# UPC-CORE: What Can Machine Translation Evaluation Metrics and Wikipedia Do for Estimating Semantic Textual Similarity?

**Alberto Barrón-Cedeño**[1,2] **Lluís Màrquez**[1] **Maria Fuentes**[1] **Horacio Rodríguez**[1] **Jordi Turmo**[1]
[1] TALP Research Center, Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, 08034, Barcelona, Spain
[2] Facultad de Informática, Universidad Politécnica de Madrid
Boadilla del Monte, 28660 Madrid, Spain
albarron, lluism, mfuentes, horacio, turmo @lsi.upc.edu

## Abstract

In this paper we discuss our participation to the 2013 Semeval Semantic Textual Similarity task. Our core features include (*i*) a set of metrics borrowed from automatic machine translation, originally intended to evaluate automatic against reference translations and (*ii*) an instance of explicit semantic analysis, built upon opening paragraphs of Wikipedia 2010 articles. Our similarity estimator relies on a support vector regressor with RBF kernel. Our best approach required 13 machine translation metrics + explicit semantic analysis and ranked 65 in the competition. Our post-competition analysis shows that the features have a good expression level, but overfitting and —mainly— normalization issues caused our correlation values to decrease.

## 1 Introduction

Our participation to the 2013 Semantic Textual Similarity task (STS) (Agirre et al., 2013)[1] was focused on the CORE problem: GIVEN TWO SENTENCES, $s_1$ AND $s_2$, QUANTIFIABLY INFORM ON HOW SIMILAR $s_1$ AND $s_2$ ARE. We considered real-valued features from four different sources: (*i*) a set of linguistic measures computed with the Asiya Toolkit for Automatic MT Evaluation (Giménez and Màrquez, 2010b), (*ii*) an instance of explicit semantic analysis (Gabrilovich and Markovitch, 2007), built on top of Wikipedia articles, (*iii*) a dataset predictor, and (*iv*) a subset of the features available in Takelab's Semantic Text Similarity system (Šarić et al., 2012).

Our approaches obtained an overall modest result compared to other participants (best position: 65 out of 89). Nevertheless, our post-competition analysis shows that the low correlation was caused mainly by a deficient data normalization strategy.

The paper distribution is as follows. Section 2 offers a brief overview of the task. Section 3 describes our approach. Section 4 discuss our experiments and obtained results. Section 5 provides conclusions.

## 2 Task Overview

Detecting two similar text fragments is a difficult task in cases where the similarity occurs at semantic level, independently of the implied lexicon (e.g in cases of dense paraphrasing). As a result, similarity estimation models must involve features other than surface aspects. The STS task is proposed as a challenge focused in short English texts of different nature: from automatic machine translation alternatives to human descriptions of short videos. The test partition also included texts extracted from news headlines and FrameNet–Wordnet pairs.

The range of similarity was defined between 0 (no relation) up to 5 (semantic equivalence). The gold standard values were averaged from different human-made annotations. The expected system's output was composed of a real similarity value, together with an optional confidence level (our confidence level was set constant).

Table 1 gives an overview of the development (2012 training and test) and test datasets. Note that both collections extracted from SMT data are highly biased towards the maximum similarity values (more than 75% of the instances have a similar-

[1] http://ixa2.si.ehu.es/sts/

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 143–147, Atlanta, Georgia, June 13-14, 2013.

Table 1: Overview of sub-collections in the development and test datasets, including number of instances and distribution of similarity values (in percentage) as well as mean, minimum, and maximum lengths.

| | | | | **similarity distribution** | | | | **length** | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | instances | $[0,1)$ | $[1,2)$ | $[2,3)$ | $[3,4)$ | $[4,5]$ | mean | min | max |
| **dev-[train + test]** | | | | | | | | | |
| MSRpar | 1,500 | 1.20 | 8.13 | 17.13 | **48.73** | 24.80 | 17.84 | **5** | 30 |
| MSRvid | 1,500 | 31.00 | 14.13 | 15.47 | 20.87 | 18.53 | 6.66 | 2 | 24 |
| SMTEuroparl | 1,193 | 0.67 | 0.42 | 1.17 | 12.32 | **85.4** | 21.13 | 1 | 72 |
| OnWN | 750 | 2.13 | 2.67 | 10.40 | 25.47 | 59.33 | 7.57 | 1 | 34 |
| SMTnews | 399 | 1.00 | 0.75 | 5.51 | 13.03 | 79.70 | 11.72 | 2 | 28 |
| **test** | | | | | | | | | |
| headlines | 750 | 15.47 | 22.00 | 16.27 | 24.67 | 21.60 | 7.21 | 3 | 22 |
| OnWN | 561 | **36.54** | 9.80 | 7.49 | 17.11 | 29.05 | 7.17 | **5** | 22 |
| FNWN | 189 | 34.39 | **29.63** | **28.57** | 6.88 | 0.53 | 19.90 | 3 | 71 |
| SMT | 750 | 0.00 | 0.27 | 3.47 | 20.40 | 75.87 | **26.40** | 1 | **96** |

ity higher than 4) and include the longest instances. On the other hand, the FNWN instances are shifted towards low similarity levels (more than 60% have a similarity lower than 2).

## 3 Approach

Our similarity assessment model relies upon SVM$^{light}$'s support vector regressor, with RBF kernel (Joachims, 1999).[2] Our model estimation procedure consisted of two steps: parameter definition and backward elimination-based feature selection. The considered features belong to four families, briefly described in the following subsections.

### 3.1 Machine Translation Evaluation Metrics

We consider a set of linguistic measures originally intended to evaluate the quality of automatic translation systems. These measures compute the quality of a translation by comparing it against one or several reference translations, considered as gold standard. A straightforward application of these measures to the problem at hand is to consider $s_1$ as the reference and $s_2$ as the automatic translation, or vice versa. Some of the metrics are not symmetric so we compute similarity between $s_1$ and $s_2$ in both directions and average the resulting scores.

The measures are computed with the Asiya Toolkit for Automatic MT Evaluation (Giménez and Màrquez, 2010b). The only pre-processing carried out was tokenization (Asiya performs additional inbox pre-processing operations, though). We considered a sample from three similarity families, which was proposed in (Giménez and Màrquez, 2010a) as a varied and robust metric set, showing good correlation with human assessments.[3]

**Lexical Similarity** Two metrics of Translation Error Rate (Snover et al., 2006) (i.e. the estimated human effort to convert $s_1$ into $s_2$): -TER and -TER$_{pA}$. Two measures of lexical precision: BLEU (Papineni et al., 2002) and NIST (Doddington, 2002). One measure of lexical recall: ROUGE$_W$ (Lin and Och, 2004). Finally, four variants of METEOR (Banerjee and Lavie, 2005) (*exact*, *stemming*, *synonyms*, and *paraphrasing*), a lexical metric accounting for $F$-Measure.

**Syntactic Similarity** Three metrics that estimate the similarity of the sentences over dependency parse trees (Liu and Gildea, 2005): DP-HWCMi$_c$-4 for grammatical categories chains, DP-HWCMi$_r$-4 over grammatical relations, and DP-O$_r(\star)$ over words ruled by non-terminal nodes. Also, one measure that estimates the similarity over constituent parse trees: CP-STM$_4$ (Liu and Gildea, 2005).

**Semantic Similarity** Three measures that estimate the similarities over semantic roles (i.e. arguments and adjuncts): SR-O$_r$, SR-M$_r(\star)$, and SR-O$_r(\star)$. Additionally, two metrics that estimate similarities over discourse representations: DR-O$_r(\star)$ and DR-O$_{rp}(\star)$.

[2] We also tried with linear kernels, but RBF always obtained better results.

[3] Asiya is available at http://asiya.lsi.upc.edu. Full descriptions of the metrics are available in the Asiya Technical Manual v2.0, pp. 15–21.

### 3.2 Explicit Semantic Analysis

We built an instance of Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) with the first paragraph of $100k$ Wikipedia articles (dump from 2010).Pre-processing consisted of tokenization and lemmatization.

### 3.3 Dataset Prediction

Given the similarity shifts in the different datasets (cf. Table 1), we tried to predict what dataset an instance belonged to on the basis of its vocabulary. We built binary maxent classifiers for each dataset in the development set, resulting in five dataset likelihood features: dMSRpar, dSMTeuroparl, dMSRvid, dOnWN, and dSMTnews.[4] Pre-processing consisted of tokenization and lemmatization.

### 3.4 Baseline

We considered the features included in the Takelab Semantic Text Similarity system (Šarić et al., 2012), one of the top-systems in last year competition. This system is used as a black box. The resulting features are named tklab_n, where $n = [1, 21]$.

Our runs departed from three increasing subsets of features: **AE** machine translation evaluation metrics and explicit semantic analysis, **AED** the previous set plus dataset prediction, and **AED_T** the previous set plus Takelab's baseline features (cf. Table 3). We performed a feature normalization, which relied on the different feature's distribution over the entire dataset. Firstly, features were bounded in the range $\mu \pm 3 * \sigma^2$ in order to reduce the potentially negative impact of outliers. Secondly, we normalized according to the $z$-score (Nardo et al., 2008, pp. 28, 84); i.e. $x = (x - \mu)/\sigma$. As a result, each real-valued feature distribution in the dataset has $\mu = 0$ and $\sigma = 1$. During the model tuning stage we tried with other numerous normalization options: normalizing each dataset independently, together with the training set, and without normalization at all. Normalizing according to the entire dev-test dataset led to the best results

Table 2: Tuning process: parameter definition and feature selection. Number of features at the **b**eginning and **e**nd of the feature selection step included.

| run | parameter def. | | | | feature sel. | | |
|---|---|---|---|---|---|---|---|
| | **c** | $\gamma$ | $\epsilon$ | **corr** | **b** | **e** | **corr** |
| AE | 3.7 | 0.06 | 0.3 | 0.8257 | 19 | 14 | 0.8299 |
| AED | 3.8 | 0.03 | 0.2 | 0.8413 | 24 | 19 | 0.8425 |
| AED_T | 2.9 | 0.02 | 0.3 | 0.8761 | 45 | 33 | 0.8803 |

## 4 Experiments and Results

Section 4.1 describes our model tuning strategy. Sections 4.2 and 4.3 discuss the official and post-competition results.

### 4.1 Model Tuning

We used only the dev-train partition (2012 training) for tuning. By means of a 10-fold cross validation process, we defined the trade-off (c), gamma ($\gamma$), and tube width ($\epsilon$) parameters for the regressor and performed a backward-elimination feature selection process (Witten and Frank, 2005, p. 294), independently for the three experiments.

The results for the cross-validation process are summarized in Table 2. The three runs allow for correlations higher than 0.8. On the one hand, the best regressor parameters obtain better results as more features are considered, still with very small differences. On the other hand, the low correlation increase after the feature selection step shows that a few features are indeed irrelevant.

A summary of the features considered in each experiment (also after feature selection) is displayed in Table 3. The correlation obtained over the dev-test partition are $corr_{AE} = 0.7269$, $corr_{AED} = 0.7638$, and $corr_{AED_T} = 0.8044$ —it would have appeared in the top-10 ranking of the 2012 competition.

### 4.2 Official Results

We trained three new regressors with the features considered relevant by the tuning process, but using the entire development dataset. The test 2013 partition was normalized again by means of $z$-score, considering the means and standard deviations of the entire test dataset. Table 4 displays the official results. Our best approach —**AE**—, was positioned in rank 65. The worst results of run **AED** can be explained by the difference in the nature of the test respect to

[4] We used the Stanford classifier; http://nlp.stanford.edu/software/classifier.shtml

Table 3: Features considered at the beginning of each run, represented as empty squares (□). Filled squares (■) represent features considered relevant after feature selection.

| Feature | AE | AED | AED_T | Feature | AE | AED | AED_T | Feature | AED_T |
|---|---|---|---|---|---|---|---|---|---|
| DP-HWCM_c-4 | ■ | ■ | ■ | METEOR-pa | ■ | ■ | ■ | tklab_7 | ■ |
| DP-HWCM_r-4 | ■ | ■ | ■ | METEOR-st | □ | ■ | □ | tklab_8 | ■ |
| DP-Or(*) | ■ | ■ | ■ | METEOR-sy | ■ | ■ | □ | tklab_9 | ■ |
| CP-STM-4 | □ | □ | ■ | ESA | ■ | ■ | ■ | tklab_10 | □ |
| SR-Or(*) | □ | □ | ■ | dMSRpar |  | ■ | □ | tklab_11 | ■ |
| SR-Mr(*) | ■ | ■ | ■ | dSMTeuroparl |  | ■ | ■ | tklab_12 | ■ |
| SR-Or | ■ | ■ | ■ | dMSRvid |  | ■ | □ | tklab_13 | ■ |
| DR-Or(*) | □ | ■ | ■ | dOnWN |  | □ | □ | tklab_14 | ■ |
| DR-Orp(*) | ■ | ■ | ■ | dSMTnews |  | □ | □ | tklab_15 | ■ |
| BLEU | ■ | ■ | □ | tklab_1 |  |  | □ | tklab_16 | ■ |
| NIST | ■ | ■ | ■ | tklab_2 |  |  | ■ | tklab_17 | ■ |
| -TER | ■ | ■ | ■ | tklab_3 |  |  | ■ | tklab_18 | ■ |
| -TERp-A | ■ | ■ | ■ | tklab_4 |  |  | ■ | tklab_19 | ■ |
| ROUGE-W | ■ | ■ | □ | tklab_5 |  |  | ■ | tklab_20 | □ |
| METEOR-ex | □ | □ | ■ | tklab_6 |  |  | □ | tklab_21 | ■ |

Table 4: Official results for the three runs (rank included).

| run | headlines | OnWN | FNWN | SMT | mean |
|---|---|---|---|---|---|
| AE (65) | 0.6092 | 0.5679 | -0.1268 | 0.2090 | 0.4037 |
| AED (83) | 0.4136 | 0.4770 | -0.0852 | 0.1662 | 0.3050 |
| AED_T (72) | 0.5119 | 0.6386 | -0.0464 | 0.1235 | 0.3671 |

Table 5: Post-competition experiments results

| run | headlines | OnWN | FNWN | SMT | mean |
|---|---|---|---|---|---|
| AE (*a*) | 0.6210 | 0.5905 | -0.0987 | 0.2990 | 0.4456 |
| AE (*b*) | 0.6072 | 0.4767 | -0.0113 | 0.3236 | 0.4282 |
| AE (*c*) | 0.6590 | 0.6973 | 0.1547 | 0.3429 | 0.5208 |

the development dataset. **AED_T** obtains worst results than **AE** on the *headlines* and *SMT* datasets. The reason behind this behavior can be in the difference of vocabularies respect to that stored in the Takelab system (it includes only the vocabulary of the development partition). This could be the same reason behind the drop in performance with respect to the results previously obtained on the dev-test partition (cf. Section 4.1).

### 4.3 Post-Competition Results

Our analysis of the official results showed the main issue was normalization. Thus, we performed a manifold of new experiments, using the same configuration as in run **AE**, but applying other normalization strategies: (*a*) $z$-score normalization, but ignoring the FNWN dataset (given its shift through low values); (*b*) $z$-score normalization, but considering independent means and standard deviations for each test dataset; and (*c*) without normalizing any of dataset (including the regressor one).

Table 5 includes the results. (*a*) makes evident that the instances in FNWN represent "anomalies" that harm the normalized values of the rest of subsets. Run (*b*) shows that normalizing the test sets independently is not a good option, as the regressor is trained considering overall normalizations, which explains the correlation decrease. Run (*c*) is completely different: not normalizing any dataset — both in development and test— reduces the influence of the datasets to each other and allows for the best results. Indeed, this configuration would have advanced practically forty positions at competition time, locating us in rank 27.

Estimating the adequate similarities over *FNWN* seems particularly difficult for our systems. We observe two main factors. (*i*) *FNWN* presents an important similarity shift respect to the other datasets: nearly 90% of the instances similarity is lower than 2.5 and (*ii*) the average lengths of $s_1$ and $s_2$ are very different: 30 vs 9 words. These characteristics made it difficult for our MT evaluation metrics to estimate proper similarity values (be normalized or not).

We performed two more experiments over FNWN: training regressors with ESA as the only feature, before and after normalization. The correlation was 0.16017 and 0.3113, respectively. That is, the normalization mainly affects the MT features.

## 5 Conclusions

In this paper we discussed on our participation to the 2013 Semeval Semantic Textual Similarity task. Our approach relied mainly upon a combination of automatic machine translation evaluation metrics and explicit semantic analysis. Building an RBF support vector regressor with these features allowed us for a modest result in the competition (our best run was ranked 65 out of 89).

## Acknowledgments

We would like to thank the organizers of this challenging task for their efforts.

This research work was partially carried out during the tenure of an ERCIM “Alain Bensoussan” Fellowship. The research leading to these results received funding from the EU FP7 Programme 2007-2013 (grants 246016 and 247762). Our research work is partially supported by the Spanish research projects OpenMT-2 and SKATER (TIN2009-14675-C03, TIN2012-38584-C06-01).

## References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 Shared Task: Semantic Textual Similarity, including a Pilot on Typed-Similarity. In **SEM 2013: The Second Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein et al. (Goldstein et al., 2005), pages 65–72.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Francisco, CA. Morgan Kaufmann Publishers Inc.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jesús Giménez and Lluís Màrquez. 2010a. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94).

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4):209–240.

Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors. 2005. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics.

Thorsten Joachims, 1999. *Advances in Kernel Methods – Support Vector Learning*, chapter Making large-Scale SVM Learning Practical. MIT Press.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Stroudsburg, PA. Association for Computational Linguistics.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In Goldstein et al. (Goldstein et al., 2005), pages 25–32.

Michela Nardo, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffmann, and Enrico Giovannini. 2008. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD Publishing.

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 (ACL 2002)*, pages 311–318, Philadelphia, PA. Association for Computational Linguistics.

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.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. TakeLab: Systems for Measuring Semantic Text. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 441–448, Montréal, Canada. Association for Computational Linguistics.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, 2 edition.