Learning the Impact of Machine Translation Evaluation Metrics for Semantic Textual Similarity

Simone Magnolini
University of Brescia,
Fondazione Bruno Kessler
Trento, Italy
magnolini@fbk.eu

Ngoc Phuoc An Vo
University of Trento,
Fondazione Bruno Kessler
Trento, Italy
ngoc@fbk.eu

Octavian Popescu
IBM Research, T.J. Watson
Yorktown, US
o.popescu@us.ibm.com

Abstract

We present a work to evaluate the hypothesis that automatic evaluation metrics developed for Machine Translation (MT) systems have significant impact on predicting semantic similarity scores in Semantic Textual Similarity (STS) task for English, in light of their usage for paraphrase identification. We show that different metrics may have different behaviors and significance along the semantic scale [0-5] of the STS task. In addition, we compare several classification algorithms using a combination of different MT metrics to build an STS system; consequently, we show that although this approach obtains state of the art result in paraphrase identification task, it is insufficient to achieve the same result in STS.

1 Introduction

Semantic related tasks have become a noticed trend in Natural Language Processing (NLP) community. Particularly, the Semantic Textual Similarity (STS) task has captured a huge attention in the NLP community despite being recently introduced since SemEval 2012 and continuing in SemEval 2013, 2014 and 2015 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015). Basically, the task requires to build systems which can compute the similarity degree between two given sentences. The similarity degree is scaled as a real score from 0 (no relevance) to 5 (semantic equivalence). The evaluation is done by computing the correlation between human judgment scores and systems' predictions by the mean of Pearson correlation method.

In contrast, Machine Translation evaluation metrics are designed to assess if the output of a MT system is semantically equivalent to a set of reference translations. In SemEval 2012, the system made by (de Souza et al., 2012) and then the system (Barrón Cedeño et al., 2013) in SemEval 2013 introduced the approach of using a set of MT evaluation metrics together with other lexical and syntactic features to predict the semantic similarity scores in STS. Although this approach shows promising results, there was no in-depth analysis on the impact of the evaluation metrics to the overall performance and how each metric behaves on STS data. Moreover, as being inspired by the literature (Madnani et al., 2012) for paraphrase recognition, which obtains the state of art result on the Microsoft Research paraphrase corpus (MSRP) (Dolan et al., 2004), we decide to analyze the impact of MT evaluation metrics in STS.

Our aim consists of two folds, (1) to obtain a clear idea of how each individual metric behaves and correlates with the human-judgement semantic similarity, and (2) to examine the approach of combining a set of chosen metrics to build regression models for predicting the semantic similarity scores and analyze the incorporation of these metrics in regarding to the overall performance of the system. To achieve our goal, we divide our research in two main aspects: first, we evaluate the correlation between each single MT metric and the human-annotation scores; and second, we evaluate how different classification algorithms perform using these metrics as features.

The remainder of this paper is organized as follows: Section 2 presents the description of different MT evaluation metrics, Section 3 reports the experimental settings, Section 4 is the evaluation and discussion, and finally, Section 5 is conclusions and future work.

Proceedings of Recent Advances in Natural Language Processing, pages 398–403,
Hissar, Bulgaria, Sep 7–9 2015.

2 Machine Translation Evaluation Metrics

Technically, the MT evaluation metric assesses the semantic equivalence between the translation hypothesis produced by a MT system and the reference translation. In STS task, the idea of using MT evaluation metrics is adopted to improve the word alignment job between two given sentences which consequently leads to better prediction of semantic similarity scores. In this study, we employ four commonly used metrics from two different groups of MT evaluation metrics, (1) the n-gram based metrics (METEOR and BLEU), and (2) the edit-distance based metrics (TER and TERp).

METEOR (Metric for Evaluation of Translation with Explicit ORdering). We use the latest version (1.5) of METEOR (Denkowski and Lavie, 2014) that finds alignments between sentences based on exact, stem, synonym and paraphrase matches between words and phrases. Segment and system level metric scores are calculated based on the alignments between sentence pairs.

BLEU (BiLingual Evaluation Understudy). We use BLEU (Papineni et al., 2002) because it is one of the most commonly used metrics and it has a high reliability. The BLEU metric computes as the amount of n-gram overlap, for different values of n=1,2,3 and 4, between the system output and the reference translation, in our case between sentence pairs. The score is tempered by a penalty for translations that might be too short. BLEU relies on exact matching and has no concept of synonymy or paraphrasing.

TER (Translation Error Rate). We use the 0.7.25 version of TER (Snover et al., 2006). TER computes the number of edits needed to "fix" the translation output so that it matches the reference. TER differs from word error rate (WER) in which it includes a heuristic algorithm to deal with shifts in addition to insertions, deletions and substitutions.

TERp (TER-Plus). The last metrics that we use is TERp (Snover et al., 2009) building upon the core TER algorithm and providing additional edit operations based on stemming, synonymy and paraphrase.

year	dataset	pairs	source
2012	MSRpar	1500	newswire
2012	MSRvid	1500	video descriptions
2012	OnWN	750	OntoNotes, WordNet glosses
2012	SMTnews	750	Machine Translation evaluation
2012	SMTeuroparl	750	Machine Translation evaluation
2013	headlines	750	newswire headlines
2013	FNWN	189	FrameNet, WordNet glosses
2013	OnWN	561	OntoNotes, WordNet glosses
2013	SMT	750	Machine Translation evaluation
2014	headlines	750	newswire headlines
2014	OnWN	750	OntoNotes, WordNet glosses
2014	Deft-forum	450	forum posts
2014	Deft-news	300	news summary
2014	Images	750	image descriptions
2014	Tweet-news	750	tweet-news pairs
2015	image	750	image description
2015	headlines	750	news headlines
2015	answers-students	750	student answers,reference answers
2015	answers-forum	375	answers in stack exchange forums
2015	belief	375	forum data exhibiting committed belief

Table 1: Summary of STS datasets in 2012, 2013, 2014, 2015.

3 Experiments

3.1 Datasets

The STS (for English) dataset consists of several datasets: STS 2012, STS 2013, STS 2014 and STS 2015 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014; Agirre et al., 2015). Each sentence pair is annotated with the semantic similarity score in the scale [0-5]. Table 1 shows the summary of STS datasets and sources over the years. For training, we use all data in STS 2012, 2013 and 2014; and for testing, we use STS 2015 datasets.

3.2 Evaluation Methods

We use two different evaluation methods to evaluate the impact of the metrics on our training dataset, (1) the Pearson correlation between the metric outputs and the gold standards which is the official evaluation method used in STS task; and (2) the RELIEF (Robnik-Sikonja and Kononenko, 1997) analysis implemented in WEKA (Hall et al., 2009) to estimate the quality of MT evaluation metric output in regression.

3.3 Settings

Firstly, we employ the four metrics to compute the semantic similarity between given sentences on the training dataset. We use the default configuration for all metrics, except the "-norm" option for METEOR that tokenizes and normalizes punctuation and lowercase, as suggested in its documentation; and the "-c" option for TER and TERp that roofs the score to 100. Then we normalize all

the output results to the scale [0-1].

Next, we combine the outputs of these four metrics to build eight different regression models using different classification algorithms in WEKA (e.g. IsotonicRegression, LeastMedSq, MultilayerPerceptron, SimpleLinearRegression, LinearRegression, M5Rules, M5 Model Trees, and DecisionTable). We only use the default settings of each algorithm without tuning any parameter because our goal is to compare the results of different approaches, not to obtain high performance. We evaluate each model twice, (i) by a 10-fold cross validation on training data, and (ii) we evaluate the model on the test data (STS 2015 dataset). For the comparison, we use the official baseline of STS task which uses the bag-of-words approach to represent each sentence as a vector in the multi-dimensional token space (each dimension has 1 if the token is present in the sentence, 0 otherwise) and computes the cosine similarity between vectors.

4 Evaluations and Discussions

4.1 Evaluation of Individual Metric

The Pearson correlation and RELIEF analysis of each single metric compared to the human-annotation scores are presented in Table 2. According to both methods, the METEOR tends to be the superior metric, while in contrast TERp has low values in both. We split the BLEU metric into four values for 1-gram, 2-gram, 3-gram and 4-gram. The Pearson correlation shows that the smaller size of n-gram overlap, the more correlation with the human judgment obtained. In overall, except TER that has inverse correlation which is the more negative result, the better correlation with human annotation scores, other metrics have reasonable correlation. Nevertheless, another error metric, TERp does not perform well and returns a positive correlation, opposite to the TER metric.

We also investigate the behaviour of each metric deeper inside each score bracket in the STS semantic scale. We plot the output of each metric in corresponding to each score bracket [0-1], [1-2], [2-3], [3-4] and [4-5] to see how each MT metric behaves on each score bracket. The results of RELIEF analysis and Pearson correlation in Figure 1 and 2 show that most of the metrics perform well in two particular score brackets [0-1] and [4-5]. This means that by deploying MT evaluation metrics for STS task, the system will be able to obtain a high precision of predicting the semantic similarity for two cases "not/almost not relevant" and "equivalent/almost equivalent". This investigation can help to significantly improve the overall performance of a STS system by increasing the accuracy of predicting the scores in brackets [0-1] and [4-5]. In contrast, both figures have a central region where the correlation scores decrease significantly, and even worst for TERp where Pearson correlation changes signs, that means that in some regions this metric switches from direct to inverse correlation.

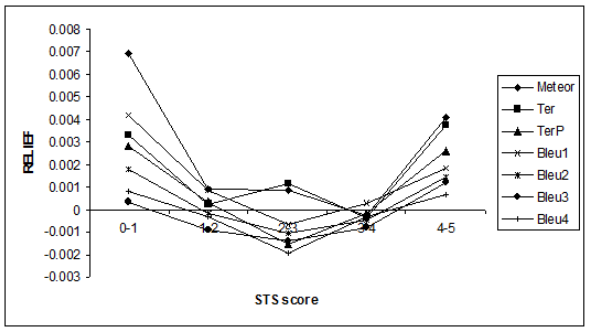

Figure 1: RELIEF analysis.

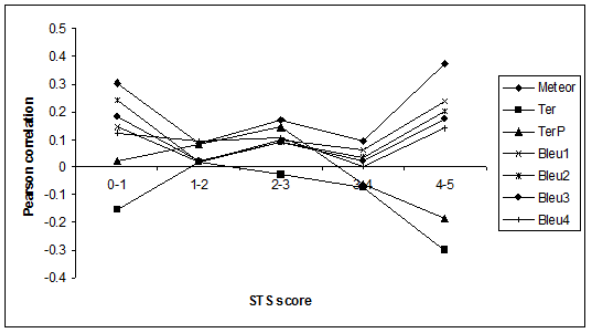

Figure 2: Pearson correlation.

	RELIEF	Pearson
METEOR	0.00503	0.56065
TER	-0.00157	-0.25673
TERp	-0.00098	0.21047
BLEU-1	-0.00145	0.36800
BLEU-2	-0.00201	0.31801
BLEU-3	-0.00203	0.27074
BLEU-4	-0.00249	0.27233

Table 2: Evaluation of the different features on the training dataset.

	IR	LMS	MLP	SLR	LR	M5R	M5P	DT	Baseline	BestSys
Cross-validation	0.610	0.629	0.606	0.560	0.653	0.737	0.739	0.698	0.382	-
Test set	0.702	0.643	0.694	0.688	0.612	0.609	0.611	0.588	0.587	**0.801**
Standard deviation	0.429	0.458	0.475	0.444	0.404	0.363	0.363	0.386	0.579	-

Table 3: Evaluation of the different algorithms: Pearson coefficient (IR: IsotonicRegression, LMS: LeastMedSq, MLP: MultilayerPerceptron, SLR: SimpleLinearRegression, LR: LinearRegression, M5R: M5Rules, M5P: M5 Model Trees, DT: DecisionTable, Baseline: STS Baseline, BestSys: 1st ranked system in STS 2015).

This enlightens an important difference between the impact of these metrics on STS task and on the paraphrase recognition task: while MT metrics show acceptable performance distinguishing the border regions, i.e. the most similar (almost paraphrase) and the most dissimilar, they have worse performance in the middle regions.

4.2 Evaluation of Metric Combination

We examine the impact of the combination of all metrics to the overall performance in STS by building several regression models using all the metric outputs as features. Since every metric and also the STS score is a numeric value we use normal regression algorithms. The results of these analysis are reported in Table 3 which shows, (i) the average of the 10-Folds cross-validation on the training data, (ii) the overall performance on the test data, and (iii) to better describe the different algorithms, we also report the standard deviation (SD) of the ten standard deviations from ten folds; we use this measure as an index to evaluate if the performances of the classifier during cross-validation are uniform or present some instability due to specific fold.

We group the models into two groups by a threshold of the standard deviation (SD $= 0.41$) in which the lower SD, the more reliable model is and vice versa. It is interesting to notice that more stable models (on the right hand side) perform well the cross-validation on the training dataset, but obtain low performance on the test dataset, in a margin of 10% (except the LR having margin of 1%). Nevertheless, the less stable models (on the left hand side) obtain better results on the test dataset and low performance on the cross-validation, in a margin of 2-10%. From our observation, another important aspect is that not all the algorithms use all given features in the same way, but during the training phase Isotonic Regression (IR) and Simple Linear Regression (SLR) discard other features and use only METEOR metric.

Another interesting observation is the different learning approaches of different algorithms taking advantage from MT metrics. Some algorithms can learn more information from the combination of these metrics and perform well the cross validation on training data, but when being evaluated on the test data, the model is strongly penalized by the domain-independence datasets in STS. In our case the STS 2012, 2013 and 2014 datasets are different from the STS 2015, which leads to an overfitting of the systems that builds the model using all these features. On the other hand, algorithms which are not so optimized can use MT metrics in a more flexible way to obtain good result on the test dataset.

In overall, all the regression models using combination of MT metrics outperform the task baseline in both cross validation on training dataset (by a large margin of 22-36%) and performance on test dataset (by a margin of 0.1-12%). However, none of these models can compare to the best system on the test dataset, the difference between the best model and the best system is a large margin of 10%. This proves that using only MT metric is not sufficient and efficient enough to solve the STS task. But combining MT metrics with other linguistic features may return promising result.

5 Conclusions and Future Work

In this study, we show the notable characteristic of the MT metrics as features for the STS task. The distribution of correlation between MT metrics and STS human judgment indicates that this feature is reliable only in the border regions of the [0-5] scale, in particular in [0-1] and [4-5]. This result means that, MT metrics have interesting degrees of correlation with STS, so they are useful features for the task, but from the other side it means that they can not be used alone, because

their performance are very low in the [1-4] range. Among the different metrics, METEOR has superior property compared to others and it proves to be an useful feature, even alone, to build acceptable STS systems.

In future we want to investigate more on the impact of other MT metrics on STS task. In this paper we have focused on the distribution of correlation on the [0-5] scale, but a study of the distribution on the different domain would give other important information on these features. Our aim is to find the most useful MT metric or the best combination of metrics among others, and the most reliable and effective algorithm to obtain better performance on the STS task. We also want to extend the study to multilingual STS, for instance, STS for Spanish, to learn if the impact and behavior of MT evaluation metrics remain the same in other languages.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Proceedings of the Main Conference and the Shared Task*.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, June. Association for Computational Linguistics.

Luis Alberto Barrón Cedeño, Lluís Màrquez Villodre, Maria Fuentes Fort, Horacio Rodríguez Hontoria, Jorge Turmo Borras, et al. 2013. UPC-CORE: What can machine translation evaluation metrics and wikipedia do for estimating semantic textual similarity? In *Proceedings of the Joint Conference on Lexical and Computational Semantics. "*SEM 2013: The Second Joint Conference on Lexical and Computational Semantics"*.

José Guilherme C de Souza, Matteo Negri, and Yashar Mehdad. 2012. Fbk: machine translation evaluation and word similarity metrics for semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 624–630. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190. Association for Computational Linguistics.

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*, pages 311–318. Association for Computational Linguistics.

Marko Robnik-Sikonja and Igor Kononenko. 1997. An adaptation of relief for attribute estimation in regression. In Douglas H. Fisher, editor, *Fourteenth International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Matthew G Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Ter-plus: paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23(2-3):117–127.