DTSim at SemEval-2016 Task 2: Interpreting Similarity of Texts Based on Automated Chunking, Chunk Alignment and Semantic Relation Prediction

Rajendra Banjade*, Nabin Maharjan*, Nobal B. Niraula, Vasile Rus
Department of Computer Science / Institute for Intelligent Systems
The University of Memphis
Memphis, TN, USA
{rbanjade, nmharjan, nbnraula, vrus}@memphis.edu

Abstract

In this paper we describe our system (DTSim) submitted at SemEval-2016 Task 2: Interpretable Semantic Textual Similarity (iSTS). We participated in both gold chunks category (texts chunked by human experts and provided by the task organizers) and system chunks category (participants had to automatically chunk the input texts). We developed a Conditional Random Fields based chunker and applied rules blended with semantic similarity methods in order to predict chunk alignments, alignment types and similarity scores. Our system obtained F1 score up to 0.648 in predicting the chunk alignment types and scores together and was one of the top performing systems overall.

1 Introduction

Measuring the semantic similarity of texts is to quantify the degree of semantic similarity between a given pair of texts, such as two words or two sentences (Rus et al., 2008; Agirre et al., 2015). For example, a similarity score of 0 means that the texts are not similar at all and 5 means that they have same meaning. While useful, such quantitative or even qualitative assessments are hard to interpret because they do not provide details, i.e. they do not explain or justify why the similarity score was assigned high or low. One way to provide an explanatory layer to text similarity assessment methods is to align chunks between texts and assigning semantic relation to each alignment. To this end, Brockett (2007) and Rus et al. (2012) produced datasets where corresponding words (or multiword expressions) were aligned and in the latter case their semantic relations were explicitly labeled

In interpretable Semantic Textual Similarity (iSTS) tasks (Agirre et al., 2016), participants had first to identify the chunks in each sentence (sys chunks) or use the chunks given by the task organizers (gold chunks), and then, align chunks across the two sentences indicating the semantic relation and similarity score of each alignment. The chunk alignment types were EQUI (semantically equivalent), OPPO (opposite in meaning), SPE (one chunk is more specific than other), SIMI (similar meanings, but not EQUI, OPPO, SPE), REL (related meanings, but not SIMI, EQUI, OPPO, SPE), and NOALI (has no corresponding chunk in the other sentence). The relatedness/similarity scores were assigned in the range of 0 to 5.

A pilot on iSTS task was organized in 2015 (Agirre et al., 2015). In 2016, the iSTS allowed many-to-many chunk alignment while in the pilot task of 2015 they only allowed one-to-one or no alignment. Also, a new dataset consisting of student answers given to a tutoring system was added in 2016. For further details about the task, please see Agirre et al. (2016).

We participated in both categories: system chunks and gold chunks. Our system first preprocesses texts, creates chunks (in sys chunks category) using our own chunking tool, and performs alignments and labels them with semantic relations and similarity scores. In this paper, we describe our system DTSim[1] and the submitted three different runs in the

* These authors contributed equally to this work

[1] Available for download from http://semanticsimilarity.org

Proceedings of SemEval-2016, pages 809–813,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

shared task. Our system was one of the top performing systems.

2 Preprocessing

Hyphens were replaced with whitespaces if they were not composite words (e.g. video-gamed). Also, the words starting with co-, pre-, meta-, multi-, re-, pro-, al-, anti-, ex-, and non- were left intact. Then, the texts were tokenized, lemmatized, POS-tagged and annotated with Named Entity (NE) tags using Stanford CoreNLP Toolkit (Manning et al., 2014). We also marked each word as whether it was a stop word. In the system chunks category, we had plain texts and we created chunks using our own Conditional Random Fields (CRF) based chunking tool (see Section 3). We normalized texts using mapping data. For example, *pct* and % were changed to *percent*. These preprocessing steps were performed for both gold chunks and system chunks category.

In student-answers dataset which consists of student answers given to a computer based logic tutor (Agirre et al., 2016), we replaced symbol *A/B/C* with *bulb A/B/C*. Similarly, *X/Y/Z* was replaced by *switch X/Y/Z*. We used this domain knowledge based on the notes found in student-answers training data description file.

3 Chunking

We developed a CRF based chunker[2] using both CoNLL-2000 shared task training and test data[3]. This data consists of a Wall Street Journal corpus: sections 15-18 as training data (211727 tokens) and section 20 as test data (47377 tokens). We generated shallow parsing features such as previous and next words from current word, current word itself, current word POS tag, previous and next word POS tags and their different combinations as described in Tjong Kim Sang and Buchholz (2000). We used CRF++ tool[4] to build the CRF models.

Furthermore, we analyzed its output (i.e. chunks) and added the following rules in the system to merge some of the chunks, resulting in chunks that make more sense and are consistent with iSTS gold chunks.

DataSet	**Chunker**	**CL**	**SL**
Training data			
Headlines	O-NLP	53.74	13.49
	EO-NLP	80.67	59.39
	CRF	**82.60**	**62.56**
Image	O-NLP	52.35	5.06
	EO-NLP	89.13	72.66
	CRF	**89.74**	**74.13**
Test data			
Headlines	O-NLP	53.88	16.13
	EO-NLP	80.96	60.18
	CRF	**83.32**	**63.23**
Image	O-NLP	52.71	5.33
	EO-NLP	89.30	72.13
	CRF	**90.29**	**74.93**

Table 1: Comparison of chunking accuracies of the various chunkers at chunk level (CL) and at sentence level (SL) with iSTS 2015 gold data.

(a) PP + NP => PP
(b) VP + PRT => VP
(c) NP + CC + NP => NP

For example, it merges chunks [on] and [Friday] to form single PP chunk [on Friday] using rule (a).

We evaluated the chunking accuracy of the CRF chunker by comparing its output against the gold chunks of iSTS 2015 data: the training and test data sets each consist of 375 pairs of Images annotation data and 378 pairs of Headlines texts. This chunker yielded the highest average accuracies on both the training and test datasets compared to other chunkers which are described next. The accuracies on the training dataset were 86.20% and 68.34% at chunk level and sentence level respectively. For the test dataset, the accuracies were 86.81% and 69% at chunk and sentence level, respectively. The results are presented in Table 1.

We also chunked the input texts using the Open-NLP chunking tool (O-NLP). The average (of Images and Headlines data) accuracies were 53.04% at chunk level and a modest 9.27% at sentence level for the training dataset. It yielded similar results on test data. We extended Open-NLP chunker (EO-NLP) using the rules described before. The results were improved and resulted in 84.% chunk level

[2] Our chunker is available at http://semanticsimilarity.org

[3] http://www.cnts.ua.ac.be/conll2000/chunking/

[4] https://taku910.github.io/crfpp/

and 66.02% sentence level average accuracies on the training dataset, respectively. The accuracy on the test data was comparable at 85.13% chunk level and 66.15% at sentence level. However, the results of our CRF based chunker were superior in all cases.

4 Chunk Alignment System

For a given sentence pair, the chunks of the first sentence were mapped to those from the second by assigning different semantic relations and scores based on a set of rules, similarity functions, and lookup resources. Before preforming alignments, we preprocessed texts as described in Section 2.

We built upon a previous system called NeRoSim (Banjade et al., 2015). The limitation of their system was that the alignments were restricted to 1:1. We modified it to support many to many alignments as well. Also, the NeRoSim system was able to process only gold chunks (i.e., chunks provided by the organizers). Now, the system can take input in the form of plain texts as well and create chunks on the fly. In addition to the chunking feature described in Section 3, the updates made to the systems are described below.

Many-to-Many Alignments:
MULTI1: If there is any ALIC chunk (i.e., chunk which does not have any corresponding chunk in the other sentence because of the 1:1 alignment restriction) in sentence A whose content words are subsumed by the content words of any already aligned chunk (C) in another sentence B, merge ALIC chunk with the chunk in A paired with C. If the content words of merged chunk and those of C are same/equal, realign chunk C with merged chunk as EQUI and update the score to 5.0.
For example:
// [Iran] [hopes] [nuclear talks] [. . .].
// [Iran Nuclear Talks] [spur] [. . .].
Step 1:
nuclear talks <=> Iran Nuclear Talks // SPE2
Iran <=> //ALIC
Step2:
Iran nuclear talks <=> Iran Nuclear Talks // EQUI

MULTI2: In *MULTI1*, if all the content words of merged chunk and those of C are not matching completely, then realign chunk C with merged chunk but keep the previous alignment type and score.

Furthermore, we have expanded the rules for SIMI and EQUI.
EQx: If unmatched words are morphological inflections of each other and all other words in the chunks are already matched, assign the EQUI relation.
E.g. *Korean Air <=> Air Korea*
SIMIx: If nouns are matching but not the adjective or vice-versa, assign SIMI label.
E.g. *red carpet <=> brown carpet*

Runs
R1: We included many-to-many alignment in NeRoSim (i.e., MULTI1 and MULTI2 were added).
R2: Same as R1 in alignment but the alignment scores were assigned based on the average scores for each alignment type in the full training data.
R3: Same as R2 but SIMIx and EQx rules added.

5 Results

The test data consisted of 1,094 sentence pairs which included texts from headlines (375), image annotations (375), and student-answers (344). The results (F1 scores)[5] on test datasets are presented in Table 2 and Table 3. Further details about the test data and the evaluation metrics can be found in Agirre et al. (2016).

Table 2 presents the results in terms of F1 scores on test set with gold chunks. We can see that the alignment scores are higher compared to the baseline system and are very close to the best results from all submissions in those categories. However, the alignment type score in each case is relatively lower than the alignment-only score and it ultimately impacted the F1 score calculated for type and score together (i.e. T+S).

Similar to Table 2, the Table 3 presents the results on the test set but this time with system chunks. In image and headlines data, our system obtained the best results. However, following the same pattern as in gold chunk results, the F1 scores for alignments

[5] Based on (Melamed, 1998) which was proposed in the context of alignment for Machine Translation.

System	**A**	**T**	**S**	**T+S**
Headlines				
Baseline	0.8462	0.5462	0.7610	0.5461
R1	0.9072	0.6650	0.8187	0.6385
R2	0.9072	0.6650	0.836	**0.6487**
R3	0.9072	0.6583	0.8329	0.6405
Max	0.9278	0.7031	0.8382	0.6960
Image				
Baseline	0.8556	0.4799	0.7456	0.4799
R1	0.8766	0.6530	0.7955	0.6238
R2	0.8766	0.6530	0.8144	0.6362
R3	0.8766	0.6675	0.8156	**0.6483**
Max	0.9077	0.6867	0.8552	0.6708
Student-answers				
Baseline	0.8203	0.5566	0.7464	0.5566
R1	0.8584	0.5552	0.7686	0.5432
R2	0.8584	0.5552	0.7809	**0.5458**
R3	0.8614	0.5468	0.7798	0.5374
Max	0.8922	0.6511	0.8433	0.6385

Table 2: F1 scores on test data with gold chunks. A, T and S refer to Alignment, Type, and Score, respectively. Max score is the best score for each metric given by any of the participating systems in the shared task including the system submitted by the team involved in organizing the task.

System	**A**	**T**	**S**	**T+S**
Headlines				
Baseline	0.6486	0.4379	0.5912	0.4379
R1	0.8366	0.5605	0.7394	0.5384
R2	0.8366	0.5605	0.7595	**0.5467**
R3	0.8376	0.5595	0.7586	0.5446
Max	0.8366	0.5605	0.7595	0.5467
Image				
Baseline	0.7127	0.4043	0.6251	0.4043
R1	0.8429	0.6148	0.7591	0.5870
R2	0.8429	0.6148	0.7806	0.5990
R3	0.8429	0.6276	0.7813	**0.6095**
Max	0.8557	0.6276	0.7961	0.6095
Student-answers				
Baseline	0.6188	0.4431	0.5702	0.4431
R1	0.8165	0.5157	0.7248	0.5049
R2	0.8165	0.5157	0.7367	**0.5074**
R3	0.8181	0.5112	0.7360	0.5029
Max	0.8166	0.5651	0.7589	0.5547

Table 3: Results on test data with system chunks.

are high but the scores for predicting the alignment types are relatively lower. It indicates that the systems overall performance will be improved greatly if improvements can be made in predicting the alignment types. Also, the scores for student-answers are lower than headlines and image texts and it requires further analysis to fully understand why this is the case. One of the reasons might be that we did not use this dataset while developing the system and no prior information about such dataset was modeled. Additionally, more errors might have been introduced in our NLP pipeline as the texts in this dataset were not standard written texts.

In addition to the difficulty of the task of aligning the chunks and assigning relation types, we found some discrepancies in the annotation which we think induced some errors. For example, *on a sofa <=> on a blue sofa* (#65 in image data), the human annotated label is SIMI but arguably the SPE2 label best describes the relation. Similarly, *in a field <=> in a green field* (#693 in image data), the SPE1 label has been found in the training set but it should be SPE2. In another example (#193 in image data), *A young boy <=> A young blonde girl* has been assigned a label SPE2 in the training data. Though the second chunk gives some additional details, the question is whether we should really compare them (and decide which one is more specific) because these two chunks are referring to different objects and therefore it sounds more like comparing apples and oranges.

6 Conclusion

This paper presented the system DTSim and three different runs submitted in SemEval 2016 Shared Task on Interpretable Textual Semantic Similarity (iSTS). We described our chunking tool and results on gold chunks as well as on system chunks categories. Our system was one of the best systems submitted in the shared task. However, there is room for improvement particularly on assigning alignment labels which we intend to address in future work. Furthermore, the annotated dataset is now quite big and it will certainly be useful in applying alternative approaches to predict chunk alignments and alignment types.

References

Eneko Agirre, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Weiwei Guof, Inigo Lopez-Gazpioa, Montse Maritxalara, Rada Mihalceab, et al. 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)*, pages 252–263.

Eneko Agirre, Aitor Gonzalez-Agirre, (Iñigo) Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, June.

Rajendra Banjade, Nobal B Niraula, Nabin Maharjan, Vasile Rus, Dan Stefanescu, Mihai Lintean, and Dipesh Gautam. 2015. Nerosim: A system for measuring and interpreting semantic textual similarity. *SemEval-2015*, page 164.

Chris Brockett. 2007. Aligning the rte 2006 corpus. *Microsoft Research*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

I Dan Melamed. 1998. Manual annotation of translational equivalence: The blinker project. *arXiv preprint cmp-lg/9805005*.

Vasile Rus, Philip M McCarthy, Mihai C Lintean, Danielle S McNamara, and Arthur C Graesser. 2008. Paraphrase identification with lexico-syntactic graph subsumption. In *FLAIRS conference*, pages 201–206.

Vasile Rus, Mihai Lintean, Cristian Moldovan, William Baggett, Nobal Niraula, and Brent Morgan. 2012. The similar corpus: A resource to foster the qualitative understanding of semantic similarity of texts. In *Semantic Relations II: Enhancing Resources and Applications, The 8th Language Resources and Evaluation Conference (LREC 2012), May*, pages 23–25.

Erik F Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics.