TINE: A Metric to Assess MT Adequacy

Miguel Rios, Wilker Aziz and **Lucia Specia**
Research Group in Computational Linguistics
University of Wolverhampton
Stafford Street, Wolverhampton, WV1 1SB, UK
{m.rios, w.aziz, l.specia}@wlv.ac.uk

Abstract

We describe TINE, a new automatic evaluation metric for Machine Translation that aims at assessing segment-level adequacy. Lexical similarity and shallow-semantics are used as indicators of adequacy between machine and reference translations. The metric is based on the combination of a lexical matching component and an adequacy component. Lexical matching is performed comparing bags-of-words without any linguistic annotation. The adequacy component consists in: i) using ontologies to align predicates (verbs), ii) using semantic roles to align predicate arguments (core arguments and modifiers), and iii) matching predicate arguments using distributional semantics. TINE's performance is comparable to that of previous metrics at segment level for several language pairs, with average Kendall's tau correlation from 0.26 to 0.29. We show that the addition of the shallow-semantic component improves the performance of simple lexical matching strategies and metrics such as BLEU.

1 Introduction

The automatic evaluation of Machine Translation (MT) is a long-standing problem. A number of metrics have been proposed in the last two decades, mostly measuring some form of matching between the MT output (hypothesis) and one or more human (reference) translations. However, most of these metrics focus on fluency aspects, as opposed to adequacy. Therefore, measuring whether the meaning of the hypothesis and reference translation are the same or similar is still an understudied problem.

The most commonly used metrics, BLEU (Papineni et al., 2002) and alike, perform simple exact matching of n-grams between hypothesis and reference translations. Such a simple matching procedure has well known limitations, including that the matching of non-content words counts as much as the matching of content words, that variations of words with the same meaning are disregarded, and that a perfect matching can happen even if the order of sequences of n-grams in the hypothesis and reference translation are very different, changing completely the meaning of the translation.

A number of other metrics have been proposed to address these limitations, for example, by allowing for the matching of synonyms or paraphrases of content words, such as in METEOR (Denkowski and Lavie, 2010). Other attempts have been made to capture whether the reference translation and hypothesis translations share the same meaning using shallow semantics, i.e., Semantic Role Labeling (Giménez and Márquez, 2007). However, these are limited to the exact matching of semantic roles and their fillers.

We propose TINE, a new metric that complements lexical matching with a shallow semantic component to better address adequacy. The main contribution of such a metric is to provide a more flexible way of measuring the overlap between shallow semantic representations that considers both the semantic structure of the sentence and the content of the semantic elements. The metric uses SRLs such as in (Giménez and Márquez, 2007). However, it analyses the content of predicates and arguments seeking for either exact or "similar" matches. The

inexact matching is based on the use of ontologies such as VerbNet (Schuler, 2006) and distributional semantics similarity metrics, such as Dekang Lin's thesaurus (Lin, 1998) .

In the remainder of this paper we describe some related work (Section 2), present our metric - TINE - (Section 3) and its performance compared to previous work (Section 4) as well as some further improvements. We then provide an analysis of these results and discuss the limitations of the metric (Section 5) and present conclusions and future work (Section 6).

2 Related Work

A few metrics have been proposed in recent years to address the problem of measuring whether a hypothesis and a reference translation share the same meaning. The most well-know metric is probably METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2010). METEOR is based on a generalized concept of unigram matching between the hypothesis and the reference translation. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases. However, the structure of the sentences is not considered.

Wong and Kit (2010) measure word choice and word order by the matching of words based on surface forms, stems, senses and semantic similarity. The informativeness of matched and unmatched words is also weighted.

Liu et al. (2010) propose to match bags of unigrams, bigrams and trigrams considering both recall and precision and F-measure giving more importance to recall, but also using WordNet synonyms.

Tratz and Hovy (2008) use transformations in order to match short syntactic units defined as Basic Elements (BE). The BE are minimal-length syntactically well defined units. For example, nouns, verbs, adjectives and adverbs can be considered BE-Unigrams, while a BE-Bigram could be formed from a syntactic relation (e.g. subject+verb, verb+object). BEs can be lexically different, but semantically similar.

Padó et al. (2009) uses Textual Entailment features extracted from the Standford Entailment Recognizer (MacCartney et al., 2006). The Textual Entailment Recognizer computes matching and mismatching features over dependency parses. The metric then predicts the MT quality with a regression model. The alignment is improved using ontologies.

He et al. (2010) measure the similarity between hypothesis and reference translation in terms of the Lexical Functional Grammar (LFG) representation. The representation uses dependency graphs to generate unordered sets of dependency triples. Calculating precision, recall, and F-score on the sets of triples corresponding to the hypothesis and reference segments allows measuring similarity at the lexical and syntactic levels. The measure also matches WordNet synonyms.

The closest related metric to the one proposed in this paper is that by Giménez and Márquez (2007) and Giménez et al. (2010), which also uses shallow semantic representations. Such a metric combines a number of components, including lexical matching metrics like BLEU and METEOR, as well as components that compute the matching of constituent and dependency parses, named entities, discourse representations and semantic roles. However, the semantic role matching is based on exact matching of roles and role fillers. Moreover, it is not clear what the contribution of this specific information is for the overall performance of the metric.

We propose a metric that uses a lexical similarity component and a semantic component in order to deal with both word choice and semantic structure. The semantic component is based on semantic roles, but instead of simply matching the surface forms (i.e. arguments and predicates) it is able to match similar words.

3 Metric Description

The rationale behind TINE is that an adequacy-oriented metric should go beyond measuring the matching of lexical items to incorporate information about the semantic structure of the sentence, as in (Giménez et al., 2010). However, the metric should also be flexible to consider inexact matches of semantic components, similar to what is done with lexical metrics like METEOR (Denkowski and Lavie, 2010). We experiment with TINE having English as target language because of the availability of linguistic processing tools for this language. The metric is particularly dependent on semantic role label-

ing systems, which have reached satisfactory performance for English (Carreras and Márquez, 2005). TINE uses semantic role labels (SRL) and lexical semantics to fulfill two requirements by: (i) compare both the semantic structure and its content across matching arguments in the hypothesis and reference translations; and (ii) propose alternative ways of measuring inexact matches for both predicates and role fillers. Additionally, it uses an exact lexical matching component to reward hypotheses that present the same lexical choices as the reference translation. The overall score s is defined using the simple weighted average model in Equation (1):

$$s(H, \mathbf{R}) = max \left\{ \frac{\alpha L(H,R) + \beta A(H,R)}{\alpha + \beta} \right\}_{R \in \mathbf{R}} \quad (1)$$

where H represents the hypothesis translation, R represents a reference translation contained in the set of available references $\mathbf{R}$; L defines the (exact) lexical match component in Equation (2), *A* defines the adequacy component in Equation (3); and α and β are tunable weights for these two components. If multiple references are provided, the score of the segment is the maximum score achieved by comparing the segment to each available reference.

$$L(H,R) = \frac{|H \bigcap R|}{\sqrt{|H| * |R|}} \quad (2)$$

The lexical match component measures the overlap between the two representations in terms of the cosine similarity metric. A segment, either a hypothesis or a reference, is represented as a bag of tokens extracted from an unstructured representation, that is, bag of unigrams (words or stems). Cosine similarity was chosen, as opposed to simply checking the percentage of overlapping words (POW) because cosine does not penalize differences in the length of the hypothesis and reference translation as much as POW. Cosine similarity normalizes the cardinality of the intersection $|H \cap R|$ using the geometric mean $\sqrt{|H| * |R|}$ instead of the union $|H \cup R|$. This is particularly important for the matching of arguments - which is also based on cosine similarity. If an hypothesized argument has the same meaning as its reference translation, but differs from it in length, cosine will penalize less the matching than POW. That is specially interesting when core arguments get merged with modifiers due to bad semantic role labeling (e.g. *[A0 I] [T bought] [A1 something to eat yesterday]* instead of *[A0 I] [T bought] [A1 something to eat] [AM-TMP yesterday]*).

$$A(H,R) = \frac{\sum_{v \in V} verb\_score(H_v, R_v)}{|V_r|} \quad (3)$$

In the adequacy component, *V* is the set of verbs aligned between *H* and *R*, and $|V_r|$ is the number of verbs in *R*. Hereafter the indexes h and r stand for hypothesis and reference translations, respectively. Verbs are aligned using VerbNet (Schuler, 2006) and VerbOcean (Chklovski and Pantel, 2004). A verb in the hypothesis v_h is aligned to a verb in the reference v_r if they are related according to the following heuristics: (i) the pair of verbs share at least one class in VerbNet; or (ii) the pair of verbs holds a relation in VerbOcean.

For example, in VerbNet the verbs *spook* and *terrify* share the same class *amuse-31.1*, and in VerbOcean the verb *dress* is related to the verb *wear*.

$$verb\_score(H_v, R_v) = \frac{\sum_{a \in A_r \cap A_t} arg\_score(H_a, R_a)}{|A_r|} \quad (4)$$

The similarity between the arguments of a verb pair (v_h, v_r) in *V* is measured as defined in Equation (4), where A_h and A_t are the sets of labeled arguments of the hypothesis and the reference respectively and $|A_r|$ is the number of arguments of the verb in R. In other words, we only measure the similarity of arguments in a pair of sentences that are annotated with the same role. This ensures that the structure of the sentence is taken into account (for example, an argument in the role of *agent* would not be compared against an argument in a role of *experiencer*). Additionally, by restricting the comparison to arguments of a given verb pair, we avoid argument confusion in sentences with multiple verbs.

The *arg_score*(H_a, R_a) computation is based on the cosine similarity as in Equation (2). We treat the tokens in the argument as a bag-of-words. However, in this case we change the representation of the segments. If the two sets do not match exactly, we expand both of them by adding similar words. For every mismatch in a segment, we retrieve the

20-most similar words from Dekang Lin's distributional thesaurus (Lin, 1998), resulting in sets with richer lexical variety.

The following example shows how the computation of $A(H, R)$ is performed, considering the following hypothesis and reference translations:

H: The lack of snow discourages people from ordering ski stays in hotels and boarding houses.

R: The lack of snow is putting people off booking ski holidays in hotels and guest houses.

1. extract verbs from H: V_h = {discourages, ordering}
2. extract verbs from R: V_r = {putting, booking}
3. similar verbs aligned with VerbNet (shared class get-13.5.1): V = $\{(v_h = order, v_r = book)\}$
4. compare arguments of ($v_h = order$,$v_r = book$):
   A_h = {A0, A1, AM-LOC}
   A_r = {A0, A1, AM-LOC}
5. $A_h \cap A_r$ = {A0, A1, AM-LOC}
6. exact matches:
   H_{A0} = {people} and R_{A0} = {people}
   argument_score = 1
7. different word forms: expand the representation:
   H_{A1} = {ski, stays} and R_{A1} = {ski, holidays}
   expand to:
   H_{A1} = {{ski},{stays, remain... journey...}}
   R_{A1} = {{ski},{holidays, vacations, trips... journey...}}
   argument_score = 0.5
8. similarly to H_{AM-LOC} and R_{AM-LOC}
   argument_score = 0.72
9. verb_score (*order*, *book*) = $\frac{1+0.5+0.72}{3}$ = 0.74
10. $A(H, R) = \frac{0.74}{2}$ = 0.37

Different from previous work, we have not used WordNet to measure lexical similarity for two main reasons: problems with lexical ambiguity and limited coverage in WordNet (instances of named entities are not in WordNet, e.g. *Barack Obama*). For example, in WordNet the aligned verbs (*order/book*) from the previous hypothesis and reference translations have: 9 senses - *order* (e.g. give instructions to or direct somebody to do something with authority, make a request for something, etc.) - and 4 senses - *book* (engage for a performance, arrange for and reserve (something for someone else) in advance, etc.). Thus, a WordNet-based similarity measure would require disambiguating segments, an additional step and a possible source of errors. Second, a thresholds would need to be set to determine when a pair of verbs is aligned. In contrast, the structure of VerbNet (i.e. clusters of verbs) allows a binary decision, although the VerbNet heuristic results in some errors, as we discuss in Section 5.

4 Results

We set the weights α and β by experimental testing to $\alpha = 1$ and $\beta = 0.25$. The lexical component weight is prioritized because it has shown a good average Kendall's tau correlation (0.23) on a development dataset (Callison-Burch et al., 2010). Table 1 shows the correlation of the lexical component with human judgments for a number of language pairs.

Table 1: Kendall's tau segment-level correlation of the lexical component with human judgments

Metric	cz-en	fr-en	de-en	es-en	avg
Lexical	0.27	0.21	0.26	0.19	0.23

We use the SENNA[1] SRL system to tag the dataset with semantic roles. SENNA has shown to have achieved an F-measure of 75.79% for tagging semantic roles over the CoNLL 2005 [2] benchmark.

We compare our metric against standard BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2010) and other previous metrics reported in (Callison-Burch et al., 2010) which also claim to use some form of semantic information (see Section 2 for their description). The comparison is made in terms of Kendall's tau correlation against the human judgments at a segment-level. For our submission to the shared evaluation task, system-level scores are obtained by averaging the segment-level scores.

TINE achieves the same average correlation with BLUE, but outperforms it for some language pairs. Additionally, TINE outperforms some of the previous which use WordNet to deal with synonyms as part of the lexical matching.

The closest metric to TINE (Giménez et al., 2010), which also uses semantic roles as one of its

[1] http://ml.nec-labs.com/senna/

[2] http://www.lsi.upc.edu/ srlconll/

Table 2: Comparison with previous semantically-oriented metrics using segment-level Kendall's tau correlation with human judgments

Metric	cz-en	fr-en	de-en	es-en	avg
(Liu et al., 2010)	0.34	0.34	0.38	0.34	0.35
(Giménez et al., 2010)	0.34	0.33	0.34	0.33	0.33
(Wong and Kit, 2010)	0.33	0.27	0.37	0.32	0.32
METEOR	0.33	0.27	0.36	0.33	0.32
TINE	0.28	0.25	0.30	0.22	0.26
BLEU	0.26	0.22	0.27	0.28	0.26
(He et al., 2010)	0.15	0.14	0.17	0.21	0.17
(Tratz and Hovy, 2008)	0.05	0.0	0.12	0.05	0.05

components, achieves better performance. However, this metric is a rather complex combination of a number of other metrics to deal with different linguistic phenomena.

4.1 Further Improvements

As an additional experiment, we use BLEU as the lexical component $L(H, R)$ in order to test if the shallow-semantic component can contribute to the performance of this standard evaluation metric. Table 3 shows the results of the combination of BLEU and the shallow-semantic component using the same parameter configuration as in Section 4. The addition of the shallow-semantic component increased the average correlation of BLEU from 0.26 to 0.28.

Table 3: TINE-B: Combination of BLEU and the shallow-semantic component

Metric	cz-en	fr-en	de-en	es-en	avg
TINE-B	0.27	0.25	0.30	0.30	0.28

Finally, we improve the tuning of the weights of the components (α and β parameters) by using a simple genetic algorithm (Back et al., 1999) to select the weights that maximize the correlation with human scores on a development set (we use the development sets from WMT10 (Callison-Burch et al., 2010)). The configuration of the genetic algorithm is as follows:

- Fitness function: Kendall's tau correlation
- Chromosome: two real numbers, α and β
- Number of individuals: 80
- Number of generations: 100
- Selection method: roulette
- Crossover probability: 0.9
- Mutation probability: 0.01

Table 4 shows the parameter values obtaining from tuning for each language pair and the correlation achieved by the metric with such parameters. With such an optimization step the average correlation of the metric increases to 0.29.

Table 4: Optimized values of the parameters using a genetic algorithm and Kendall's tau and final correlation of the metric on the test sets

Language pair	Correlation	α	β
cz-en	0.28	0.62	0.02
fr-en	0.25	0.91	0.03
de-en	0.30	0.72	0.1
es-en	0.31	0.57	0.02
avg	0.29	–	–

5 Discussion

In what follows we discuss with a few examples some of the common errors made by TINE. Overall, we consider the following categories of errors:

1. Lack of coverage of the ontologies.
   R: This year, women were awarded the Nobel Prize in all fields except physics
   H: This year the women received the Nobel prizes in all categories less physical
   The lack of coverage in VerbNet prevented the detection of the similarity between *receive* and *award*.

2. Matching of unrelated verbs.
   R: If snow falls on the slopes this week, Christmas will sell out too, says Schiefert.
   H: If the roads remain snowfall during the week, the dates of Christmas will dry up, said Schiefert.
   In VerbOcean *remain* and *say* are incorrectly

said to be related. VerbOcean was created by a semi-automatic extraction algorithm (Chklovski and Pantel, 2004) with an average accuracy of 65.5%.

3. Incorrect tagging of the semantic roles by SENNA.

   R: Colder weather is forecast for Thursday, so if anything falls, it should be snow.

   H: On Thursday , must fall temperatures and, if there is rain, in the mountains should.

   The position of the predicates affects the SRL tagging. The predicate *fall* has the following roles (A1, V, and S-A1) in the reference, and the following roles (AM-ADV, A0, AM-MOD, and AM-DIS) in the hypothesis. As a consequence, the metric cannot attempt to match the fillers. Also, SRL systems do not detect phrasal verbs such as in the example of Section 3, where the action *putting people off* is similar to *discourages*.

6 Conclusions and Future Work

We have presented an MT evaluation metric based on the alignment of semantic roles and flexible matching of role fillers between hypothesis and reference translations. To deal with inexact matches, the metric uses ontologies and distributional semantics, as opposed to lexical databases like WordNet, in order to minimize ambiguity and lack of coverage. The metric also uses an exact lexical matching component to reward hypotheses that present lexical choices similar to those of the reference translation.

Given the simplicity of the metric, it has achieved competitive results. We have shown that the addition of the shallow-semantic component into a lexical component yields absolute improvements in the correlation of 3%-6% on average, depending on the lexical component used (cosine similarity or BLEU).

In future work, in order to improve the performance of the metric we plan to add components to address a few other linguistic phenomena such as in (Giménez and Márquez, 2007; Giménez et al., 2010). In order to deal with the coverage problem of an ontology, we plan to use distributional semantics (i.e. word space models) also to align the predicates. We consider using a backoff model for the shallow-semantic component to deal with the very frequent cases where there are no comparable predicates between the reference and hypothesis translations, which result in a 0 score from the semantic component. Finally, we plan to improve the lexical component to better tackle fluency, for example, by adding information about the word order.

References

Thomas Back, David B. Fogel, and Zbigniew Michalewicz, editors. 1999. *Evolutionary Computation 1, Basic Algorithms and Operators*. IOP Publishing Ltd., Bristol, UK, 1st edition.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. 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, Uppsala, Sweden, July.

Xavier Carreras and Lluís Márquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the 9th Conference on Natural Language Learning, CoNLL-2005*, Ann Arbor, MI USA.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July.

Michael Denkowski and Alon Lavie. 2010. 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.

Jesús Giménez and Lluís Márquez. 2007. Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 256–264, Stroudsburg, PA, USA.

Jesús Giménez, Lluís Márquez, Elisabet Comelles, Irene Castellón, and Victoria Arranz. 2010. Document-level automatic mt evaluation based on discourse representations. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 333–338, Stroudsburg, PA, USA.

Yifan He, Jinhua Du, Andy Way, and Josef van Genabith. 2010. The dcu dependency-based metric in wmt-metricsmatr 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 349–353, Stroudsburg, PA, USA.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, pages 768–774, Stroudsburg, PA, USA.

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla: translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 354–359, Stroudsburg, PA, USA.

Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 41–48, New York City, USA, June.

Sebastian Padó, Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23:181–193, September.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Stephen Tratz and Eduard Hovy. 2008. Summarisation evaluation using transformed basic elements. In *Proceedings TAC 2008*.

Billy T.-M. Wong and Chunyu Kit. 2010. The parameter-optimized atec metric for mt evaluation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 360–364, Stroudsburg, PA, USA.