IHS-RD-Belarus at SemEval-2016 Task 1: Multistage Approach for Measuring Semantic Similarity

Maryna Beliuha, Maryna Chernyshevich
IHS Inc. / IHS Global Belarus
131 Starovilenskaya St
220123, Minsk, Belarus
{Marina.Beliuga}@ihs.com,
{Marina.Chernyshevich}@ihs.com

Abstract

This paper describes the system for rating the degree of semantic equivalence between two text snippets developed by IHS-RD-Belarus for the SemEval 2016 STS shared task (Task 1). To predict the human ratings of text similarity we use a support vector regression model with multiple features representing similarity and difference scores calculated for each pair of sentences.

1 Introduction

Measuring semantic equivalence between two texts has become an emerging research subject in recent years. Graded textual similarity notion can be applied to a wide range of NLP tasks such as paraphrase recognition, automatic machine translation evaluation, question answering, text summarization, information retrieval, etc.

The SemEval STS shared task has been held annually since 2012 (Agirre et al., 2012; Agirre et al., 2013; Agirre et al., 2014, Agirre et al., 2015) attracting numerous participating teams with various approaches, such as alignment of related content word sequences (Sultan et al., 2015), measuring similarity between vector representations of texts and using machine learning algorithms for computing multiple lexical, syntactic and semantic features.

In this article, we present the system developed by IHS-RD-Belarus for automated measuring of semantic similarity between two sentences using support vector regression (SVR) implemented in LIBSVM toolbox[1] (Chang and Lin, 2011) with multiple features representing similarity and difference scores calculated for each pair of sentences.

The rest of the paper is structured as follows. In Section 2 we describe the task and the data used to train our system. Section 3 describes in detail the features used by our system. Results and a short conclusion are presented in Section 4 and 5, respectively.

2 Task description

Semantic Textual Similarity (STS) measures the degree of equivalence in the underlying semantics of paired snippets of text. It can range from complete unrelatedness to exact semantic equivalence (Agirre et al., 2015).

Given two sentences, participating systems are asked to return a continuous valued similarity score on a scale from 0 to 5, with 0 indicating that the semantics of the sentences are completely unrelated and 5 signifying semantic equivalence. For example, the sentence "*Military plane crashes in south France*" and the sentence "*Military plane crashes in southeastern Turkey, 1 dead*" have a very low similarity score despite many equal words, thus scored 1.0, while the sentence "*Sarkozy announces re-election bid*" and the sentence "*France's Nicolas Sarkozy makes his reelection bid official*" are scored 4.2, as they are considered very similar even though there are many word differences between them.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Proceedings of SemEval-2016, pages 1280–1285,
San Diego, California, June 16-17, 2016.

2.1 Dataset

The dataset used for training of our system consists of the datasets provided by the organizers of the STS shared task, specifically 1500 pairs of news-wire headlines, 1500 pairs of image descriptions, 450 pairs of sentences from forum posts, 300 pairs of sentences from news summary, 750 pairs of students answers, 375 pairs of Q&A forum answers and 375 pairs of sentences from committed belief annotation. We excluded some of the datasets provided by the organizers from our training data as they had a negative influence on all the data we used for testing in preliminary experiments. As a development test set we used 20% of each dataset provided by the organizers as training data.

Our intention was to create a universal system, therefore we didn't use any domain specific features and didn't train separate models for predicting the similarity scores of the text snippets provided for evaluation depending on their domain. Thus, we used the same set of data listed above to train our SVR model and the same model was used to predict the similarity scores of all test datasets.

2.2 Evaluation

Gold standard scores are averaged over multiple human annotations. Performance of the systems is assessed by computing the Pearson correlation between machine assigned semantic similarity scores and gold standard scores.

3 System description

To predict the scores of the test set we use supervised machine learning, specifically a support vector regression model, to combine a large amount of features computed from pairs of sentences. Each feature represents either a similarity or difference score between two text snippets. To obtain the optimal values for SVR parameters *C*, *g* and *p*, we used grid search. The system we end up submitting had parameters $C = 10$, $g = 0.1$ and $p = 0.2$.

3.1 Semantic similarity features

The workflow for computing the similarity measures is based on a multistage aligner using various internal and external resources that align identical or similar words and phrases. As these resources have different degrees of reliability, we calculate similarity measures, described in Section 3.1.1, after each alignment step and use them as separate features.

The aligning steps performed by our system are illustrated in Figure 1.

Stage 1. First, to identify semantically similar words and phrases we use the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013), which is a large database of lexical, phrasal and syntactic paraphrases.

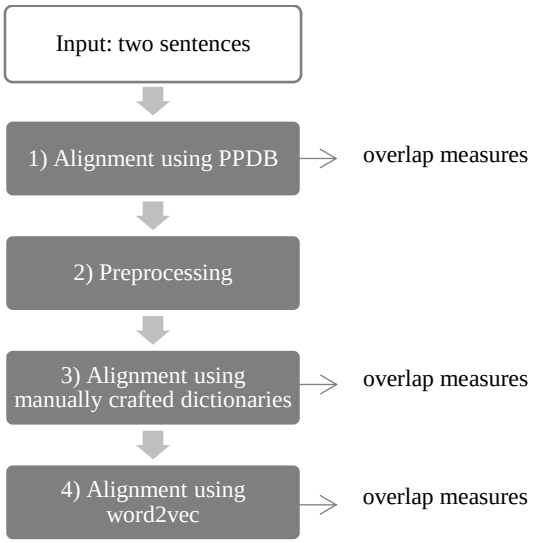

Figure 1: Aligning steps performed by the system

Stage 2. To obtain additional word alignments our system performs the following preprocessing steps:

- words are tokenized with IHS Goldfire[2] linguistic processor (Чеусов, 2006);
- words are POS tagged with IHS Goldfire linguistic processor;
- stop words, uninformative adverbs, discourse markers and words that do not contain at least one letter or number are deleted;
- words are lowercased;
- punctuation marks are removed.

On **Stage 3** we apply manually crafted domain independent wordlists such as 14870 pairs of synonymous adjectives, verbs and nouns *("wrong" – "incorrect", "link" – "connect", "seller" – "vendor"*), 1435 pairs of adjectives and adverbs derived from them *("clear" – "clearly"*), 2953 pairs of ac-

[2] https://www.ihs.com/products/design-standards-software-goldfire.html

tions and their agents (*"connect" – "connector"*), 17556 pairs of verbs and deverbal nouns (*"pulsate" – "pulsation"*), 563 pairs of nouns and denominal adjectives (*"sinusoid" – "sinusoidal"*). The lists were automatically generated using derivational affixes and then validated on random corpus of two million sentences (a derived synonym was considered valid if it appeared in the corpus more than 3 times). We also used Wikipedia lists of 264 paired country names and nationalities (*"british" – "uk"*) and 262 paired country names and capitals (*"uk" – "London"*).

During this stage we also align words having a Levenshtein distance of less than 1 to catch common misspellings.

On **Stage 4** we align words in the sentences using the GoogleNews vectors dataset, available on the word2vec web site[3], which has a 3,000,000 word vocabulary of 300-dimensional word vectors trained on about 100 billion words. We consider two words to be semantically similar if the cosine between the words is more than 0.7.

3.1.1 Word overlap measures

To calculate similarity scores we use two similarity measures which we apply after each alignment step described above.

Jaccard similarity coefficient. We measure the similarity score using Jaccard index, which is defined as the amount of word overlap normalized by the union of the sets of words present in the two sentences. It is calculated using the formula:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

where S_1 and S_2 are the vectors of the first and the second sentence, respectively. We consider the intersection of vectors S_1 and S_2 to be to be the words that are aligned to each other.

Similarity score using TF-IDF. A word's TF-IDF score reflects the importance of the word for the particular sentence offset by the frequency of the word in all sentences. It can be used as a weighting factor when calculating semantic similarity of the sentences. Therefore, as one of the features, we calculate the TF-IDF weighted proportion of aligned content words over the sum of the TF-IDF scores for all words in the two sentences. In other words, given sentences S_1 and S_2,

$$sts(S_1, S_2) = \frac{\sum_{w \in (S_1^a \cup S_2^a)} tfidf(w)}{\sum_{w \in (S_1 \cup S_2)} tfidf(w)}$$

where $\sum_{w \in (S_1^a \cup S_2^a)} tfidf(w)$ is a sum of TF-IDF values of all the aligned words in S_1 and S_2, while $\sum_{w \in (S_1 \cup S_2)} tfidf(w)$ is a sum of TF-IDF values of all words in both sentences.

Intuitively, the lower is the sum of TF-IDF values of the aligned words, the less important the aligned words, which makes the total similarity score smaller and vice versa.

3.2 Cosine similarity measures

Similarity can also be defined by the cosine of the angle between two vectors. Cosine similarity is one of the most well-known similarity measures as has been broadly applied to numerous information retrieval tasks (Strehl, 2000). We calculated the similarity of pairs of text snippets using the following equation:

$$cos(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where *A* and *B* are TF-IDF vectors. The cosine similarity score is non-negative and bounded between [0,1].

Following our multi-stage principle we generate word vectors and calculate cosine similarity scores after each of 4 stages of original sentences transformation and use them as separate features. Each step aims to limit words diversity of text:

Stage 1. Normalization: change to a lowercase, removal of punctuation marks and stop words.

Stage 2. Grammatical transformation of words: nouns to singular form; verbs to infinitive form; adjectives and adverbs to positive degree.

Stage 3. Aligning, described in detail in section 3.1.

Stage 4. Words expansion: each word in a sentence is expanded with a vector of words with the word2vec similarity of more than 0.6, generated using word2vec and pre-trained Google vectors. We add only words having a zero entry in the vector. The TF-IDF scores of the added words are cal-

[3] https://code.google.com/archive/p/word2vec/

culated in the following way: the TF-IDF of the source word is multiplied by the word2vec similarity. The resulting vectors are approximately 4 times larger than the initial ones.

3.3 Syntactic similarity

Word overlap based measures may lead to discrepancies, because they do not capture syntactic relations. The similarity of the words or phrases having the same syntactic roles in two sentences may be indicative of their overall semantic similarity (Oliva et al., 2011) and vice versa. For example, in two expressions "the notebook of my mother" and "my mother cooks amazing" the word "mother" has different syntactic meaning: in the first expression it has just an attributive meaning, while in the second it is a subject. Similarly, if two sentences differ, for example, by their main predicates (e.g., *"predict"* and *"walk"*), it can indicate that they are more likely to have significant semantic differences.

To address this issue, we design features that compute similarity scores based on a syntactic analysis of the sentences.

Noun phrase similarity feature: To align noun phrases first we extracted all noun phrases from the first and the second sentence and calculated similarity between each pair of noun phrases. Noun phrase extraction was performed by IHS Goldfire linguistic processor that is based on shallow parsing but with support for head word identification.

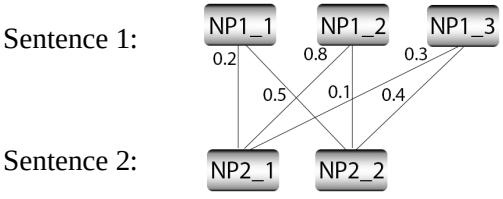

Figure 2: Example of noun phrase alignment

The similarity between two noun phrases is calculated using the WordNet path similarity provided by NLTK[4] and the Levenshtein distance. Path similarity scores denote how similar two word senses are based on the shortest path that connects the senses in the WordNet hypernym-hyponym taxonomy. The similarity between two noun phrases is the sum of the similarity between head words $Sim(mw)$ and the similarity of all attributes $Sim(att)$ divided by 2 in order to penalize the weight of attributes:

$$Sim(np_1, np_2) = Sim(mw) + \frac{Sim(att)}{2}$$

After calculating all pairwise similarity scores, we align each noun phrase from the first sentence with the noun phrase from the second one having the highest similarity score:

NP1_1 : NP2_2 (0.5),
NP1_2 : NP2_1 (0.8),
NP1_3: NULL (0)

The sentence level noun phrase similarity score is calculated as the averaged score of the individual aligned noun phrase similarities.

Parse trees comparison feature: At this level of analysis we compare binary verb-centric nodes (Subject-Action, Action-Object, etc.) of the trees extracted with IHS Goldfire linguistic processor while leaving aside the nodes of lesser importance (Main-Attribute, Main-Preposition, etc.). Complete or partial match within the contents of the important verb-centric compared nodes suggests various degrees of syntactic and semantic role equivalence.

We give a higher score to a pair of sentences if they have aligned words in the same or similar verb-centric nodes (the score then equals the count of the number of matching verb-centric nodes) and we give a penalty score if the sentences have no matching nodes or aligned words appear only in non-informative nodes (the score equals minus one). If any of the sentences has no verb-centric nodes at all and therefore the parse tree is not generated the score remains 0.

3.4 Differentiation features

Another set of features is used to reveal semantic differences between the sentences.

Part-of-speech feature: We assume sentences that differ by uninformative words to have a higher similarity score than sentences that differ by words that have informative POS-tags such as verbs or nouns. Compare a pair of sentences having different determiners *"how is this possible?" – "how is that possible?"* and a pair of sentences having different nouns *"I love cats" – "I love dogs"*. Therefore, we calculate a weighted sum of weights of all non-matching words having informative POS-tags

[4] http://www.nltk.org/howto/wordnet.html

(i.e. verbs, nouns, adjectives, adverbs and numerals) based on an empirically determined informativeness weight of the POS-tags.

Named entity feature: Comparing, for example, the pair of sentences *"Ten people killed in twin blasts in Nigeria"* and *"Ten killed in new blast in Russia"*, we assume that the impact of named entities to semantic equivalence is disproportionately high. Therefore, we count the total number of non-matching named entities as another feature.

As matching named entities we considered two named entities that match:

- exactly: *Russia - Russia*;
- partially: *Bill Torn - Mr. Torn*;
- country name and nationality pairs list: *british – UK*;
- country name and capital pairs list: UK – London.

TF-IDF feature: We calculate TF-IDF of all words that differ in two sentences. Taking its mean value as a separate feature allows us to make a conclusion on how different the sentences are: the lower is the value of this feature, the higher is the degree of similarity, and vice versa.

4 Results

To assess system performance, the organizers provided five test sets from different domains. Table 1 illustrates the performance of the system developed by our team as compare to the top and median scores of the other systems that participated in the task.

Our system outperformed most other systems achieving very promising results. As seen below, it performed well above the median for all of the datasets and achieved results that are very close to the best performing system on headlines. However, note that most systems showed relatively poor performance on the answer-answer (ans-ans) and question-question (ques-ques) datasets which can be explained by significant differences between the provided training data and these particular test sets.

Dataset	Best	Median	Our system
plagiarism	.84138	.78949	.82634
ans-ans	.69235	.48018	.55322
postediting	.86690	.81241	.83761
headlines	.82749	.76439	.82539
ques-ques	.74705	.57140	.599
ALL	**.77807**	**.68923**	**.728312**

Table 1: Performance on the 2016 STS Test Set

5 Conclusion

In this paper we present our system for automatic rating of semantic textual similarity developed by our team for the SemEval 2016 STS shared task (Task 1). To measure semantic equivalence of two text snippets we use a supervised system based on a support vector regression model to combine multiple features representing similarity and difference scores calculated for each pair of sentences. Our system performed relatively well on all of the STS 2016 evaluation datasets. We believe that introducing additional features for deeper understanding of textual semantics might further improve performance on the task.

References

Alexander Strehl, Joydeep Ghosh, Raymond Mooney. 2000. Impact of Similarity Measures on Web-page Clustering. In Proceedings of the AAAI-2000: Workshop of Artificial Intelligence for Web Search, July 2000, 58-64

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, WeiWei Guo. *SEM 2013 shared task: Semantic Textual Similarity. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM '13, pages 32-43, Atlanta,Georgia, USA.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, SemEval '12, pages 385-393, Montreal, Canada.

Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei Guo; Rada Mihalcea; German Rigau; Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval '14, Dublin, Ireland.

Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei

Guo; Inigo Lopez-Gazpio; Montse Maritxalar; Rada Mihalcea; German Rigau; Larraitz Uria; Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, June.

James Todhunter, Igor Sovpel and Dzianis Pastanohau. System and method for automatic semantic labeling of natural language texts. *U.S. Patent 8 583 422, November 12, 2013*.

Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, Ángel Iglesias. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase 152Database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 758-764.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 148–153, Denver, Colorado, June. Association for Computational Linguistics.

Чеусов, А. В. Разработка лингвистических процессоров промышленной обработки текстовых документов / А. В. Чеусов // Искусственный ин-теллект. Интеллектуальные и многопроцессорные системы: материалы научно-технической конференции, п. Кацивели, 25-30 сентября 2006 г.; в 3 т. / Мин. обр-я и науки Рсн, Мин. обр-я и науки Украины, НАН Беларуси; ред. В.О. Бронзов. - Таганрог: ТРТУ, 2006. - Т. 2. - С. 366-370.