy. In order to calculate correlations at segment level we used Pearson correlation and we took into account all segments regardless of the system providing them in order to have a more precise correlation. Table 4 shows the results obtained.

[2] By chunks we understand a group of words that go together, one next to the other, not necessarily working as a constituent

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

	NO DEP + HYP	DEP + HYP	DEP + NO HYP.
Pearson Correlation	0.734	0.755	0.759

Table 4. Pearson correlations at segment level

On the one hand, the use of the partially-implemented dependency module improved the performance of the metric. Thus, adding linguistic knowledge which deals with deep structure at clause and phrase level helped to account for certain relationships which would not be considered by means of the n-gram matching module, such as different word order of the constituents inside the sentence. On the other hand, and opposed to our hypothesis, at segment level, the metric correlates better with human judgments when lexical semantic relations are more restricted. It seems therefore that the use of direct hyperonyms and hyponyms does not help to improve the metric performance; on the contrary, it slightly degrades the correlation with human judgments. There might be a couple of reasons for this result: first, a low percentage of hyponyms and hyperonyms in the reference translations; secondly, the fact of not using any process of disambiguation might make the metric match certain words which, although being hyponyms or hyperonyms, do not share such a relationship in the domain under analysis.

For the sake of comparison and just to check that our first steps were in the right direction, we were also interested in comparing our metric with the widely-used metric BLEU. As shown in Table 5 the results obtained by our metric at system level, although being yet in its first stages, outperforms the results obtained by IBM's BLEU at both system and segment level, due to the use of more lexical semantic information by our metric and the calculation of recall.

Metric	Pearson Correlation	
	Segment	System
VERTa	**0.759**	**0.970**
BLEU	0.683	0.931

Table 5. Metric comparison at segment and system level

4. Conclusions and Future Work

In this paper we have describe the work in progress of the metric we are developing. We have described the modules of the metric which have been designed and implemented so far and we reported the results obtained in some preliminary experiments. The scores obtained in the correlations with human judgments show that the use of linguistic information dealing with different types of linguistic phenomena and at different levels helps in improving the metric performance. Although they are preliminary results, they will be extremely helpful to continue with our on-going research. Moreover, the figures obtained by our primary metric implementation when compared to BLEU show promising results for the combination and use of a wide variety of linguistic features.

In a near future, we plan to keep working on the development of the metric by exploring the use of other linguistic information (i.e. multi-words treatment, the importance of function and content words and the use of semantic information at sentence level). In addition, we also expect to improve the metric performance by finishing the implementation of the dependency module (i.e. refining the type of dependency labels and matches to take into account, and implementing the set of similarity rules) and continue working on the tuning of the weights used both inside the modules and in metrics combination. Regarding the meta-evaluation of the metric, we will analyze the coverage of each level separately and we will evaluate our metric not only in terms of adequacy but also in terms of fluency. Finally, we would also like to test the robustness of VERTa with other languages with richer inflectional morphology such as Spanish.

Acknowledgments

We are very grateful to LDC for kindly providing the development data used in the MetricsMaTr 2010 shared-task.

This work has been partially funded by the Spanish Government (projects KNOW2, TIN-2009-14715-C04-03, and Holopedia, TIN2010-21128-C02-02).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments in *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, pages 65-72, Ann Arbor, Michigan.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research in *Proceedings of the EACL 2006*, pages 249–256.

Elisabet Comelles, Victoria Arranz and Irene Castellon. 2010. Constituency and Dependency Parsers Evaluation. SEPLN (ed.), *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. 45, pages 59-66. SEPLN. Valencia. ISSN: 1135-5948

Michael J. Denkowski and Alon Lavie. 2011. METEOR 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems in *Proceedings of the 6th Workshop on Statistical Machine Translation (ACL-2011)*, pages 85–91, Edinburgh, Scotland, UK.

Alon Lavie and Michael J. Denkowski. 2009. The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation, 23*.

Christian Fellbaum (Ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogeneous MT systems in *Proceedings of the 2nd Workshop on Statistical Machine Translation (ACL)*, pages 256-264, Prague, Czech Repubilc.

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation in *Proceedings of the 3rd Workshop on Statistical Machine Translation (ACL)*, pages 195-198, Columbus. OH.

Jesús Gimenez. 2008. Empirical Machine Translation and its Evaluation. Doctoral Dissertation. UPC.

Yifan He, Jinhua Du, Andy Way and Josef van Genabith. 2010. The DCU Dependency-based Metric in WMT-Metrics MATR 2010. *In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 349-353, Uppsala, Sweden.

Ding Liu and Daniel Hildea. 2005. Syntactic Features for Evaluation of Machine Translation in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32, Ann Arbor

Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses in *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation (LREC-2006)*. Genoa, Italy.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. in *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*, Manchester, UK.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation in *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, Rochester, New York.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Labelled Dependencies in Machine Translation Evaluation in *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 104– 111, Prague, Czech Republic.

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 of the Association for Computational Linguistics (ACL02), pages 311-318. Philadelphia. PA.

Lucia Specia and Jesús Giménez. 2010. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In the *Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.