A Modified Annotation Scheme for Semantic Textual Similarity

by

darshan.agarwal , vandan.mujadia , Dipti Misra Sharma, Radhika Mamidi

in

18th International Conference on Computational Linguistics and Intelligent Text Processing
(*CICLing-2017*)

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2017

A Modified Annotation Scheme for Semantic Textual Similarity

Darshan Agarwal, Vandan Mujadia, Dipti Misra Sharma, and Radhika Mamidi

Language Technologies Research Center(LTRC)
Kohli Center On Intelligent Systems (KCIS)
International Institute Of Information Technology (IIIT-H)
Hyderabad, Telangana, 500032
{agarwal.darshan95, vmujadia}@gmail.com,
{dipti, radhika.mamidi}@iiit.ac.in

Abstract This paper presents an annotation schema for annotating semantic textual similarity. Given two sentences, the goal of the annotation is to give the similarity score between two sentences on a scale of 0 to 5. Annotators faced several difficulties in assigning similarity scores by following the [1] annotation scheme. To overcome those difficulties, we propose a new set of annotation guidelines which takes into account two major aspects of a sentence: events and entities. The semantic similarity score between a pair of sentences is assigned by finding the similarity of events and relations like hypernymy, co-hyponymy, meronymy, etc. between the entities in the sentences individually. Using our scheme we annotated the degree of semantic relatedness on 750 pairs of mononlingual Hindi sentences which were collected from newspapers, essays. We observed a significant improvement in inter-annotator agreement from 0.55 to 0.81 Fleiss' kappa measure.

1 Introduction

Semantic Textual Similarity (STS) is the degree of equivalence between two sentences semantically. We may also say, it is the ability to substitute one sentence for the other without changing its meaning. Textual similarity can range from exact semantic equivalence to complete unrelatedness, corresponding to quantified discrete values between 5 and 0. While 0 and 5 capture the extremes of textual similarity, the other scores capture the intermediate levels from pairs of texts differing only in some minor aspects of meaning to significant differences in entities and events. Semantic textual similarity has application in multitude of areas such as paraphrase recognition [2], automatic machine translation evaluation [3], machine translation [4], short answer grading [5], ontology mapping and schema matching.

Earlier, when there was no human annotated data set available, semantic textual similarity systems were unsupervised and had to be evaluated extrinsically on tasks like textual entailment recognition and paraphrase detection. The SemEval STS task series [1,6,7,8,9] has made an important contribution through the large human annotated data set for English, enabling intrinsic evaluation of STS systems and making supervised STS systems a reality.

Giving a semantic similarity score is a difficult task both manually and automatically because it is relatively easy to express the same idea in different ways. [1] scoring scheme provides the method to annotate in a very abstract way. In this paper, we try to make the annotation procedure easier for annotators by making the annotation process more well defined. So, our method divides the sentence into events and its participants and then get the similarity at the lower level. Since events play an important role in a sentence, we give more weight to event similarity than to participants. Due to non-availability of human annotated data for Hindi and in order to evaluate the STS system for Hindi, there was a need for semantic similarity annotated Hindi dataset. By following our scheme, we have created a monolingual Hindi corpus of 750 pairs of sentences, with their similarity scores annotated.

2 Corpus Creation

Pairs of valid sentences which are on the same topic, same meaning, or the sentences on different topics but on the same entity were collected manually from Hindi newspapers and essays [1]. A total of 750 pairs of sentences were collected. In order to balance the corpus, some (50) pairs of sentences which are not similar were also added to the corpus. The average word length of a sentence is 8. This is the first attempt on annotation of semantic textual similarity on general news domain corpus for Hindi language.

3 Annotation

3.1 Quality of Annotation

To assess the quality of annotation we find the inter annotator agreement using two methods. The first method followed is Fleiss' Kappa [10] which is the frequently used agreement coefficient for annotation tasks on categorical data. Flesis' Kappa measure is calculated as :

$$k = \frac{P - P_e}{1 - P_e}, \tag{1}$$

The factor $1 - P_e$ gives the degree of agreement that is attainable above chance and $P - P_e$ gives the degree of agreement actually achieved above chance. In the second method, we measured the Pearson product moment correlation [2] of each annotator with the average of the rest of the annotators. The average of all the correlations is taken as the final agreement score.

[1] http://khabar.ndtv.com/, http://hindi.news18.com/, etc.

[2] https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient#Calculating_a_weighted_correlation

3.2 Annotation with Existing Schema

100 pairs of sentences were randomly picked from the collected corpus for annotation. Four language annotators, well versed in linguistics were chosen for the task. Annotators were presented with detailed instructions as mentioned in [1], to label each sentence pair on a 6 point scale from 0 (different topics) to 5 (completely equivalent). The scheme is as follows:

(5) The two sentences are completely equivalent, as they mean the same thing.
पंछी तालाब में स्नान कर रहा है ।
panchii taalaab meN snaan kar rahaa hai.
The bird is bathing in the lake.
पंछी अपने आप को तालाब में धो रहा है ।
panchii apane aap ko taalaab meN dho rahaa hai.
Bird is washing itself in lake.

(4) The two sentences are mostly equivalent, but some unimportant details differ.
बीजेपी सांसद शत्रुघ्न सिन्हा ने शनिवार रात करीब १० बजे सीएम नितीश कुमार से उनके सरकारी आवास पर मुलाक़ात की ।
biijepii saaMsad shatrughan sinhaa ne shanivaar raat kariib 10 baje CM nitish kumAr se unake sarkaarii aavaas par mulaakaat kii.
Shatrugan Sinha, the member of BJP, met Nitish Kumar at his official residence on Saturday night at 10 pm.
सिन्हा और कुमार के बीच मुलाकात मुख्यमंत्री आवास में हुई ।
sinhaa aur kumaar ke beecha mulaakaat mukhyamantrii aavaas meN huii.
The meeting between Sinha and Kumar was held in the Chief Minister's residence.

(3) Two sentences are roughly equivalent, but some important information missing/differs.
राम को गवाह माना जा रहा है संदिग्ध नहीं ।
raam ko gavaah maanaa jaa rahaa hai sandigdh nahiN.
Ram is considered a witness but not a suspect.
अब राम कोई संदिग्ध नहीं रहा ।
aba raam koii sandigh nahiN rahaa.
Ram is not a suspect anymore.

(2) The two sentences are not equivalent, but share some details.
वे दोनों पेड़ पर चढ़ रहे है ।
ve donoN peD par chaDh rahe haiN.
Both of them are climbing the tree.
वे दोनों पेड़ से उतर रहे है ।
ve donoN peD se utar rahe haiN.
Both of them are climbing down the tree.

(1) The two sentences are not equivalent, but are on the same topic.
महिला वायलिन बजा रही है ।
mahilaa vaayalin bajaa rahii hai.
The woman is playing violin

उस खूबसूरत स्त्री को गिटार सुनने में आनंद मिलता है ।
usa khubsurat strii ko giTaar sunane meN aanand miltaa hai.
The beautiful lady enjoys listening to the guitar.

(0) The two sentences are on different topics.
याकूब मेनन की फांसी के विरोध में सुपरस्टार सलमान खान भी आ गए है ।
yaakuub menan kii faaNsii ke virodh meN suparsTaar salmaan khaan bhii aa gaye hai.
Superstar Salman Khan has also come against the execution of Yakub Menon
आपने बिहार को स्पेशल स्टेटस देने का वादा किया था ।
aapne bihaar ko speshal sTeTas dene kaa vaadaa kiyaa thaa.
You had promised to give special status to Bihar.

Need for Scheme modification To assess the quality of the annotation, inter annotator agreement was obtained by the methods discussed in section 3.1. 0.55 agreement using Fleiss' Kappa measure and 60% inter tagger correlation using the second method was obtained.

This low agreement score lead us to the pre-assumptions that there are problems either in scheme or annotators or in the data. Since the data was collected from the known newspapers, we believe that the quality of data is good. To check where the agreement was mainly conflicting, we analyzed the annotated data manually. The maximum disagreement was encountered at scores 1, 2 and 3, 4. The disagreement at the scores 3, 4 was due to the difference in consideration of information missing/differing as important or unimportant.

(1.1) बीजेपी सांसद शत्रुघ्न सिन्हा ने शनिवार रात करीब १० बजे सीएम नितीश कुमार से उनके सरकारी आवास पर मुलाक़ात की ।
biijepii saaMsad shatruughan sinhaa ne shanivaar raat karib 10 baje CM nitish kumaar se unake sarkaarii aavaas par mulaakaat kii.
Shatrugan Sinha, the member of BJP, met Nitish Kumar at his official residence on Saturday night at 10 pm.

(1.2) सिन्हा और कुमार के बीच मुलाकात मुख्यमंत्री आवास में हुई ।
sinhaa aur kumaar ke beech mulaakaat mukhyamantrii aavaas meN huii.
The meeting between Sinha and Kumar was held in the Chief Minister's residence.

(2.1) घर जाते हुए मेरा टायर पंक्चर हो गया ।
ghar jaate huye meraa Taayar paMchar ho gayaa.
My tyre got punctured on my way to home.

(2.2) कॉलेज जाते हुए मेरी गाड़ी खराब हो गयी ।
kolej jaate hue merii gaadhii kharaab ho gayii.
My car broke down on my way to college.

In the above two sentence pairs, there was a disagreement between the annotators at scores 3 and 4. One group of annotators agreed with the score of 3- *The two sentences are roughly equivalent, but some important information differs/missing*, assuming that the time and location in (1) and finer details about how car broke down in (2) are important details which are missing in the other sentence. Where as the other group of annotators gave score as 4- *The two sentences are mostly equivalent, but some unimportant details differ*, considering the missing details as unimportant. Thus, there was confusion as when to consider the sentences as roughly equivalent and when mostly equivalent.

Similarly, there was also disagreement observed at scores 1- *The two sentences are not equivalent, but are on the same topic.* and 2- *The two sentences are not equivalent, but share some details* may be due to the difference in consideration of similar information as sharing details or only on same topic. This can be observed in the following example

1. वे लोग क्रिकेट खेल रहे है ।
   ve log krikeT khel rahe hai.
   They were playing cricket.
2. वे लोग फुटबॉल खेल रहे है ।
   ve log fuTbaal khel rahe hai.
   They were playing football.

Annotators were also confused in giving scores for sentences, where one sentence has more than two participants and they are missing in other sentence. As observed in the following example, there are two participants aDvaanii *(Advani)* and modii *(Modi)* in the first sentence, while there is a single participant performing the same event in the second sentence.

1. अडवाणी और मोदी ने दिल्ली में चुनावी रैली की ।
   advaaNii aur modii ne dillii meN chunaavii rally kii.
   Advani and Modi did an election rally in Delhi.

2. आडवाणी ने दिल्ली से चुनावी रैली आरम्भ की ।
   advaaNii ne dillii meN chunaavii rally aarambh kii.
   Advani started his election rally from Delhi.

The disagreement was also seen in some of the cases where few annotators used their world knowledge while annotating.

3.3 Modified Annotation Scheme

In this section, we propose an annotation scheme to make the scoring process simpler with less ambiguity. We bring semantics of the sentence into consideration. As shown in figure 1 we divided the sentence into two broader categories, events and its participating entities. We have further sub-divided the entities category into core and non-core. The entities which are core to the event and entities which are not important with respect to the event (like time and location). The events category is divided into completely equivalent, mostly equivalent, roughly equivalent and not equivalent. The definition of these are kept similar to [6]. We have not further sub-divided the events section, since on an average the number of events in the sentences are not more than 2. So, there is less ambiguity in comparison of events. We have divided events into only 4 sub categories instead of 6 (0-5), since if the events are not equivalent, there cannot be any further division into it. The core entities are divided into 3 sub-categories : Exact Match, Partial Match and No Match as shown in figure 1.

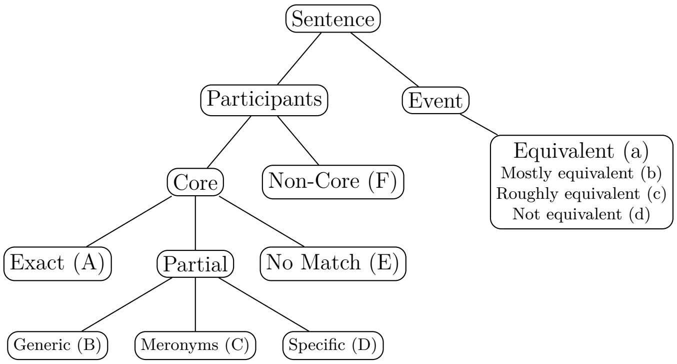

Figure 1: Semantic-tree for annotation

Exact (A) The entities which are completely similar, come under the ”Exact match” category (like *raam* (ram) and *raam* (ram)).

Partial In the ”Partial Match” category the entities which are partially similar are considered. The ”Partial” category is sub divided into three categories : Generic (Hypernym), Specific (Co-hyponyms) and Meronym.

Generic / Hypernyms (B) The generic category is the case of hypernyms where the relationship between the entities is of hypernymy. An entity is represented in one of the sentences in a generic sense like *senaa* (army), while in other sentence the entity is mentioned in a more specific sense like *bhaaratiya senaa* (Indian Army). In this category, the semantic field of the entity is less broad than the semantic field of entity in the other sentence. For example, in the following two sentences :

1. लोग ईद के दिन गले मिलते है।
   log iida ke din gale milate hai.
   People hug each other on the occassion of Eid.
2. ईद के दिन रंजना और शाहरुख गले मिले।
   iida ke din ranjanaa aur shahrukh gale mile.
   On the occasion of Eid, Ranjana and Shahrukh hugged each other.

Here, the participating entities *log* (people) and *ranjanaa aur shahrukh* (Ranjana and Shahrukh) for the event *gale milanaa* (hugging each other) are similar with respect to hypernym relation. Since at the semantic level, the entities *ranjanaa aur shahrukh* (Ranjana and Shahrukh) are part of the broader entity *log* (people).

Meronym (C) This subcategory is for the case of ”Meronymy” where an entity in one sentence denotes a constituent part of, or a member of entity in another sentence. The entities *Tayar* (Tyre) and *gaaDii* [Car] come under this category where *tyre* is a meronym of *car*. The similarity score for meronym should be given less than Generic. Since in the meronym category, the meaning of the sentence might change, if the entity is replaced by its meronym. While in the Generic category, the sentence meaning remains the same irrespective of the replacement of the entity with its hypernym.

Specific / Co-Hyponyms (D) The specific category is for the case, where the relationship between the entities is of co-hyponyms. These entities are not equivalent. The entities in both the sentences are mentioned in a specific sense and share the same hypernym but are not hyponyms of one another. For example *bhaaratiya senaa* (Indian army) and *cheenii senaa* (Chinese army) come under the specific category, since they are two different entities but both entities are hyponyms of the hypernym ”army”. The specific sub-category should be given lesser score than B and C, since here the entities are not equivalent, but share a common hypernym.

No Match (E) The core entities are also sub-divided into *No match* category, which is for the case when the participating entities are totally different in both the sentences.

Score	a	b	c	d
A	5	4	2	1
B	4.5	3.5	1.5	0.5
C	4	3	1	0.5
D	3	2.5	0.5	0
E	3	2	0	0

Table 1: Semantic similarity scores

To calculate semantic similarity between two sentences, the annotators are first required to figure out the event similarity (Figure-1 indices a, b, c, d). The next step would be to find the relationship between the entities in the sentences. For each derived entity in a sentence, the matching entity with the same semantic role in the other sentence is found, and assigned a similarity level (A-E). We then consult the lookup table (Table 1) and calculate the similarity score for each entity pair with respect to the event similarity obtained earlier. The minimum of all entity pair scores is assigned as the final similarity score between the two sentences. When core entities are missing in one of the sentences, we subtract 1 from the final score. In the cases, where two or more than two entities do not match then the score obtained from table 1 is subtracted by 1. Here, subtraction of 1 is done by taking the other values of the lookup table into consideration. Table 1 gives the similarity score between two entities by taking event similarity into consideration. Each column corresponds to the similarity between events and the rows denote the similarity between the entities. Table 1 is made by considering the fact that events are more important in a sentence than entities. Thus, reducing the scores proportionally by taking the entity similarity order as A > B > C > D > E.

3.4 Annotation with Modified Scheme

The annotators were presented with detailed instructions of the method discussed in section 3.3 as annotation guidelines. Each of the 4 annotators gave scores to 750 pairs of sentences by following the new scheme. We achieved an inter annotator agreement of 0.83 Fleiss' Kappa measure on earlier 100 pair of sentences, 0.81 Fleiss' Kappa and 75% Pearson correlation on remaining 650 pair of sentences. With the good agreement and also through the survey taken by the annotators, it reflected that this scheme had made the task much easier. But the annotators also gave the feedback that the process has become more time consuming. Upon analysis, it was seen that there was disagreement in some of the cases, where word knowledge is used. As in the following pair of sentences, few of the annotators had world knowledge that king of Ayodhya is treated as God in Hindu mythology, so they gave score as 4, treating discussion is on the same person, while there were also annotators who assumed in both the sentences the discussion is on different persons and gave score as 3.

1. राम भक्तों के अनुसार दीवाली वाले दिन अयोध्या के राजा राम लंका के अत्याचारी राजा रावण का वध करके अयोध्या लौटे थे ।
   raam bhaktoN ke anusaar diivaalii vaale din ayodhyaa ke raaja raam laNkaa ke atyaachaarii raajaa raavaN kaa vadh karake ayodhyaa lauTe the.
   According to his devotees, Lord Rama, the king of Ayodhya returned home on the day of Diwali after killing the tyrannical king of Lanka Ravana.
2. हिंदू मान्यताओं में राम भक्तों के अनुसार भगवान श्री रामचंद्रजी असुरी वृत्तियों के प्रतीक रावणादि का संहार करके अयोध्या लौटे थे ।
   hindu maanyataaoN meN raam bhaktoN ke anusaar bhagvaan shrii ramchandrajii asurii vrattiyoN ke pratiik ravaNaadi kaa saNhaar karke ayodhyaa lauTe the.
   In Hindu beliefs, according to the devotees of Ram, Lord Shri Ramchandra returned to Ayodhya after killing the symbol of daemonic instincts like Ravana

Score	Number of Pairs
5.0	63
4.5	78
4.0	138
3.5	160
3.0	57
2.5	58
2.0	49
1.5	39
1.0	39
0.5	35
0.0	32

Table 2: Semantic similarity scores

3.5 Data Statistics

In the final corpus, the score which had a maximum agreement among the annotators was assigned to each sentence pair. Table 2 shows the distribution of scores for the sentence pairs. From Table 2, we can see that the data distribution is such that sentence pairs having similarity score as 3.5 and 4 are highest, while pairs with other scores are distributed uniformly.

4 Conclusion & Future Work

In this work, we have discussed an ongoing effort of building a semantic textual similarity corpus for Hindi. The present size of this corpus is around 750 pairs of sentences which is annotated by four annotators. We discussed the need for modification of annotation scheme for sentences with higher frequency of words. By this work, we try to open dialogue on existing annotation approaches for semantic textual similarity and also proposing the modified annotation scheme for the same. We also discussed the evaluation of the corpus by measuring the inter-annotator agreement between the two annotators using the Kappa and Pearson correlation statistics. The agreement calculated here is considered to be reliable and substantial ensuring that the annotation of the Hindi corpus for semantic similarity is consistent. We aim at increasing the size of the semantic similarity corpus and we intend to create the similar corpus for other Indian languages. In future, we would also like to analyze further why there has been disagreement among annotators and also further guide the annotators. In future we also want to take these annotation more fine-grain on the aspect of semantic roles and scoring method.

References

1. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: Semeval-2012 task 6: A pilot on semantic textual similarity. In: Proceedings of 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, Association for Computational Linguistics (2012) 385–393

2. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics (2004) 350
3. Kauchak, D., Barzilay, R.: Paraphrasing for automatic evaluation. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics (2006) 455–462
4. Sachdeva, K., Sharma, D.M.: Exploring the effect of semantic similarity for phrase-based machine translation. ACL-IJCNLP 2015 (2015) 41
5. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics (2009) 567–575
6. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In: In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, Citeseer (2013)
7. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2014 task 10: Multilingual semantic textual similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). (2014) 81–91
8. Agirre, E., Banea, C., et al.: Semeval-2015 task 2: Semantic textual similarity, english, s-panish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), June. (2015)
9. Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., Wiebe, J.: Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. Proceedings of SemEval (2016) 497–511
10. Randolph, J.J.: Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. Online submission (2005)