SemantiKLUE: Semantic Textual Similarity with Maximum Weight Matching

Nataliia Plotnikova and **Gabriella Lapesa** and **Thomas Proisl** and **Stefan Evert**
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Professur für Korpuslinguistik
Bismarckstr. 6, 91054 Erlangen, Germany
{nataliia.plotnikova,gabriella.lapesa,thomas.proisl,stefan.evert}@fau.de

Abstract

This paper describes the SemantiKLUE system (Proisl et al., 2014) used for the SemEval-2015 shared task on Semantic Textual Similarity (STS) for English. The system was developed for SemEval-2013 and extended for SemEval-2014, where it participated in three tasks and ranked 13th out of 38 submissions for the English STS task. While this year's submission ranks 46th out of 73, further experiments on the selection of training data led to notable improvements showing that the system could have achieved rank 22 out of 73. We report a detailed analysis of those training selection experiments in which we tested different combinations of all the available STS datasets, as well as results of a qualitative analysis conducted on a sample of the sentence pairs for which SemantiKLUE gave wrong STS predictions.

1 Introduction

The SemEval-2015 task on ”Semantic Textual Similarity for English” (Agirre et al., 2015) is a rerun of the corresponding task from SemEval-2014 with new test data and updated categories. The predictions of participating systems were evaluated against manually annotated and subsequently filtered data. STS was measured on a scale ranging from 0 (no similarity at all) to 5 (total equivalence). SemantiKLUE, developed in 2014, uses a distributional bag-of-words model as well as a word-to-word alignment for each pair of sentences based on a maximum weight matching algorithm.

Our SemEval-2015 submission for all 5 test categories (headlines, images, belief, answers-forums, answers-students) was based on the training data set from 2014 with 2234 sentence pairs from 3 categories, namely paraphrase sentence pairs (MSRpar), sentence pairs from video descriptions (MSRvid) and MT evaluation sentence pairs (SMTeuroparl). Follow up experiments conducted after the submission deadline showed us that this training configuration was far from optimal, and that our system would have benefited a lot from a better training, as we managed to significantly improve the overall scores. With the best training configuration, SemantiKLUE would have ranked 22nd out of 73 submissions (11th out of 28 teams), with a weighted mean of Pearson correlation coefficients over all test categories of 0.7508 (best system: 0.8015)

In the following sections, we first give a short overview of the system (Section 2), and then we describe the follow-up experiments that allowed us to define the best training data set in terms of its subsets (Section 3); finally, we present the results of a qualitative analysis of the performance of our system (Section 4).

2 System Description

SemantiKLUE combines supervised and unsupervised approaches for the computation of textual similarity: a number of similarity measures are computed and passed to a support vector regression learner, which is trained on the available training data and test sets of previous years. The learnt weights are then used to generate semantic similarity scores for the test data in the desired range.

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 111–116,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

2.1 Training Data and Preprocessing

The system was trained on manually annotated sentence pairs from the STS task at SemEval 2014. All sentence pairs were preprocessed with Stanford CoreNLP[1] for part-of-speech annotation and lemmatization. Each sentence was represented as a graph using the CCprocessed variant of the Stanford Dependencies (collapsed dependencies with propagation of conjunct dependencies) implemented with the NetworkX[2] module. This graph representation was involved in the computation of all 39 similarity measures for words and tokens in each sentence. Prepositions, articles, conjunctions as well as auxiliary verbs like *be* and *have* were ignored in the computation of token-based measures.

2.2 Similarity Measures: Overview

A detailed description of all 39 similarity measures used as features in SemantiKLUE is provided in Proisl et al., 2014 (Sections 2.2 - 2.7). Similarity measures used by our system include:

- **Heuristic similarity measures**: word form overlap and lemma overlap between two texts computed with Jaccard coefficient; difference in text length used by Gale and Church (1993); a binary feature to treat negation in each sentence pair.
- **Document similarity measures** based on two distributional models: a model based on non-lemmatized information, built from the second release of the Google Books N-Grams database (Lin et al., 2012); a lemmatized model, built from a 10-billion word Web corpus[3].
- **Alignment-based measures**: one-to-one alignment and one-to-many alignment for both words and lemmata, computed via maximum weight matching, based on cosine similarity between two words in paired sentences as edge weight. Figure 1 visualizes a one-to-many alignment based on lemmatized data. The colors of the connections correspond to different cosine ranges, reported in the legend to the right of the plot.
- **WordNet-based similarity measures**: Leacock and Chodorow's (1998) normalized path length and Lin's (1998) universal similarity measure. Using these similarity measures, the best one-to-one and the best one-to-many alignment are computed. After that, the arithmetic mean of the similarities between the aligned words from text A and text B with and without identical word pairs is calculated. An additional WordNet-based feature is the number of unknown words in both texts.

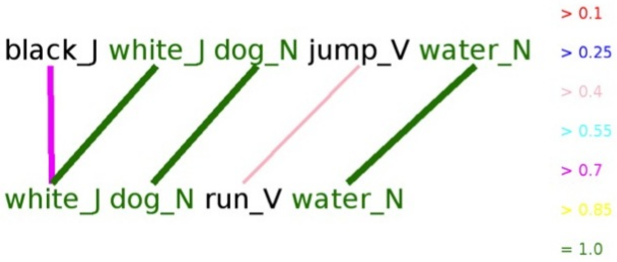

Figure 1: One-to-many alignment plot. Sentences: "A black and white dog is jumping into the water" , "A white dog runs across the water"; Subset: Images; Gold Score: 2.8; SemantiKLUE score: 2.93.

- **Dependency-based heuristic measures**: overlap of dependency relation labels between the two texts; arithmetic mean of the similarities between the best aligned one-to-one dependency relations based on Leacock and Chodorow's normalized path lengths; average overlap of neighbors for all aligned word pairs based on one-to-one alignment created with similarity scores from the lemma-based DSM.
- **Experimental features**: cosine similarities for each pair of sentences; average neighbor rank based on the rank of text A among the nearest neighbors from text B and vice versa.

The feature set described above was processed by the support vector regressor implemented in the scikit-learn[4] (Pedregosa et al., 2011) library. All the experiments presented in this paper rely on the best support vector setting identified by Proisl et al. (2014), namely: RBF kernel of degree 2 and penalty C = 0.7. In what follows, we describe the procedures adopted to adjust training data and find the best training configurations.

3 Experiments

This section describes all post-hoc experiments on the STS 2015 test data performed to improve the

[1] http://nlp.stanford.edu/software/corenlp.shtml

[2] http://networkx.github.com

[3] Wackypedia and UkWaC (Baroni et al., 2009), UMBC WebBase (Han et al., 2013), and UKCOW 2012 (Schäfer and Bildhauer, 2012).

[4] http://scikit-learn.org/

predictions of the system. The abbreviations used in the following tables reporting experiment results are listed in Table 1.

short	full name	source
mp	MSRpar[5]	train 2014
mv	MSRvid[6]	train 2014
smt	SMTeuroparl[7]	train 2014
img	images[8]	test 2014
hl	headlines[9]	test 2014
ow	OnWN[10]	test 2014
df	deft-forum[11]	test 2014
dn	deft-news[12]	test 2014
tn	tweet-news	test 2014
fn	FNWN [13]	test 2013
ans-f	answers-forums	test 2015
ans-s	answers-students	test 2015
head	headlines	test 2015

Table 1: Training set categories: abbreviations.

All 39 similarity measures were used by the regression learner to train the system. SemantiKLUE was tested on different training data with various combinations of training and test sets from 2013 and 2014. Results for the submitted system are typeset in italics in Table 2, the best results in each column are typeset in bold font.

The best results would have been obtained by training on the MSR data from SemEval 2014 for all test sets. Considerable improvements can be achieved removing the SMTeuroparl category from the training set. This category consists of MT pairs of sentences whose exclusion would have given the system rank 37 (weighted mean of .7148) instead of 46 (.6717) out of 73 submissions.

We turned the test data from SemEval 2014 into a training set for the 2015 test data (see Table 3). The figures in Table 3 show that training sets for images and headlines perform best with the corresponding categories of the test set (images and headlines) from SemEval 2014.

	ans-f	ans-s	head	belief	images	mean
mp	-.2533	.5944	.4515	.3102	.6497	.4310
mv	.3262	.5990	.6044	.5021	.7879	.6014
smt	.2603	.5263	.4073	.3177	.4715	.4235
mp + mv	**.5509**	**.7259**	**.7009**	**.6961**	**.8088**	**.7148**
mv + smt	.4891	.6849	.6822	.5658	.7991	.6734
smt + mp	-.0893	.4989	.2947	.1296	.3781	.2980
mp + mv + smt	*.4913*	*.7005*	*.6681*	*.5617*	*.7915*	*.6717*

Table 2: Evaluation results for different training sets from 2014.

	ans-f	ans-s	head	belief	images	mean
img	.2673	.6549	.6574	.5669	**.8180**	.6367
hl	.5760	**.6760**	**.7734**	.6439	.7249	**.6960**
ow	.3446	.6661	.5960	.5386	.7334	.6093
df	.3743	.5884	.5618	.6023	.5818	.5551
dn	.2620	.6746	.5765	.5804	.7246	.5992
tn	**.6484**	.6134	.6968	**.6858**	.7018	.6698

Table 3: Evaluation results for different training sets based on the 2014 test categories.

STS results appear to be extremely sensitive to the choice of the training dataset. For this reason, we conducted more fine-grained experiments to look for the best combination of training data for the 2015 test sets. We combined training and test data of SemEval 2014 with the best training categories of SemEval 2013 (see Table 4) to test the performance of the system on the optimal training subset defined for SemEval 2014[14]. That optimal training configuration consists of the FNWN, headlines, MSR and OnWN data sets: the corresponding performance is typeset in italics. Comparable or even better results can be achieved with a combination of test and train categories of SemEval 2014 only. Thus, combining the training category MSR (*mp + mv*) with another test category of 2014 (such as tweets or headlines) results in about 1.5%-2% improvement. A more precise investigation helped us to find the best test combination with MSR, headlines, images, and tweet-news categories. This brought our system to the weighted mean of .7508, corresponding to the 11th place out of 28 teams. We tried to further improve these results, by adding the optimal categories for training found in 2014 and extended the best training set defined for 2015 with FNWN (*mp+mv+hl+img+tn+fn*), but this led to slightly worse results in all test categories.

A further set of experiments was aimed at testing different subsets of similarity measures used at the

[5] Microsoft Research Paraphrase Corpus.

[6] Microsoft Research Video Description Corpus.

[7] WMT2008 development dataset.

[8] Image descriptions from the Flickr dataset.

[9] Headlines mined from news sources.

[10] Sense definitions from OntoNotes and WordNet .

[11] Forum posts.

[12] News summaries.

[13] Sense definitions from FrameNet and WordNet

[14] For space reasons we list only the combinations resulting in the best scores. Combinations with SMTeuroparl, for example, led to consistently worse results and are therefore left out.

	answers-forums	answers-students	headlines	belief	images	mean
img+hl	.5119	.6995	.7663	.6296	.8262	.7157
tn+img	.6158	.6949	.7354	.6982	.8187	.7265
tn+hl	.6313	.6625	.7736	.6887	.7350	.7078
tn+mp+mv	**.6460**	.7213	.7462	**.7118**	.8136	.7400
tn+hl+img	.6223	.7028	.7682	.7004	.8247	.7392
mp+mv+img	.4853	.7297	.7110	.6596	.8302	.7108
mp+mv+hl	.6246	**.7336**	.7766	.7057	.8210	.7491
mp+mv+hl+fn	.5426	.7335	**.7775**	.6664	.8147	.7326
mp+mv+tn+hl	.6458	.6961	.7734	.7106	.8180	.7414
mp+mv+tn+img	.6319	.7292	.7434	.7076	.8269	.7423
mp+mv+tn+fn	.5891	.7212	.7459	.6895	.8087	.7288
mp+mv+img+fn	.3337	.6693	.4005	.5791	.7756	.5755
mp+mv+ow+fn+hl	*.5906*	*.7225*	*.7600*	*.6762*	*.8135*	*.7324*
mp+mv+hl+img+tn	.6341	.7325	.7686	.7067	**.8315**	**.7508**
mp+mv+hl+img+tn+fn	.5931	.7313	.7684	.6869	.8291	.7422

Table 4: Evaluation results for different training sets based on train and test categories of 2014 and 2013.

	answers-forums	answers-students	headlines	belief	images	mean
token (one to one)	.5377	.6483	.6393	.6663	.6608	.6375
token (one to many)	.3930	.6566	.5744	.5449	.5901	.5725
lemma (one to one)	**.6423**	.6484	**.6610**	**.7075**	**.7774**	**.6904**
lemma (one to many)	.6043	**.6749**	.6082	.6777	.7469	.6677

Table 5: Single-feature experiments with different alignments: correlation based on cosine similarity.

	img	hl	ow	df	dn	tn
img	*.8689*	.6141	.6767	.3363	.4479	.5183
hl	.7249	*.8173*	.6754	.4179	.6028	.6763
ow	.7039	.5707	*.8926*	.3790	.5666	.5760
df	.5497	.4931	.5969	*.7818*	.5193	.4836
dn	.6957	.5582	.6428	.4008	*.8588*	.3935
tn	.6823	.6453	.6321	.3816	.5222	*.8697*

Table 6: Test data categories of 2014 against each other (columns = training sets, lines = test sets).

machine learning stage. Results showed that the use of fewer similarity features (exclusion of all identical words in each pair of sentences from the calculation of similarity scores) resulted in worse performance of the whole system.

Our system is based on a relatively large feature set, but we were also interested in discovering how well SemantiKLUE would have performed if trained on a single feature. We tested a feature based on cosine similarity between the two centroid vectors as a measure of semantic similarity for each sentence pair as suggested by Schütze (1998) using either tokens or lemmas (see Table 5). We selected cosine between centroid vectors as a candidate feature, because it is most intuitive and naturally connects to the representation of topical information, crucial in capturing textual similarity.

We found that regardless of the alignment (one to one or one to many both for lemma and tokens), the weighted mean of Pearson correlation coefficients is low (.6904 for the one-to-one alignment) for the cosine similarity value calculated with lemma based centroid vectors, but still higher than what is achieved by the more complex system with a large set of features with a poor training set (.6717) in the submission with *mp+mv+smt* used for the training set (see Table 4 for comparison).

As we were interested in identifying the most balanced training sets in the test categories of 2014, we tested all categories against each other. Results are shown in Table 6: the rows of the table correspond to test subsets, while columns represent training sets. The results typeset in italics show that there is a high level of overtraining for the cases in which training and test data are identical. The most balanced and robust test data are those of the image and OnWN categories: they can be used as training data for future experiments.

To sum up, our results show that the best training configuration for SemEval 2015 involves **MSR, headlines, images, and tweet-news categories** (see Table 4). The scatter plots in Figures 2 to 4 relate the similarity score in the gold standard (*x-axis*) to the relatedness score produced by SemantiKLUE (*y-axis*) in its best training configuration, for three of

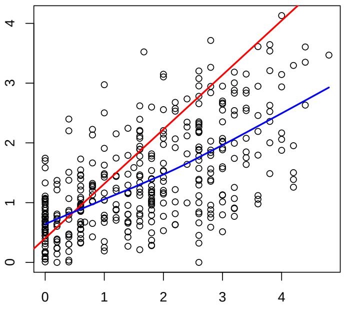

Figure 2: Answers Forums.

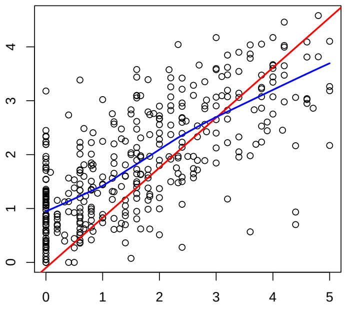

Figure 3: Belief.

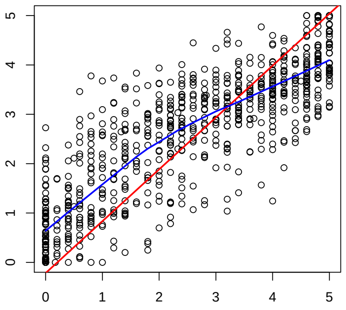

Figure 4: Images.

the five Semeval 2015 test sets. For each plot we show the regression line (drawn in red) as well as a smoother, drawn (in blue) with the LOWESS function from R[15]. Smoothed lines show different nonlinear patterns for the different subsets.

4 Qualitative Analysis

In this section, we report the results of a qualitative analysis conducted on sentence pairs for which SemantiKLUE, in the optimal training configuration identified in Section 2.2, made wrong predictions.

Our goal was to identify a taxonomy of SemantiKLUE's problems. Broadly speaking, there are two possibilities for SemantiKLUE to make a wrong similarity guess: the system can **overestimate** the similarity between the two sentences - thus generating a relatedness score higher than the speakers' judgments - or it can **underestimate** similarity - generating a score lower than the gold standard. In the process of interpretation/classification, we relied on the inspection of alignment plots (cf. Figure 1) and on our knowledge of the dynamics of the features within SemantiKLUE.

The analysis was conducted manually on a selected sample of sentence pairs from the test data. We selected sentences for which the absolute difference between the similarity score in the gold standard and the relatedness score produced by SemantiKLUE was between 1.5 and 2.5 points. That range was identified by inspecting the distribution of gold standard/relatedness score differences in the five subsets (corresponding plots are not shown here for reasons of space). Within this range, we randomly picked 40 items (sentence pairs) per subset, 20 with positive difference (underestimation), 20 with negative difference (overestimation)[16].

Let us start with the cases in which SemantiKLUE overestimated STS. We list the identified mistake categories, providing a short description for the cases in which the label is not self-explanatory, and report the percentage of affected sentences. Each item can be affected by more than one mistake type.

- **One or two words** (often very frequent and with generic meaning) **dominate the alignment**, or one sentence is practically a subset of the other: **56%** of the items.
- **Wrong alignments**: **25%** of the items.
- **Modification**: presence of identical modifiers with different heads boosts overall similarity. This mistake type affects **7%** of the cases.
- **Same frame, different participants**: the sentences depict the same event, but the participants (or the background) determine a significant difference in meaning that our system fails to capture. This problem affects **8%** of the items.
- **Same participants, different frames**: **11%** of the items.
- **Negation**: **10%** of the items.
- **(Near) Antonyms**: **8%** of the items.
- **Proper Names**: **18%** of the items.
- **Amounts**: when building the alignment, SemantiKLUE ignores numerical values, which are in some cases crucial in determining (dis)similarities between sentences otherwise near identical (e.g., "2 people killed.." vs. "100 people killed"). This problem affects **18%** of the items.

We now proceed to cases of underestimation, for which we identified the following mistake types:

[15] http://www.r-project.org/

[16] In two cases, we had to enlarge the range to ensure that at least 20 items would have been selected: belief/positive, between 1.4 and 3.5; answers-forums/negative, between 1 and 2.5.

- **Collocations** (e.g, “heads up”, “make sense”) negatively affect the alignment process: SemantiKLUE would have performed better if multiwords had entered the alignment process as a whole, and not as individual edges. This mistake type affects **10%** of the items.
- **Crucial alignments missing or weaker than expected**: **17%** of the items.
- **The similarity between the sentences is due to logical form, compositionality or world knowledge**. This problem affects **16%** of the items.
- **Different register** makes alignment problematic, even if the sentences are content-wise similar: **12%** of the items.
- **Displacement of different pieces of information between two sentences otherwise centered on the same topic** makes them less similar for SemantiKLUE then for the raters: **28%** of the items.
- **Spelling mistakes** prevent otherwise straightforward alignments: **10%** of the items.
- **Difficult cases**, for which the alignment would simply suggest a score higher than the one predicted by the regressor. Such cases, (**15%**), require further investigation.

5 Conclusion

In this paper, we presented the results of our evaluation experiments on the performance of the SemantiKLUE system (Proisl et al., 2014) on the SemEval-2015 STS task. Our experiments showed that the performance of our system is heavily dependent on the choice of the training set, as we managed to significantly improve the performance of our system with respect to the original submission. The qualitative evaluation sketched in Section 4 provided interesting insights into specific features of the STS data and it allowed us to identify some idiosyncracies (e.g., the behavior of the system in case of alignment of identical words) and weaknesses (e.g., the treatment of multiwords in the process of alignment) that we are already working on improving.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, 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.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Lushan Han, Abhay L. Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. 2013. UMBC_EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*. Atlanta, GA.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 265–283. MIT Press, Cambridge, MA.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 296–304. San Francisco, CA.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the Google Books Ngram Corpus. In *Proceedings of the ACL 2012 Syst. Demonstrations*, pages 169–174, Jeju Island, Korea.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Douard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Thomas Proisl, Stefan Evert, Paul Greiner, and Besim Kabashi. 2014. SemantiKLUE: Robust semantic similarity at multiple levels using maximum weight matching. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 532–540. Dublin, Ireland.

Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC ’12)*, pages 486–493. Istanbul, Turkey.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.