FCICU at SemEval-2017 Task 1: Sense-Based Language Independent Semantic Textual Similarity Approach

Basma Hassan[1] Samir AbdelRahman[2] Reem Bahgat[2] Ibrahim Farag[2]

[1]Faculty of Computers and Information
Fayoum University, Fayoum, Egypt
bhassan@fayoum.edu.eg

[2] Faculty of Computers and Information
Cairo University, Giza, Egypt
{s.abdelrahman, r.bahgat, i.farag}@fci-cu.edu.eg

Abstract

This paper describes FCICU team systems that participated in SemEval-2017 Semantic Textual Similarity task (Task1) for monolingual and cross-lingual sentence pairs. A sense-based language independent textual similarity approach is presented, in which a proposed alignment similarity method coupled with new usage of a semantic network (BabelNet) is used. Additionally, a previously proposed integration between sense-based and surface-based semantic textual similarity approach is applied together with our proposed approach. For all the tracks in Task1, Run1 is a string kernel with alignments metric and Run2 is a sense-based alignment similarity method. The first run is ranked 10th, and the second is ranked 12th in the primary track, with correlation 0.619 and 0.617 respectively.

1 Introduction

Semantic Textual Similarity (STS) is the task of measuring the similarity between two short texts semantically. STS is very important because a wide range of Natural Language Processing (NLP) applications rely heavily on such task.

This paper describes our participation in the STS task (Task1) at SemEval 2017 in all the six monolingual and cross-lingual tracks (Cer et al., 2017). The STS task seeks to calculate a graded similarity score from 0 to 5 between two sentences according to their meaning, i.e. semantically. The monolingual tracks are Arabic, English, and Spanish sentence-pairs (track1, track3, and track5 respectively), while the cross-lingual tracks are Arabic, Spanish, and Turkish sentences paired with English sentences (track2, track4a-4b, and track6 respectively). An additional Primary track is provided that presents the mean score of the results of all the other tracks.

The similarity between two natural language sentences can be inferred from the quantity/quality of aligned constituents in both sentences. Such alignments provide valuable information regarding how and to what extent the two sentences are related or semantically similar, where semantically equivalent text pairs are likely to have a successful alignment between their words. Our proposed sense-based approach employs this aspect to calculate the similarity between sentence-pairs regardless of their language. This is achieved through a proposed word-sense aligner that relies mainly on a new usage of the semantic network BabelNet. BabelNet utilization compensates the need of a machine translation module that is most commonly used to transfer cross-lingual STS to monolingual. Besides, the proposed sense-based similarity score is combined with a surface-based similarity score.

The paper is organized as follows. Section 2 explains our main multilingual sense-based aligner. Section 3 describes our system that participated in all tracks. Section 4 shows the experiments conducted and analyzes the results achieved. Section 5 concludes the paper and mentions some future directions.

2 Multilingual Sense-Based Aligner

Highly semantically similar sentences should also have a high degree of conceptual alignment between their semantic units: words, tokens,

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 125–129,
Vancouver, Canada, August 3 - 4, 2017.

phrases, etc. Several STS methods that use alignments in their calculations have been proposed in literature. Many of those methods were very successful and were among the top performing methods during the last years of SemEval 2013-2016 (Han et al., 2013; Han et al., 2015; Hänig et al., 2015; Sultan et al., 2014a; Sultan et al., 2014b; Sultan et al., 2015).

From this point, we present a sense-based STS approach that produces a similarity score between texts by means of a multilingual word-sense aligner. The following subsections describe in detail the main resource utilized in our STS approach, namely BabelNet (details in subsection 2.1), and our proposed word-sense aligner that our sense-based similarity method relies on (subsection 2.2).

2.1 BabelNet

BabelNet[1] is a rich semantic knowledge resource that covers a wide range of concepts and named entities connected with large numbers of semantic relations (Navigli and Ponzetto, 2010). Concepts and relations are gathered from different lexical resources such as: WordNet, Wikipedia, Wikidata, Wiktionary, FrameNet, ImageNet, and others.

BabelNet is made up of about 14 million entries called *Babel synsets*. Each Babel synset is a set of multilingual lexicalizations (each being a Babel Sense) that represents a given meaning, either concept or named entity, and contains all the synonyms which express that meaning in a range of different languages. For example, the concept '*A motor vehicle with four wheels*' is represented by the synset {car$_{en}$, auto$_{en}$, automobile$_{en}$, automobile$_{fr}$, voiture$_{fr}$, auto$_{fr}$, automóvil$_{es}$, auto$_{es}$, coche$_{es}$, otomobil$_{tr}$, araba$_{tr}$, سيارة$_{ar}$, مركبة$_{ar}$, عربة$_{ar}$}[2], this synset contains synonyms in English (EN), French (FR), Spanish (ES), Turkish (TR), and Arabic (AR) languages.

BabelNet semantic knowledge is encoded as a labeled directed graph, where vertices are Babel synset (concepts or named entities), and edges connect pairs of synsets with a label indicating the type of the semantic relation between them.

2.2 Word-Sense Aligner

Alignment is the task of discovering and aligning similar semantic units in a pair of sentences expressed in a natural language.

[1] http://babelnet.org/

[2] Each word is a Babel sense in the subscripted language.

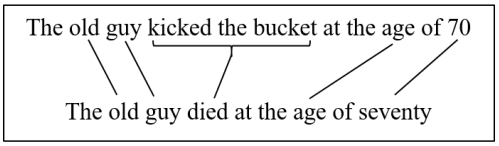

Figure 1: Token alignments using our aligner between monolingual English - English sentence pair example.

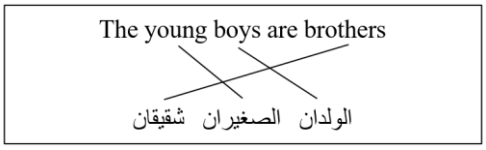

Figure 2: Token alignments using our aligner between cross-lingual English - Arabic sentence pair from SemEval 2017-Track2 dataset.

Our proposed multilingual aligner aligns tokens across two sentences based on the similarity of their corresponding Babel synsets. A token can be in the form of a single word or a multi-words token. When alignment of a single word token fails, its multi-words synonyms are retrieved from BabelNet. The proposed aligner aligns only a token that is neither a stop word nor a punctuation mark.

Figure 1 shows an example of alignments between English monolingual sentence-pairs using our aligner. In this figure the idiom "kicked the bucket" is considered as a single token of multiple words, and it was successfully aligned with the token "died" in the other sentence because both tokens are synonyms to each other in BabelNet. Figure 2 illustrates an example of direct token alignments between English-Arabic cross-lingual sentence pairs without using any machine translation module for translating one sentence language to the other.

Token-pairs are aligned one-to-one in decreasing order of their Babel synsets similarity score (s) using Equation (1). The most commonly used Babel synset of each token is selected.

$Al_{S1,S2} = \{(t, t', s) : t \in T_1, t' \in T_2, \text{and } s > \delta\}$; where T_i is a set of tokens of sentence i, and δ is a threshold parameter for alignment score ($\delta = 0.5$)[3]

2.3 Synset Similarity Measure

Finding similarity between synsets is a fundamental part of our aligner. Hence, we proposed a synset similarity measure based on the hypothesis

[3] According to experimental results conducted, we found that the best value for this threshold is 0.5.

that highly semantically similar concepts have high degree of common neighbor synsets. From this standpoint, this measure calculates the similarity between Babel synset pairs (bs_i, bs_j) based on the overlap between their directly connected synsets. The overlap-coefficient is used, which is defined as the size of the intersection divided by the smaller of the size of the two sets. That is:

$$sim_{synset}(bs_i, bs_j) = \frac{\left|NS_i \cap NS_j\right|}{\min(|NS_i|, |NS_j|)} \quad (1)$$

where NS_i and NS_j are the sets of all neighbor Babel synsets having a connected edge with bs_i and bs_j in the BabelNet network respectively. Since synonyms are belong to the same synset, their similarity score is equal to 1.

3 System Description

Our systems are based on the past successful integrated architecture of sense-based and surface-based similarity functions presented in SemEval-2015 system (Hassan et al., 2015). We use the integration in the latter system unchanged (Equation 2), where the current results are provided by taking the arithmetic mean of: 1) $sim_{proposed}$: a proposed sentence-pair semantic similarity score (differs in each Run, details in subsection 3.2), and 2) sim_{SC} : the surface-based similarity function proposed by Jimenez et al. (2012). Hence,

$$sim(S_1, S_2) = \frac{sim_{proposed}(S_1, S_2) + sim_{SC}(S_1, S_2)}{2} \quad (2)$$

The approach presented in (Jimenez et al., 2012) represents sentence words as sets of q-grams and measures semantic similarity based on soft cardinality computed from sentence q-grams similarity. Our system employs this approach, with the following parameters setup: p=2, bias=0, and α=0.5.

In this section, the text preprocessing details is firstly explained in subsection 3.1, and then each submitted Run is described in subsection 3.2.

3.1 Text Preprocessing

The given multilingual input sentences are preprocessed beforehand to map the raw natural language text into structured representation that can be processed. This process is including only four different tasks: (1) tokenization, (2) stopwords removal, (3) lemmatization, and (4) sense tagging.

Tokenization: is carried out using Stanford CoreNLP[4] (Manning et al., 2014), in which the input raw sentence text, in any language, is broken down into a set of tokens.

Stopwords removal: is the task of removing all tokens that are either a stop word or a punctuation mark.

Lemmatization: is a language-dependent task, in which each token is annotated with its lemma. English tokens are lemmatized using Stanford CoreNLP (Manning et al., 2014). Spanish tokens are lemmatized using a freely available lemma-token pairs dataset[5]. Arabic tokens are lemmatized using Madamira[6] (Pasha et al., 2014). For Turkish tokens, lemmatization is not carried out.

Sense tagging: is the task of attaching the Babel synsets (bs) to each sentence token (t). It is achieved by retrieving all the Babel synsets of token's lemma.

On completion of the text preprocessing phase, each sentence is represented by a set of tokens (T), in which each token (t) is annotated by its original word (t_{word}), lemma (t_{lemma}), and a set of Babel synsets (bs_t). This structured representation is then used as an input to our proposed aligner (subsection 2.3), and from which a set of aligned tokens across two sentences S_1 and S_2 is formed ($Al_{S1,S2}$).

3.2 Submitted Runs

We made two system submissions to participate in all the provided monolingual and cross-lingual tracks, named Run1 and Run2. Each run proposes a new different sense-based similarity method between sentence-pairs. The proposed similarity score is then applied in Equation (2), $sim_{proposed}$, resulting in the final similarity score between two sentences in each run. In the following, each of the two runs is described.

Run1: String Kernel with Alignments

A kernel can be interpreted as a similarity measure between two sentences, it is a simple way of computing the inner product of two data points in a feature space directly as a function of their original space variables (Liang et al., 2011). At SemEval 2015, a string kernel was presented, which relied on the hypothesis that the greater the

[4] http://nlp.stanford.edu/software/corenlp.shtml
[5] http://www.lexiconista.com/datasets/lemmatization/
[6] http://camel.abudhabi.nyu.edu/madamira/

similarity of word senses between two texts, the higher their semantic equivalence will be (Hassan et al., 2015). Accordingly, this run employs the string kernel presented in (Hassan et al., 2015) in which the alignments obtained from our proposed aligner is used in mapping a sentence to feature space. The changed kernel mapping function is given by:

$$\phi_t(S) = \max_{1 \le i \le n} \{ sim(t, t_i) \} \quad (3)$$

where $sim(t, t_i)$ is the alignment score s of the two tokens if $(t, t_i, s) \in Al_{S1,S2}$, and is equal to 0 otherwise, and n is the number of tokens contained in sentence S, i.e. $|T|$.

The normalized string kernel between two sentences S_1 and S_2 is calculated as follows (Shawe-Taylor and Cristianini, 2004):

$$\kappa_{NS}(S_1, S_2) = \frac{\kappa_S(S_1, S_2)}{\sqrt{\kappa_S(S_1, S_1)\kappa_S(S_2, S_2)}} \quad (4)$$

$$\kappa_S(S_1, S_2) = \langle \phi(S_1), \phi(S_2) \rangle = \sum_{t \in T} \phi_t(S_1) \cdot \phi_t(S_2)$$

where T is the set of all tokens in both S_1 and S_2.

Given two sentences, S_1 and S_2, our similarity score between S_1 and S_2 proposed by this run is the value of the normalized string kernel function between the two sentences (Equation 4). That is:

$$sim_{proposed}(S_1, S_2) = \kappa_{NS}(S_1, S_2) \quad (5)$$

Run2: Alignment-Based Similarity Metric

Alignment-based semantic similarity approaches presented in (Sultan et al., 2014a; Sultan et al., 2014b; Sultan et al., 2015) relied only on the proportions of the aligned content words on the two sentences. We hypothesized that alignments are not of the same importance, an alignment of synonym tokens with alignment score 1 is not the same as an alignment of two semantically related tokens with score 0.5. Hence, the proposed similarity score between S_1 and S_2 proposed for this run is based on the alignment scores as well as their proportion to the number of tokens in both sentences. It is given by:

$$sim_{proposed}(S_1, S_2) = \frac{2 * \sum_{al \in Al_{S1,S2}} al.s}{|T_1| + |T_2|} \quad (6)$$

where T_i is a set of tokens in sentence i, and $al.s$ is the score calculated for the alignment al.

Track	Run1	Run2	Baseline	Best Score
1 : AR-AR	.7158	**.7158**	.6045	.7543
2 : AR-EN	**.6782**	.6781		.7493
3 : SP-SP	.8484	**.8489**	.7117	.8559
4a: SP-EN	**.6926**	.6854		.8302
4b: SP-EN	**.0254**	.0214		.3407
5 : EN-EN	.8272	**.8280**	.7278	.8547
6 : TR-EN	**.5452**	.5390		.7706
Primary	**.6190**	**.6166**		**.7316**

Table 1: System performance on SemEval-2107 datasets.

4 Experimental Results

The main evaluation measure selected by the task organizers was the Pearson correlation between the system scores and the gold standard scores. Table 1 presents the official results of our submissions in SemEval2017-Task1 for both Run1 and Run2 in the six tracks as well as the primary track. The best performing score obtained in each track is included as well alongside with the baseline system results announced by the task organizers. Our best system (Run1) achieved 0.619 correlation and ranked the 10th run and the 5th team out of 84 runs and 31 teams respectively.

Although the performance of the two Runs differs slightly, it is noticeable from the table that Run1 (Kernel) performs better with cross-lingual sentence-pairs, while Run2 (Alignments) performs better with monolingual sentence-pairs. Hence, relying on aligned tokens only in cross-lingual sentences is insufficient.

5 Conclusions and Future work

Experimental results proved that, in spite of the fact that our proposed simple unsupervised approach relies only on BabelNet and token alignments, it is capable of assessing the semantic similarity between two sentences in different languages with good performance, 10th run rank and 5th team rank. Also, the proposed approach demonstrates the effectiveness and usefulness of using the BabelNet semantic network in solving the STS task. Some potential future work includes enhancing our proposed synset similarity method, and exploiting the extraction of promising content words in the given sentences.

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, pages 1–14, Vancouver, Canada.

Lushan Han, Abhay Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. UMBC EBIQUITY-CORE: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages. 44-52, Atlanta, Georgia, USA

Lushan Han, Justin Martineau, Doreen Cheng, and Christopher Thomas. 2015. Samsung: Align-and-differentiate approach to semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 172–177, Denver, Colorado, USA

Christian Hänig, Robert Remus, and Xose De La Puente. 2015. ExB Themis: Extensive feature extraction from word alignments for semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 264–268, Denver, Colorado, USA

Basma Hassan, Samir AbdelRahman, and Reem Bahgat. (2015). FCICU: The Integration between Sense-Based Kernel and Surface-Based Methods to Measure Semantic Textual Similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 154–158, Denver, Colorado, USA.

Sergio Jimenez, Claudia Becerra, and Alexander Gelbukh. 2012. Soft Cardinality: A Parameterized Similarity Function for Text Comparison. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 449–453, Montreal, Canada.

Yizeng Liang, Qing-Song Xu, Hong-Dong Li, and Dong-Sheng Cao. 2011. *Support Vector Machines and Their Application in Chemistry and Biotechnology*. CRC Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), pages 1094–1101, Reykjavik, Iceland.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014a. Back to Basics for Monolingual Alignment: Exploiting Word Similarity and Contextual Evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014b. DLS@CU: Sentence Similarity from Word Alignment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 241–246, Dublin, Ireland.

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, USA.

John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.