UMDuluth-BlueTeam : SVCSTS - A Multilingual and Chunk Level Semantic Similarity System

Sakethram Karumuri
Viswanadh Kumar Reddy Vuggumudi
Sai Charan Raj Chitirala
Department of Computer Science
University of Minnesota Duluth
{karum006, vuggu001, chiti001}@d.umn.edu

Abstract

This paper describes SVCSTS, a system that was submitted in SemEval-2015 Task 2: Semantic Textual Similarity(STS)(Agirre et al., 2015). The task has 3 subtasks viz., English STS, Spanish STS and Interpretable STS. SVCSTS uses Monolingual word aligner (Sultan et al., May 2014), supervised machine learning, Google and Bing translator API's. Various runs of the system outperformed all other participating systems in Interpretable STS for non-chunked sentence input.

1 Introduction

Semantic Textual Similarity gives a quantifier to evaluate semantic equivalence between two sentences. Earlier SemEval tasks (Agirre et al., 2012), (Agirre et al., 2013), (Agirre et al., 2014) focused on finding the semantic equivalence between sentences in English and Spanish. A new pilot task was introduced this year to find which parts (chunks) of the sentences are equivalent in meaning.

SVCSTS is an extension to (Sultan et al., 2014) and it handles both Spanish STS and Interpretable STS. SVCSTS uses Monolingual word aligner (Sultan et al., May 2014), supervised machine learning techniques, Google and Bing translator API's.

Section 2 describes a brief overview of SVCSTS's approach for various subtasks. Section 3 outlines the performance of SVCSTS in various subtasks of SemEval 2015 Task-2.

2 System Description

Following 3 sub sections describe SVCSTS's approach for the 3 subtasks.

2.1 English STS

This task was about finding the semantic similarity between English sentences. (Sultan et al., 2014) system was used to find the semantic equivalence between two sentences and a score on a scale of 0-5 was given.

2.2 Spanish STS

Spanish STS is built upon English STS to calculate similarity scores for a given pair of Spanish sentences on a scale of 0 to 4. Spanish sentences were translated to English, fed to English STS system and the scores are scaled accordingly. Translations were done using Bing Translator API (Bing Translator API) and Google Translate API. Two translators were used to improve the accuracy of the translations.

Google Translate API was obtained from (Kashyap et al., 2014). We used this system to get multiple translations of each chunk in a sentence. Multiple sentences are generated by combining the top two translations of each chunk. We then randomly pick a maximum of ten sentences for each Spanish sentence. Translation pairs are formed by choosing corresponding numbered sentences from sentence 1 and sentence 2 translations. We limited the number of translations to 10 to reduce the overall computation time.

Translation pairs were then passed to English STS system. Final score was obtained as the average

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 107–110,
Denver, Colorado, June 4-5, 2015.

taken from all translation pairs for a given Spanish sentence pair and the score is scaled accordingly.

2.3 Interpretable STS

Existing STS systems report similarity for a pair of sentences.

This is a pilot task where the challenge is to find the semantic relationships between the chunks of sentence 1 and sentence 2. Chunks from the input sentence pair are to be aligned, labeled with the type (described here) of alignment and are to be scored on a scale of 0-5 based on their semantic similarity.

The type of alignments defined in the task description are:

1. EQUI : both chunks are semantically similar.
2. OPPO : both chunks are semantically opposite.
3. SPE1 : both chunks are semantically similar but chunk1 has more information.
4. SPE2 : both chunks are semantically similar but chunk2 has more information.
5. SIMI : similar chunks but no EQUI, OPPO, SPE1 or SPE2.
6. REL : related chunks but no SIMI, EQUI, OPPO, SPE1, SPE2.
7. ALIC : when 1:1 alignment of chunks is not possible extra chunks are given ALIC
8. NOALI: a chunk has no corresponding semantically similar chunk

There are two variations in the input for this subtask:

1. Raw input - Plain sentences are provided and the system has to identify the chunks
2. Chunked input - Chunked sentences are provided by the task organizers

2.3.1 Identifying Chunks

OpenNLP chunker was used to chunk the input sentences and some post processing was done. For the post processing we observed a few rules from gold standard chunks. Those rules include combining chunks of specific chunk tags given by OpenNLP chunker. A large number of rules were discovered but the following were the rules, which maximized accuracy.

- PP + NP + PP + NP
- PP + NP
- VP + PRT
- NP + O + NP
- VP + ADVP
- VP + PP + NP + O
- NP + O

Applying these rules we have increased accuracy from 86.58% to 90.16% against the gold standard chunks.

2.3.2 Aligning Chunks

Monolingual word aligner (Sultan et al., May 2014) was used to find word alignments in the two input sentences. For chunked input, sentences are generated from the chunks prior to running the word aligner. For words aligned their corresponding chunks are aligned.

2.3.3 Labeling Aligned Chunks

Supervised machine learning was performed using Scikit-Learn (scikit-learn). We used the following features for each chunk alignment to assign a type for the alignment.

1. Length of sentence 1 chunk
2. Length of sentence 2 chunk
3. Number of nouns in sentence 1 chunk
4. Number of nouns in sentence 2 chunk
5. Number of verbs in sentence 1 chunk
6. Number of verbs in sentence 2 chunk
7. Number of adjectives in sentence 1 chunk
8. Number of adjectives in sentence 2 chunk
9. Number of prepositions in sentence 1 chunk
10. Number of prepositions in sentence 2 chunk

Type of Alignment	Score
EQUI	5
SPE1	3.75
SPE2	3.55
ALIC	NIL
NOALI	0
SIMI	2.94
REL	2.82
OPPO	4

Table 1: Avg. alignment type scores

Runs	Features Used
Run - 1	3,4,5,6,7,8,9,10,11,12
Run - 2	3,4,5,6,7,8,9,10,11,12,13
Run - 3	1,2,3,4,5,6,7,8,9,10,11,12,13

Table 2: Features used in various runs

11. The path similarity between words of sentence 1 and sentence 2 chunks

12. Unigram overlap between sentence 1 and sentence 2 chunks

13. Bigram overlap between sentence 1 and sentence 2 chunks

We experimented the classification of labels using 3 classifiers LinearSVC, SVC with RBF (Radial Basis Function) Kernel and SVC with Polynomial Kernel. But the classifier SVC with RBF (with parameters C = 1.0, gamma=0.7) proved to give better results.

2.3.4 Scoring Aligned Chunks

Average score for each alignment type was calculated from the gold standard data. The average scores that were used to score chunk alignment are described in Table 1.

2.3.5 Multiple Runs

We tried various combination of features (described in Section 2.3.3) for training the classifier. The details of three runs that resulted in better accuracy on training data are described in Table 2.

3 Results

The results of all the subtracks were very encouraging. For English STS, the results are outlined in Table 3. SVCSTS was ranked 14th among 73 runs. The results of Spanish STS are shown in Table 4. We were ranked 4th among 16 runs. Table 5 and Table 6 summarize the results of Interpretable STS for chunked and non-chunked input respectively. Runs 2 and 3 seemed to outperform many other participating systems for non-chunked sentence input.

Inputs	Baseline	SVCSTS
answers-forums	0.4453	0.6561
answers-students	0.6647	0.7816
belief	0.6517	0.7363
headlines	0.5312	0.8085
images	0.6039	0.8236
Mean	0.5871	0.7775
Rank	59	14

Table 3: Scores for English STS

Inputs	SVCSTS
Wikipedia	0.59364
Newswire	0.65471
Mean	0.63430
Rank	4

Table 4: Scores for Spanish STS

Acknowledgments

We thank Dr. Ted Pedersen for introducing us to SemEval shared tasks.

Inputs	Baseline	SVCSTS
For Headlines - Run 2		
F1 Ali	0.6701	0.7820
F1 Type	0.4571	0.5154
F1 Score	0.6066	0.7024
F1 Type+Score	0.4571	0.5098
For Images - Run 3		
F1 Ali	0.7060	0.8336
F1 Type	0.3696	0.5759
F1 Score	0.6092	0.7511
F1 Type+Score	0.3693	0.5634

Table 5: Scores for Interpretable STS (Chunked Input)

Inputs	Baseline	SVCSTS
For Headlines - Run 1		
F1 Ali	0.8448	0.8861
F1 Type	0.5556	0.5962
F1 Score	0.7551	0.7960
F1 Type+Score	0.5556	0.5887
For Images - Run 2		
F1 Ali	0.8388	0.8853
F1 Type	0.4328	0.6095
F1 Score	0.7210	0.7968
F1 Type+Score	0.4326	0.5964

Table 6: Scores for Interpretable STS (Raw Input)

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 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 2015. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: 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 (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. *sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

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

Abhay Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya Satyapanich, Sunil Gandhi, and Tim Finin. Meerkat mafia: Multilingual and cross-level semantic textual similarity systems. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 416–423, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.

Md. Arafat Sultan, Steven Bethard, and Tamara Sumner. Dls@cu: Sentence similarity from word alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland, 2014. Association for Computational Linguistics and Dublin City University. Winner of the shared task.

Md. Arafat Sultan and Steven Bethard and Tamara Sumner Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence *Transactions of the Association for Computational Linguistics*, Vol. 2, (May), pages 219–230.

Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python *Journal of Machine Learning Research*, Vol 12, pages 2825–2830 2011

https://github.com/openlabs/Microsoft-Translator-Python-API