SAGAN: An approach to Semantic Textual Similarity based on Textual Entailment

Julio Castillo[†‡] **Paula Estrella**[‡]
[‡]FaMAF, UNC, Argentina
[†]UTN-FRC, Argentina
jotacastillo@gmail.com
pestrella@famaf.unc.edu.ar

Abstract

In this paper we report the results obtained in the Semantic Textual Similarity (STS) task, with a system primarily developed for textual entailment. Our results are quite promising, getting a run ranked 39 in the official results with overall Pearson, and ranking 29 with the Mean metric.

1 Introduction

For the last couple of years the research community has focused on a deeper analysis of natural languages, seeking to capture the meaning of the text in different contexts: in machine translation preserving the meaning of the translations is crucial to determine whether a translation is useful or not, in question-answering understanding the question leads to the desired answers (while the opposite case makes a system rather frustrating to the user) and the examples could continue. In this newly defined task, Semantic Textual Similarity, there is hope that efforts in different areas will be shared and united towards the goal of identifying meaning and recognizing equivalent, similar or unrelated texts. Our contribution to the task, is from a textual entailment point of view, as will be described below.

The paper is organized as follows: Section 2 describes the relevant tasks, Section 3 describes the architecture of the system, then Section 4 shows the experiments carried out and the results obtained, and Section 5 presents some conclusions and future work.

2 Related work

In this section we briefly describe two different tasks that are closely related and in which our system has participated with very promising results.

2.1 Textual Entailment

Textual Entailment (TE) is defined as a generic framework for applied semantic inference, where the core task is to determine whether the meaning of a target textual assertion (hypothesis, H) can be inferred from a given text (T). For example, given the pair (T,H):
T: Fire bombs were thrown at the Tunisian embassy in Bern
H: The Tunisian embassy in Switzerland was attacked
we can conclude that T entails H.

The recently created challenge "Recognising Textual Entailment" (RTE) started in 2005 with the goal of providing a binary answer for each pair (H,T), namely whether there is entailment or not (Dagan et al., 2006). The RTE challenge has mutated over the years, aiming at accomplishing more

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 667–672,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

accurate and specific solutions; for example, in 2008 a three-way decision was proposed (instead of the original binary decision) consisting of "entailment", "contradiction" and "unknown"; in 2009 the organizers proposed a pilot task, the Textual Entailment Search (Bentivogli et al, 2009), consisting in finding all the sentences in a set of documents that entail a given Hypothesis and since 2010 there is a Novelty Detection Task, which means that RTE systems are required to judge whether the information contained in each H is novel with respect to (i.e., not entailed by) the information contained in the corpus.

2.2 Semantic Textual Similarity

The pilot task STS was recently defined in Semeval 2012 (Aguirre et al., 2012) and has as main objective measuring the degree of semantic equivalence between two text fragments. STS is related to both Recognizing Textual Entailment (RTE) and Paraphrase Recognition, but has the advantage of being a more suitable model for multiple NLP applications.

As mentioned before, the goal of the RTE task (Bentivogli et al, 2009) is determining whether the meaning of a hypothesis H can be inferred from a text T. Thus, TE is a directional task and we say that T entails H, if a person reading T would infer that H is most likely true. The difference with STS is that STS consists in determining how similar two text fragments are, in a range from 5 (total semantic equivalence) to 0 (no relation). Thus, STS mainly differs from TE in that the classification is graded instead of binary. In this manner, STS is filling the gap between several tasks.

3 System architecture

Sagan is a RTE system (Castillo and Cardenas, 2010) which has taken part of several challenges, including the Textual Analysis Conference 2009 and TAC 2010, and the Semantic Textual Similarity and Cross Lingual Textual Entailment for content synchronization as part of the Semeval 2012.

The system is based on a machine learning approach and it utilizes eight WordNet-based (Fellbaum, 1998) similarity measures, as explained in (Castillo, 2011), with the purpose of obtaining the maximum similarity between two WordNet concepts. A concept is a cluster of synonymous terms that is called a synset in WordNet. These text-to-text similarity measures are based on the following word-to-word similarity metrics: (Resnik, 1995), (Lin, 1997), (Jiang and Conrath, 1997), (Pirrò and Seco, 2008), (Wu & Palmer, 1994), Path Metric, (Leacock & Chodorow, 1998), and a semantic similarity to sentence level named SemSim (Castillo and Cardenas,2010).

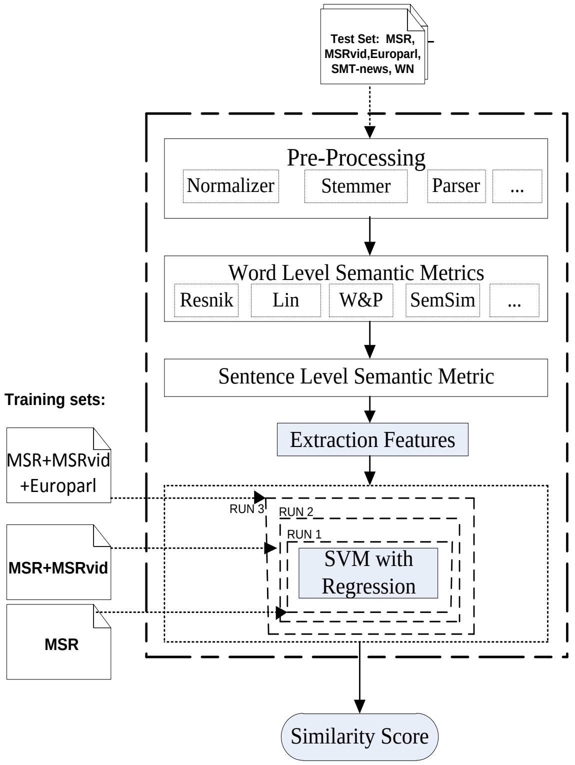

Fig.1. System architecture

The system construct a model of the semantic similarity of two texts (T,H) as a function of the semantic similarity of the constituent words of both phrases. In order to reach this objective, we used a text to text similarity measure which is based on word to word similarity. Thus, we expect that combining word to word similarity metrics to text level would be a good indicator of text to text similarity.

Additional information about how to produce feature vectors as well as each word- and sentence-level metric can be found in (Castillo, 2011). The architecture of the system is shown in Figure 1.

The training set used for the submitted runs are those provided by the organizers of the STS. However we also experimented with RTE datasets as described in the next Section.

4 Experiments and Results

For preliminary experiments before the STS Challenge, we used the training set provided by the organizers, denoted with "_train", and consisting of 750 pairs of sentences from the MSR Paraphrase Corpus (MSRpar), 750 pairs of sentences from the MSRvid Corpus (MSRvid), 459 pairs of sentences of the Europarl WMT2008 development set (SMT-eur). We also used the RTE datasets from Pascal RTE Challenge (Dagan et al., 2006) as part of our training sets. Additionally, at the testing stage, we used the 399 pairs of news conversation (SMT-news) and 750 pairs of sentences where the first one comes from Ontonotes and the second one from a WordNet definition (On-WN).

In STS Challenge it was required that participating systems do not use the test set of MSR-Paraphrase, the text of the videos in MSR-Video, and the data from the evaluation tasks at any WMT to develop or train their systems. Additionally, we also assumed that the dataset to be processed was unknown in the testing phase, in order to avoid any kind of tuning of the system.

4.1 Preliminary Experiments

In a preliminary study performed before the final submission, we experimented with three machine learning algorithms Support Vector Machine (SVM) with regression and polynomial kernel, Multilayer perceptron (MLP), and Linear Regression (LR). Table 1 shows the results obtained with 10-fold cross validation technique and Table 2 shows the results of testing them with two datasets and 3 classifiers over MSR_train.

Classifier	Pearson c.c
SVM with regression	0.54
MLP	0.51
LinearRegression	0.54

Table 1. Results obtained using MSR training set (MSRpar + MSRvid) with 10 fold-cross validation.

Training set & ML algorithm	Pearson c.c
Europarl + SVM w/ regression	0.61
Europarl + MLP	0.44
Europarl + linear regression	0.61
MSRvid + SVM w/ regression	0.70
MSRvid + MLP	0.52
MSRvid + linear regression	0.69

Table 2. Results obtained using MSR training set

Results reported in Table 1 show that we achieved the best performance with SVM with regression and Linear Regression classifiers and using MLP we obtained the worst results to predict each dataset. To our surprise, a linear regression classifier reports better accuracy that MLP, it may be mainly due to the correlation coefficient used, namely Pearson, which is a measure of a linear dependence between two variables and linear regression builds a model assuming linear influence of independent features. We believe that using Spearman correlation should be better than using the Pearson coefficient given that Spearman assumes non-linear correlation among variables. However, it is not clear how it behaves when several dataset are combined to obtain a global score. Indeed, further discussion is needed in order to find the best metric to the STS pilot task. Given these results, in our submission for the STS pilot task we used a combination of STS datasets as training set and the SVM with regression classifier.

Because our approach is mainly based on machine learning the quality and quantity of dataset is a key factor to determine the performance of the system, thus we decided to experiment with RTE datasets too (Bentivogli et el., 2009) with the aim of increasing the size of the training set.

To achieve this goal, first we chose the RTE3 dataset because it is simpler than subsequent datasets and it was proved to provide a high accuracy predicting other datasets (Castillo, 2011). Second, taking into account that RTE datasets are binary classified as YES or NO entailment, we assumed that a non entailment can be treated as a value of 2.0 in the STS pilot task and an entailment can be thought of as a value of 3.0 in STS. Of course, many pairs classified as 3.0 could be mostly equivalent (4.0) or completely equivalent (5.0) but we ignored this fact in the following experiment.

Training set	Test set	Pearson c.c.
RTE3	MSR_train	0.4817
RTE3	MSRvid_train	0.5738
RTE3	Europarl_train	0.4746
MSR_train+RTE3	MSRvid_train	0.5652
MSR_train+RTE3	Europarl_train	0.5498
MSRvid_train+RTE3	MSR_train	0.4559
MSRvid_train+RTE3	Europarl_train	0.4964

Table 3. Results obtained using RTE in the training sets and SVM w/regression as classifier

From these experiments we conclude that RTE3 alone is not enough to adequately predict neither of the STS datasets, and it is understandable if we note that only one pair with 2.0 and 3.0 scores are present in this dataset.

On the other hand, by combining RTE3 with a STS corpus we always obtain a slight decrease in performance in comparison to using STS alone. It is likely due to an unbalanced set and possible contradictory pairs (e.g: a par in RTE3 classified as 3.0 when it should be classified 4.3). Thus, we conclude that in order to use the RTE datasets our system needs a manual annotation of the degree of semantic similarity of every pair <T,H> of RTE dataset.

Having into account that in our training phase we obtained a decrease in performance using RTE datasets we decided not to submit any run using the RTE datasets.

4.2 Submission to the STS shared task

Our participation in the shared task consisted of three different runs using a SVM classifier with regression; the runs were set up as follows:
- Run 1: system trained on a subset of the Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), named MSR and consisting of 750 pairs of sentences marked with a degree of similarity from 5 to 0.
- Run 2: in addition to the MSR corpus we incorporated another 750 sentences extracted from the Microsoft Research Video Description Corpus (MSRvid), annotated in the same way as MSR.
- Run 3: to the 1500 sentences from the MSR and MSRvid corpus we incorporated 734 pairs of sentences from the Europarl corpus used as development set in the WMT 2008; all sentences are annotated with the degree of similarity from 5 to 0.

It is very interesting to note that we used the same system configurations for every dataset of each RUN. In this manner, we did not perform any kind of tuning to a particular dataset before our submission. We decided to ignore the "name" of each dataset and apply our system regardless of the particular dataset. Surely, if we take into account where each dataset came from we can develop a particular strategy for every one of them, but we assumed that this kind of information is unknown to our system.

The official scores of the STS pilot task is the Pearson correlation coefficient, and other variations of Pearson which were proposed by the organizers with the aim of better understanding the behavior of the competing systems among the different scenarios.

These metric are named ALL (overall Pearson), ALLnrm (normalized Pearson) and Mean (weighted mean), briefly described below:
- ALL: To compute this metric, first a new dataset with the union of the five gold datasets is created and then the Pearson correlation is calculated over this new dataset.
- ALLnrm: In this metric, the Pearson correlation is computed after the system outputs for each dataset are fitted to the gold standard using least squares.
- Mean: This metric is a weighted mean across the five datasets, where the weight is given by the quantity of pairs in each dataset.

Table 5 report the results achieved with these metrics followed by an individual Pearson correlation for each dataset.

Interestingly, if we analyze the size of data sets, we see that the larger the training set used, the greater the efficiency gains with ALL metric. In effect, RUN3 used 2234 pairs, RUN2 used 1500 pairs and RUN1 was composed by 750 pairs. This highlights the need for larger datasets for the purpose of building more accurate models.

With ALLnrm our system achieved better results but since this metric is based on normalized Pearson correlation which assumes a linear correlation, we believe that this metric is not representative of the underlying phenomenon. For example, conducting manual observation we can see that pairs from SMT-news are much harder to classify than MSRvid pairs. This results can also be evidenced from others participating teams who almost always achieved better results with MSRvid than SMT-news dataset.

The last metric proposed is the Mean and we are ranked 29 among participating teams. It is probably due to the weight of SMT-news (399 pairs) is smaller than MSR or MSRvid.

Mean metrics seems to be more suitable for this task but lack an important issue, do not have into account the different "complexity" of the datasets. It is also a issue for all metrics proposed. We believe that incorporating to Mean metric a complexity factor weighting for each dataset based on a human judge assignment could be more suitable for the STS evaluation. We think in complexity as an underlying concept referring to the difficulty of determine how semantically related two sentences are to one another. Thus, two sentences with high lexical overlap should have a low complexity and instead two sentences that requires deep inference to determine similarity should have a high complexity. This should be heighted by human annotators and could be a method for a more precise evaluation of STS systems.

Finally, we suggested measuring this new challenging task using a weighted Mean of the Spearman's rho correlation coefficient by incorporating a factor to weigh the difficulty of each dataset.

Run	ALL	Rank	ALLnrm	Rank Nrm	Mean	Rank Mean	MSR par	MSR vid	SMT-eur	On-WN	SMT-news
Best Run	,8239	1	,8579	2	,6773	1	,6830	,8739	,5280	,6641	,4937
Worst Run	-,0260	89	,5933	89	,1016	89	,1109	,0057	,0348	,1788	,1964
Sagan-RUN1	,5522	57	,7904	47	,5906	29	,5659	,7113	,4739	,6542	,4253
Sagan-RUN2	,6272	42	,8032	37	,5838	34	,5538	,7706	,4480	,6135	,3894
Sagan-RUN3	,6311	39	,7943	45	,5649	46	,5394	,7560	,4181	,5904	,3746

Table 5. Official results of the STS challenge

5 Conclusions and future work

In this paper we present Sagan, an RTE system applied to the task of Semantic Textual Similarity. After a preliminary study of the classifiers performance for the task, we decided to use a combination of STS datasets for training and the classifier SVM with regression. With this setup the system was ranked 39 in the best run with overall Pearson, and ranked 29 with Mean metric. However, both rankings are based on the Pearson correlation coefficient and we believe that this coefficient is not the best suited for this task, thus we proposed a Mean Spearman's rho correlation coefficient weighted by complexity, instead. Therefore, further application of other metrics should be one in order to find the most representative and fair evaluation metric for this task. Finally, while promising results were obtained with our system, it still needs to be tested on a diversity of settings. This is work in progress, as the system is being tested as a metric for the evaluation of machine translation, as reported in (Castillo and Estrella, 2012).

References

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. *Accelerated DP Based Search For Statistical Translation*. In Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH-97).

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 AnnualMeeting of the Association for Computational Linguistics(ACL-02), pages 311–318.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. *A Evaluation Tool for Machine Translation:Fast Evaluation for MT Research*. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000).

G. Doddington. 2002. *Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics.* In Proceedings of the 2nd International Conference on Human Language Technology Research (HLT-02), pages 138–145, San Francisco, CA, USA.

Satanjeev Banerjee and Alon Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.* In Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05), pages 65–72.

Michael Denkowski and Alon Lavie. 2011. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR.

He Yifan, Du Jinhua, Way Andy, and Van Josef . 2010. *The DCU dependency-based metric in WMT-MetricsMATR 2010*. In: WMT 2010 - Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, ACL, Uppsala, Sweden.

Chi-kiu Lo and Dekai Wu. 2011. *MEANT: inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles*. 49th Annual Meeting of the Association for Computational Linguistic (ACL-2011). Portland, Oregon, US.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA-06), pages 223–231.

Ido Dagan, Oren Glickman and Bernardo Magnini. 2006. *The PASCAL Recognising Textual Entailment Challenge*. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.) Machine Learning Challenges. Lecture Notes in Computer Science , Vol. 3944, pp. 177-190, Springer.

Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. *Source-language entailment modeling for translating unknown terms*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL. Stroudsburg, PA, USA, 791-799.

Wilker Aziz and Marc Dymetmany and Shachar Mirkin and Lucia Specia and Nicola Cancedda and Ido Dagan. 2010. *Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-VocabularyWords*. In: Proceedings of the 14th annual meeting of the European Association for Machine Translation (EAMT), Saint-Rapha, France.

Dahlmeier, Daniel and Liu, Chang and Ng, Hwee Tou. 2011.TESLA at WMT 2011: Translation Evaluation and Tunable Metric.In: Proceedings of the Sixth Workshop on Statistical Machine Translation. ACL, pages 78-84, Edinburgh, Scotland.

S. Pado, D. Cer, M. Galley, D. Jurafsky and C. Manning. 2009. *Measuring Machine Translation Quality as Semantic Equivalence: A Metric Based on Entailment Features*. Journal of MT 23(2-3), 181-193.

S. Pado, M. Galley, D. Jurafsky and C. Manning. 2009a. *Robust Machine Translation Evaluation with Entailment Features*. Proceedings of ACL 2009.

Eneko Agirre, Daniel Cer, Mona Diab and Aitor Gonzalez-Agirre. 2012. *SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity*. In Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), in conjunction with the First Joint Conference on Lexical and Computational Semantics (*SEM 2012).

Bentivogli, Luisa, Dagan Ido, Dang Hoa, Giampiccolo, Danilo, Magnini Bernardo.2009.*The Fifth PASCAL RTE Challenge*. In: Proceedings of the Text Analysis Conference.

Fellbaum C. 1998. *WordNet: An Electronic Lexical Database*, volume 1. MIT Press.

Castillo Julio. 2011. *A WordNet-based semantic approach to textual entailment and cross-lingual textual entailment*. International Journal of Machine Learning and Cybernetics - Springer, Volume 2, Number 3.

Castillo Julio and Cardenas Marina. 2010. *Using sentence semantic similarity based onWordNet in recognizing textual entailment*. Iberamia 2010. In LNCS, vol 6433. Springer, Heidelberg, pp 366–375.

Castillo Julio. 2010. *A semantic oriented approach to textual entailment using WordNet-based measures*. MICAI 2010. LNCS, vol 6437. Springer, Heidelberg, pp 44–55.

Castillo Julio. 2010. *Using machine translation systems to expand a corpus in textual entailment*. In: Proceedings of the Icetal 2010. LNCS, vol 6233, pp 97–102.

Resnik P. 1995. *Information content to evaluate semantic similarity in a taxonomy*. In: Proceedings of IJCAI 1995, pp 448–453 907.

Castillo Julio, Cardenas Marina. 2011. *An Approach to Cross-Lingual Textual Entailment using Online Machine Translation Systems*. Polibits Journal. Vol 44.

Castillo Julio and Estrella Paula. 2012. Semantic *Textual Similarity for MT evaluation*. NAACL 2012 Seventh Workshop on Statistical Machine Translation. WMT 2012, Montreal, Canada.

Lin D. 1997. *An information-theoretic definition of similarity*. In: Proceedings of Conference on Machine Learning, pp 296–304 909.

Jiang J, Conrath D.1997. *Semantic similarity based on corpus statistics and lexical taxonomy*. In: Proceedings of theROCLINGX 911

Pirro G., Seco N. 2008. *Design, implementation and evaluation of a new similarity metric combining feature and intrinsic information content*. In: ODBASE 2008, Springer LNCS.

Wu Z, Palmer M. 1994. *Verb semantics and lexical selection*. In: Proceedings of the 32nd ACL 916.

Leacock C, Chodorow M. 1998. *Combining local context and WordNet similarity for word sense identification*. MIT Press, pp 265–283 919

Hirst G, St-Onge D . 1998. *Lexical chains as representations of context for the detection and correction of malapropisms*. MIT Press, pp 305–332 922

Banerjee S, Pedersen T. 2002. *An adapted lesk algorithm for word sense disambiguation using WordNet*. In: Proceeding of CICLING-02

William B. Dolan and Chris Brockett.2005. *Automatically Constructing a Corpus of Sentential Paraphrases*. Third International Workshop on Paraphrasing (IWP2005). Asia Federation of Natural Language Processing.