Sbdlrhmn: A Rule-based Human Interpretation System for Semantic Textual Similarity Task

Samir AbdelRahman
sbdlrhmn@illinois.edu,
s.abdelrahman@fci-cu.edu.eg

Catherine Blake
clblake@illinois.edu

The Graduate School of Library and Information Science
University Of Illinois at Urbana-Champaign

Abstract

In this paper, we describe the system architecture used in the Semantic Textual Similarity (STS) task 6 pilot challenge. The goal of this challenge is to accurately identify five levels of semantic similarity between two sentences: equivalent, mostly equivalent, roughly equivalent, not equivalent but sharing the same topic and no equivalence. Our participations were two systems. The first system (rule-based) combines both semantic and syntax features to arrive at the overall similarity. The proposed rules enable the system to adequately handle domain knowledge gaps that are inherent when working with knowledge resources. As such one of its main goals, the system suggests a set of domain-free rules to help the human annotator in scoring semantic equivalence of two sentences. The second system is our baseline in which we use the Cosine Similarity between the words in each sentence pair.

1 Introduction

Accurately establishing sentence semantic similarity would provide one of the key ingredients for solutions to many text-related applications, such as automatic grading systems (Mohler and Mihalcea, 2009), paraphrasing (Fernando and Stevenson, 2008), text entailment (Corley et al., 2005) and summarization (Erkan and Radev, 2004). Current approaches for computing semantic similarity between a pair of sentences focus on analyzing their shared words (Salton, 1989), structures (Hu et al. 2011;Mandreoli et al. 2002), semantics (Mihalcea et al. 2006; Le el al. 2006; Hatzivassiloglou, 1999) or any of their combinations (Liu et al. 2008; Foltz et al. 1998). The goal is to arrive at a score which increases proportionally with the relatedness between the two sentences. Yet, they are not concerned with scoring the interpretations of such relatedness (Zhang et al. 2011; Jesus et al. 2011; Wenyin et al. 2010; Liu et al. 2008).

Semantic Textual Similarity (STS), SEMEVAL-12 Task 6 (Agirre et al. 2012), measures the degree of semantic equivalence between a pair of sentences by comparing meaningful contents within a sentence. The assigned scores range from 0 to 5 for each sentence pair with the following interpretations: (5) completely equivalent, (4) mostly equivalent pair with missing unimportant information, (3) roughly equivalent with missing important information, (2) not equivalent, but sharing some details, (1) not equivalent but sharing the same topic and (0) not equivalent and on different topics. The goal of developing our rule-based system was to identify knowledge representations which have possibly all task human interpretations. Meanwhile, the system domain-free rules aim to help the human annotator in scoring semantic equivalence of sentence pair.

The proposed rule-based solution exploits both sentence syntax and semantics. First, it uses Stanford parser (Klein and Manning, 2002) to expose the sentence structure, part-of-speech (POS) word tags, parse tree and Subject-Verb-Object (S-V-O) dependencies. Second, Illinois Coreference Package (Bengtson and Roth, 2008) is used to extract sentence named entities resolving possible men-

*First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 536–542,
Montréal, Canada, June 7-8, 2012.

tions. Third, WordNet (Miller, 1995) and Adapted Lesk Algorithm for word sense disambiguation (Banerjee and Pedersen, 2010) are used to compute each sentence word semantic relatedness to the other sentence. ReVerb (Etzioni et al. 2011) augments WordNet in case of uncovered words and helps us to discriminate the topics of sentences. We use (Blake, 2007) thought to compare the sentence pair words with each other. Finally, we evolve a rule-based module to present the human heuristics when he interprets the relatedness of the sentence pair meaningful contents.

Throughout our training and testing experiments, we used Task6 corpora (Agirre et al. 2012) namely MSRpar, MSRvid, SMTeuroparl, OnWN and SMTnews; where:

- MSRpar is 1500 pairs of sentences of MSR-Paraphrase, Microsoft Research Paraphrase Corpus; 750 for training and 750 for testing.
- MSRvid is 1500 pairs of sentences of MSR-Video, Microsoft Research Video Description Corpus; 750 for training and 750 for testing.
- SMTeuroparl is 918 pairs of sentences of WMT2008 development dataset (Europarl section); 459 for training and 459 for testing.
- OnWn is 750 pairs of sentences pairs of sentences where the first sentence comes from Ontonotes and the second sentence from a WordNet definition; it is only a testing corpus.
- SMTnews is 399 pairs of sentences of news conversation sentence pairs from WMT; it is only a testing corpus.

The reminder of this paper is organized as follows: Section 2 describes our two participations; Section 3 discusses their official results; Section 4 draws our conclusion for both systems.

2 The Proposed Systems

In this section, we focus on the rule-based system, Sections 2.1, 2.2, 2.3 and 2.4, as our main task contribution. Further, the section describes our second run, Sections 2.5, to shed light on the role of cosine similarity for solving the task problem. To establish the task semantic textual similarity, we show how the rule-based system exploits the sentence semantic, syntax and heuristics; also, we describe how our base-line system uses the sentence syntax only.

2.1 Definitions

We say the two sentences are on different topics, if all their verbs are mostly (> 50%) unrelated (Table 1). Otherwise, they are on the same topic. For example, the two sentences *"A woman is putting on makeup."*, *"A band is singing."* are on different topics as *"putting"*, *"singing"* are not equivalent. However, the two sentences *"A baby is talking."*, *A boy is trying to say firetruck."* are on the same topics as *"talking"* and *"trying to say"* are semantically equivalent.

We define the sentence important information as its head nouns, named entities or main verbs; where the main verbs are all verbs except auxiliary, model and infinitive ones. Hence, we say that two sentences miss important information if either loses at least one of these mentions from the other. Otherwise, they are candidates to be semantically equivalent. For example, the sentence *"Besides Hampton and Newport News, the grant funds water testing in Yorktown, King George County, Norfolk and Virginia Beach."* misses *"Hampton and Newport News"* compared to the sentence *"The grant also funds beach testing in King George County, Norfolk and Virginia Beach."* However, "on a table" is unimportant information which *"A woman is tapping her fingers."* misses compared to *"A woman is tapping her fingers on a table."*

Finally, we deploy a list of stop words and non-verbs as unimportant information. However, if any exists in both sentences, we match them with each other; otherwise we ignore any occurrences.

2.2 The Syntactic Module

This syntactic module is a preprocessing module in which the system calls Stanford parser, Version 2.0.1, and the Illinois coreference package, Version 1.3.2, to result in the sentence four type representations: 1) part of speech (POS) tags, 2) Subject-Verb-Object (S-V-O), Subject-Verb (S-V) and Verb-Object (V-O) dependencies, 3) parse tree and 4) coreference resolutions. All sentences are lemmatized based on their POSs. Also, verbs and CDs are utilized to determine topics/important information and numbers respectively. All noun and verb phrases are used to boost the sentence word semantic scores (Section 2.3). We consider all occurrences of S-V-O, S-V and V-O to distinguish

the topic compatibility between two comparable sentences (Section 2.3 and 2.4).

The coreference package is used to match the equivalent discourse entities between two sentences which improve the matching steps. For example, in the pair of *"Mrs Hillary Clinton explains her plan towards the Middle East countries"* and *"Mrs Clinton meets their ambassadors"*, *"Mrs Hillary Clinton"*, *"her"* and *"Mrs Clinton"* refer to the same entity where "*the Middle East countries"* and *"their"* are equivalent. Moreover, we consider the second sentence doesn't lose *"Hillary"* as missing important information since the related mentions are labeled equivalent.

2.3 The Semantic Matching Module

WordNet, Version 3.0, has approximately 5,947 entries covering around 85% of training corpora words (Agirre et al. 2012). Most of the remaining 15% words are abbreviations, named entities and incorrect POS tags. We use WordNet shortest path measure to compute the semantic similarity between two words. Also, we use Adapted Lesk algorithm to obtain the best WordNet word sense. The disambiguation algorithm compares each pair of words through their contexts (windows) of words coupled with their all overlapping glosses of all WordNet relation types.

The semantic matching module inputs are the sentence pair (S1, S2), their lemmatized words, parse trees, S-V-O/S-V/V-O dependencies and coreference mentions (Section 2.2). It matches syntactically the words with each other. For any uncovered WordNet word, the module calls ReVerb (Section 2.4) and it assigns the returned value to the word score. All numbers, e.g. million, 300,45.6, are mathematically compared with each other. This module compares the noun phrases with single words to handle the compound words, e.g. *"shot gun"* with *"shotgun"* or *"part-of-speech"* with *"part of speech"*. For those words whose scores are not equal to 1, it compares each pair of words from the sentence pair within their Subject-VP (subject with its verb phrase) contexts using Adapted Lesk algorithm to find best sense for each included word. Then, it applies WordNet shortest path measure to score such words. In our disambiguation algorithm implementation, we found that the runtime requirement is directly proportional to the input sentence length. So, we shortened the sentence length to Subject-VP which includes the underlying comparable words.

Relatedness	Score (S1, S2)
unrelated	0 <= Ws <0.3
weakly related	0.3 <= Ws <0.85
strongly related	Ws >= 0.85

Table 1 – Mapping relatedness to wordnet similarity

Table 1 describes the proposed system WordNet thresholds through our relatedness definitions. The thresholds were thoroughly selected depending on our analysis for the WordNet hierarchary and semantic similarity measures (Pedersen et al., 2004). We obsereved that while most of the nearest tree sibilings and parent-child nodes scores have more than 0.85 Wordnet semantic scores, most of the fartherest ones have scores less than 0.3. In between these extremes, there is a group of scattered tree nodes which ranges from 0.3 to 0.85. The number of nodes per each mentioned group is related to the semantic simlarity measure technique.

2.4 Semantics – Using ReVerb

Our working hypothesis is that verbs that use the same arguments are more likely to be similar. To estimate verb usage, the system uses frequencies from the ReVerb (http://openie.cs.washington.edu/) online interface to count the number of times a verb is used with two arguments. For example, consider the sentence pair *"The man fires rifle"* and *"The man cuts lemon"*. The number of sentences in ReVerb that contain the verb *fires* with the argument *rifle* is 538 and the number of sentences for the verb *cuts* with the argument *lemon* is 45, which tell us that you are more likely to find sentences that describe firing a rifle than cutting a lemon on the web. However, there a no ReVerb sentences for the verb *fires* with the argument *lemon* or the verb *cuts* with the argument *rifle*. Which tells us that people generally don't fire lemons or cut rifles.

Reverb provides the system with information about the suitability of using argument in one sentence with verbs from another. Specifically, frequencies from Reverb are retrieved for each subject-verb-object triple in each sentence, e.g. "S1-V1-O1" and "S2-V2-O2". The system then retrieves ReVerb frequencies for the verb-object in

each sentence of "V2-O1" and "V1-O2". If at least one of all of these scores equals to 0, they are considered to be weakly similar.

ReVerb is also called for any sentence word that WordNet doesn't cover. The system retrieves the Reverb frequency for is-a relation using the word missing from Wordnet, as Argument1, and each word from the other sentence as Argument2. The largest Reverb retrieved score is taken. Consider the pair of *"A group of girls are exiting a taxi"* and *"A video clip of Rihanna leaving a taxi."*. Since "*Rihanna*" is not a WordNet word, our ReVerb interface hits the web for "*Rihanna is-a girl", "Rihanna is-a group", "Rihanna is-a taxi" and "Rihanna is-a existing"* and it returns *"Rihanna is-a girl"* as the best candidate with strength score equals 0.2.

We explored several relatedness scores which specifically equal to 0, 0.2, 0.4, 0.6, 0.8 or 1 if the frequencies are less than to 10, 50, 100, 500, 1000 or 1000+ respectively.

2.5 The Rule-Based Module

Rule-based module aims at defining human-like rules to interpret how the pair similar or dissimilar from each other. Pair Similarity (P) is based on the strong relatedness values (Table 1) and the Dissimilarity (D) is based on the other types of relatedness values. As we believe that strong and not strong are proportional to the pair similarity and dissimilarity respectively

Rule-based module input is sentence pair S1, S2 word semantic scores, i.e. Ws1s and Ws2s (Table 1). Then, it calculates: 1) their three types of averages for S1 and S2 semantic scores, i.e. all word semantic scores, weakly related only and unrelated values; 2) P as the minimum percentage of strong Wss in (S1 and S1); 3) D as, 100-P, the percentage of not strong Wss in S1 and S1

This module outputs the semantic textual similarity semantic (STS) score which ranges from 0 to 5. Throughout this section, when we use "unrelated", "weak" and strong terminologies, we use Table 1 Relatedness definitions. Also, when we use "important" term, we refer to our definition (Section 2.1)

Human judgments for computing STS score of the sentence pair are based on word similarities and dissimilarities. They consider that two sentences are similar if most (> 50%) of their words are strongly related, otherwise the sentences are candidates to be dissimilar. Since all Wss range from 0 to 1, the average of strong scores is more than the average of weak scores. Likewise, the average of weak scores is more than the average of non-related scores.

Score(Sentence Ws1s, Sentence Ws2s)

AllAvg = (Ws1s+ Ws1s)/2
WeakAvg= the averaged weakly related scores of Ws1s and Ws1s
UnRElAvg=the average of unrelated scores of Ws1s and Ws1s
P = minimum (% Ws1s strong scores, % Ws2s strong scores)
D=100-P
Value=0
If 95 <= P <=100 then Value = 5;
If 80 <= P < 95 then Value = 4;
If 50 <= P < 80 then Value = 3;
If 20 <= P < 50 then Value = 2;
If 0 <= P < 20 then
  If all verbs are strongly related then Value=1
  Else Value= 0.0001;
If (Value in [4, 5]) then
  If all Ds for important words then Value= 3
If (Value ==3) then
  If all Ds for not important words then Value= 2
If (Value <> 5 AND Value <> 0) then
  If all Ds for weakly related words
    Value= Value+ AllAvg
  Else if at least half Ds for weakly related words
    Value= Value+ WeakAvg
  Otherwise
    Value = Value + UnRelAvg
Return Value

When we call Score(Ws1s,Ws2s), we take care of the following two special cases where it goes directly to Value 3: 1) if missing some words leads to missing the whole verb/noun phrases and 2) if one sentence has all past tense verbs and the other has present verbs.

When we design P inequalities, we make them have relaxed boundaries conformed with human grading values. For example, we choose P between 95 and 100 in Value (5); where 95 and 100 equal to grades 4.5 and 5 respectively. Value (3) interval are values between more than or equal 2.5 and less than 4. Then, we utilize the important information

and verb constraints to direct classifications through different groups.

When we design range conditions between values, we select D to present the distance between the sentence pair. As D weak values increase, the two sentences become closer. As D unrelated values increase, the two sentences become distant.

We carefully analyzed the training corpora to assure that the above thresholds satisfy most of the training sentence pairs. Each threshold output was manually checked and adjusted to satisfy around 55% to 75% of the training corpora.

Applying the above module, the pair of *"A man is playing football"* and *"The man plays football"* STS score equals 5.00. The pair of *"A man is singing and playing"* and *"The man plays"* STS score equals 3.00 since the first one misses *"singing"*. The pair of *"The cat is drinking milk."* and *"A white cat is licking and drinking milk kept on a plate."* STS scores equals to 3.4 since they have P=0.66, *"white"* as unimportant information but *"licking"*, *" kept"*, *"plate"* as important information words.

2.6 Our Baseline System Description

Our goal in the second run is to evaluate the relatedness of the two sentences using only the words in the sentence. Sentences are represented as a vector (i.e. based on the Vector Space Model) and the similarity between the two sentences S1 and S2 is (5* cosine similarity). We take into account all sentence words such that they are lower-case and non-stemmed.

3 Results and Discussion

3.1 Rule-based System Analysis

Our system was implemented in Python and used the Natural Language Toolkit (NLTK, www.nltk.org/), WordNet and lemmatization modules. Table 2 provides in the official results of our system Pearson-Correlation measure.

D	Para	Vid	Europ	OnWn	News
Tr	0.6011	0.7021	0.4528		
Te	0.5440	0.7335	0.3830	0.5860	0.2445

Table 2. Run1 Official Person-Correlation measure

In Table 2, the first row shows the proposed system results namely 0.6011, 0.7021 and 0.4528 for MSRpar, MSRvid and SMTeuropel training corpora respectively. The second row shows the test results, namely 0.5440, 0.7335 and 0.3830, 0.5860 and 0.2445 for MSRpar, MSRvid and SMTeuropel, On-Wn and SMTnews testing corpora respectively.

In the Task-6 results (Agirre et al. 2012), our system was ranked 21th out of 85 participants with 0.6663 Pearson-Correlation ALL competition rank. We tested two WordNet measures, namely the shortest path and WUP, the path length to the root node from the least common subsumer (LCS) of the two concepts, measures on the training corpora. In contrast to the shortest path measure, WUP measure increased the P versus the D scores on the three corpora. This overestimated many training STS scores and negatively affected the correlation with the gold standard corpora. Using WUP measure, the correlations of MSRpar, MSRvid and SMTeuropel corpora were 0.5553, 0.3488 and 0.4819 respectively. We decided to use WordNet shortest path measure due to its better correlation results. When we used WUP measure on testing corpora, the correlations were 0.5103, 0.4617, 0.4810, 0.6422 and 0.4400 for MSRpar, MSRvid and SMTeuropel, On-Wn and SMTnews testing corpora respectively. We observed that when we used WUP measure on MSRvid corpora, the correlations were degraded. This is because most of MSRvid corpus pair sentences talking about human genders which have high WUP scores when comparing with each other. Unfortunately, WordNet shortest path measure underestimated SMTnews pair sentence similarities which affected dramatically the related correlation measure. Hence, the choice of the suitable WordNet metric for the whole corpora is still under our consideration.

Thresholds and Semantic Pattern: Our current efforts are directed towards statistical modeling of the system thresholds. We intend also to use some web semantic patterns or phrases, such as ReVerb patterns, to boost the semantic scores of single words.

3.2 Baseline System Analysis

In Table 3, the first row shows the proposed system results namely 0.4688, 0.4175 and 0.5349 for

MSRpar, MSRvid and SMTeuropel training corpora respectively. The second row shows the proposed system results, namely 0.4617, 0.4489 and 0.4719, 0.6353 and 0.4353 for MSRpar, MSRvid and SMTeuropel, On-Wn and SMTnews testing corpora respectively.

D	Para	Vid	Europ	OnWn	News
Tr	0.4688	0.4175	0.5349		
Te	0.4617	0.4489	0.4719	0.6353	0.4353

Table 3. Run 2 Official Person-Correlation measure

In the Task-6 results (Agirre et al. 2012), Run2 was ranked 72th out of 85 participants with 0.4169 Pearson-Correlation ALL competition rank. As anticipated, Run2 released fair results. Its performance is penalized or awarded proportionally to the number of exact matching pair words. Accordingly, it may record considerable scores for pairs which have highly percentage exact matching words. For example, it provides competitive correlation scores compared to other participants on On-Wn and SMTnews testing corpora. Though, this doesn't imply that it is an ideal solution for STS task. It usually indicates that many corpus pairs may have some substantial exact matching words.

4 Conclusions

In this paper, we presented systems developed for SEMEVAL12- Task6. The first run used both semantics and syntax. The second run, our baseline, uses only the words in the initial two sentences and defines similarity as the cosine similarity between the two sentences. The official task results suggest that semantics and syntax (Run1) supersedes the words alone (Run 2) with 0.2494 which indicates that the words alone are not sufficient to capture semantic similarity.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. (1115774). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Catherine Blake. 2007. *The Role of Sentence Structure in Recognizing Textual Entailment.* RTE Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing:101-106.

Courtney Corley and Andras Csomai and Rada Mihalcea. 2005. *Text Semantic Similarity, with Applications.* Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP), Borovetz, Bulgaria.

Dan Klein and Christopher D. Manning. 2002. *Fast Exact Inference with a Factored Model for Natural Language Parsing.* In *Advances in Neural Information Processing Systems 15 (NIPS)*, Cambridge, MA: MIT Press:3-10.

Dong-bin Hu and Jun Ding. 2011. *Study on Similar Engineering Decision Problem Identification Based On Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS.* Systems Engineering Procedia 1: 406–413.

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)

Eric Bengtson and Dan Roth. 2008. *Understanding the Value of Features for Coreference Resolution.* EMNLP:294-303.

Federica Mandreoli and Riccardo Martoglia and Paolo Tiberio. 2002 . *A Syntactic Approach for Searching Similarities within Sentences.* Proceeding of International Conference on Information and Knowledge Management:656–637.

George A. Miller. 1995. *WordNet: A Lexical Database for English.* Communications of the ACM, 38(11): 39-41.

Gerard Salton. 1989. *Automatic Text Processing.* The Transformation, Analysis, and Retrieval of Information by Computer. Wokingham, Mass.Addison-Wesley.

Gunes Erkan and Dragomir R. Radev. 2004. *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization.* Journal of Artificial Intelligence Research 22:457-479.

Junsheng Zhang, Yunchuan Sun, Huilin Wang, Yanqing He. 2011. *Calculating Statistical Similarity between Sentences.* Journal of Convergence Information Technology, Volume 6, Number 2: 22-34.

Liu Wenyin and Xiaojun Quan and Min Feng and Bite Qiu. 2010. *A Short Text Modeling Method Combining Semantic and Statistical Information.* Information Sciences 180: 4031–4041.

Michael Mohler and Rada Mihalcea. 2009. *Text-to-text Semantic Similarity for Automatic Short Answer Grading.* Proceedings of the European Chapter of the Association for Computational Linguistics (EACL).

Oliva Jesus and Serrano I. Jose and María D. del Castillo and Ángel Iglesias .2011. *SyMSS: A Syntax-based Measure for Short-Text Semantic Similarity.* Data and Knowledge Engineering 70: 390–405.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. *Open Information Extraction: The Second Generation.* Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI).

Rada Mihalcea and Courtney Corley and Carlo Strapparava. 2006. *Corpus-based and knowledge-based measures of text semantic similarity.* Proceeding of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference.

Peter W. Foltz and Walter Kintsch and Thomas K Landauer. 1998. *The measurement of textual coherence with latent semantic analysis.* Discourse Processes Vol. 25, No. 2-3: 285-307.

Samuel Fernando and Mark Stevenson. 2008. *A Semantic Similarity Approach to Paraphrase Detection.* Computational Linguistics (CLUK) 11th Annual Research Colloquium, 2008.

Satanjeev Banerjee and Ted Pedersen. 2010. *An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet.* CICLING:136-145.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. *WordNet::Similarity-measuring the relatedness of concepts*. In Proceedings of NAACL, 2004.

Vasileios Hatzivassiloglou , Judith L. Klavans , Eleazar Eskin. 1999. *Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning.* Proceeding of Empirical Methods in natural language processing and Very Large Corpora.

Xiao-Ying Liu and Yi-Ming Zhou and Ruo-Shi Zheng. 2008. *Measuring Semantic Similarity within Sentences*. Proceedings of the Seventh International Conference on Machine Learning and Cybernetics.

Yuhua Li, David McLean, Zuhair A. Bandar, James D. O'Shea, and Keeley Crockett. *Sentence Similarity based on Semantic Nets and Corpus statistics*. 2006. IEEE Transactions on Knowledgeand Data Engineering Vol. 18, No. 8: 1138-1150.