Algorithm to Identify the Connection between Sentences

Praveen Sankarasubramanian[#1], Dr.E.N.Ganesh[#2]
[#1]*Research Scholar, Vels Institute of Science, Technology and Advanced Studies, (VISTAS),Chennai, Tamilnadu, India.,*
[#2]*Dean School of Engineering, VISTAS, Chennai, Tamilnadu, India.*
[1]praveengrb@gmail.com
[2]enganesh50@gmail.com

***Abstract*— Many applications require the affiliation of sentences (which includes text summarization, answering questions, producing natural language, analyzing natural language, and text clustering). The similarity of terms may be improved using the algorithm. Estimating the relationship among phrases at some point of the NLP data interpretation remains a notably critical problem. This paper suggests the concept of accomplice similar sentences. First, define the relations that could be used for text clustering. This could be identified by various algorithms like support vector machine, and other algorithms Secondly, terms are correlated and their similarities are recognized. Finally, similar sentences are clustered.**

***Keywords*— Sentence Relationship, NLP, Sentence Similarity, Submission**

I. INTRODUCTION

Every day the information grows vastly. This has become a motivation for many researchers to expand the text processing system. The interpretation of the sentences is vital to retrieving relevant records from the textual content. subsequently, the knowledge of sentence relation is prominent within the NLP.

II. DEFINING THE RELATIONSHIP BETWEEN THE SENTENCES

Consider two sentences S1 and S2.
As per cross-document structure theory (CST), the sentences can be classified as

1. Two sentences could be **identical** (Identity) and can have the same information.
   S1: India is a democratic country
   S2: India is a democratic country
2. Both sentences could represent the same context in different way **Paraphrase (Equivalence)**
   S1: There are ten people in a queue. Alex Standing as the eighth person.
   S2: Alex stands as second from the last in a 10 person's queue.
3. The same context is explained (**translated**) in a different language.
4. Both the sentences will share the same information, where either one of the sentences could be a "**subsumption**" of the other. It could provide additional information about the context.
   S1: School starts at 9 am
   S2: School starts at 9 am and the main gate will be closed after that.
5. Either of the sentences could provide conflicting, **contradictory** information about the context
   S1: The bus "A" carried 30 passengers.
   S2: The bus "A" drowned with 30 passengers.
6. The first sentence could provide a fact, evidence, historical background about any entity mentioned in the second sentence.
   S1: In the last 5 years, none of the students was able to clear the course X in their first attempt.
   S2: John cleared course X in his first attempt.
7. One sentence explicitly **cites** the other one.
   S1: Last month a report stated that the machinery would be under maintenance today.
   S2: Today a circular came that the machines would be under maintenance.
8. One sentence presents a **qualified version (Modality)** of another sentence.
   S1: Mr. X is reported to own several immovable assets.
   S2: Mr. X owns an estate at Darjeeling.
9. One sentence could be the **attributed version (Attribution)** of another

10. One sentence could be the **summarized version** of the other
    S1: Mr. Smith visiting the Taj Mahal second time.
    S2: Just like today, Two weeks before Mr. Smith visited the Taj Mahal.
11. The first sentence presents additional information, which has happened (**follow-up**) since the second sentence.
    S1: More than one hundred passengers are missing from the submerged flight.
    S2: So far, no casualties from the accident have been confirmed
12. Both sentences could refer to the same context in a direct and indirect way. Either of the sentences **indirectly** quotes the same context which was directly quoted in the other sentence
13. Both sentences could refer to the same context and the second sentence could **elaborate or refine** the details provided in the first sentence.
    S1: 50% of the students scored more than 70%, 30 % of the students scored more than 60%, 20% of the students were unable to clear the exam.
    S2: 80% of the students cleared the examination.
14. One sentence asserts (**fulfill**) the occurrence of the event predicted in the other.
    S1: This Thursday, Prime minister will meet Governor and then he discusses about the welfare of the state.
    S2: This Thursday, Prime minister will meet Governor
15. Either one of the sentences **describes** the entity mentioned in the other.
    S1: This Thursday, Prime minister will meet Governer and then he discusses about the welfare of the state.
    S2: This Thursday, Prime Minister will meet Governor
16. Both the sentences represent the same context written for a different set of target readers (**reader profile**)
    S1: S, a sweet that has a lot of dry fruits, sugar, milk and milk solids.
    S2: This sweet S is made with milk solids.
17. The same context could be explained from a **different perspective**
    S1: Mr. X was unhappy with the dressing style of Ms. Z
    S2: Mr. X praised Ms. Z for her beautiful client presentation.
18. The sentences could have **partially overlapping sentences**.
    S1: The building Marina has ten beautiful luxury apartments.
    S2: The billionaire Richie Rich owns ten luxury apartments at Marina.
19. No relationship Between the sentence
    S1: The sky is blue
    S2: Sun is red

III. IDENTIFYING THE STRUCTURE OF THE SENTENCES

A. Properties of the Sentences

The type of information described in two text chunks is crucial to describe the type of relationships. The property of a sentence changes based on the entities (like a person, location, date, organization, time, money, percentage), number of conjunctions, length of the sentences, type of the speech.

B. Named Entity Recognition (NER)

Location: This entity often describes information such as the place where the event occurred. It could be a business, organization, natural location, or locale. For example Company location, Government sector, lake or mountain and so on.
Date and Time entity: This entity often describes when the incident took place. It could be a calendar event (morning, evening), natural (summer, winter)
Person: This entity describes the parties involved in the event.

C. Number of Conjunctions

Conjunctions play an important role in identifying the relationship between the sentences. There are around 40 types of conjunctions. The number of conjunctions appearing between the two sentences is identified. If more conjunctions appear in a sentence pair then there could be a higher chance of similarity. So maximum weight should be given to this sentence. For example, if we represent it with a value between zero and one. Then Sentence pair with a higher number of conjunctions will be set to one else zero.

D. Length of Sentence

Split the sentence by whitespace and form a collection of words. Number of words in a sentence denotes the length of the sentence.

E. Type of Speech

We determined the type of speech, whether the text span, S1 cites another sentence by detecting the occurrence of quotation marks to identify Citation or Indirect Speech which are the sub-category of Identity.

IV. FINDING THE SIMILARITY BETWEEN SENTENCES

A. Finding the cosine similarity, term frequency, and inverse document frequency

Consider three sentences

Sentence 1: Learning is an everlasting process
Sentence 2: Life is a game of everlasting learning
Sentence 3: Please do not stop learning

The term frequency measures the number of times a word occurs in a sentence pair.

In reality, sentences or documents could be of different sizes. On a large document or sentence, the frequency of the terms would be higher. There is a need for normalizing the sentences. To achieve the best result, stop words, special characters like $# are removed, words are lemmatized. Then tf is calculated. To

Normalization: In the normalization process, term frequency is divided by the total number of terms in the document. For example, Sentence 1 has five words or terms and the term learning appears once. Hence, the normalized term frequency is 1/5

The main purpose of using Inverse document frequency is to find out the relevant documents matching the given term.

All terms in the sentence are considered as equally important.

IDF of a term is equal to the log base e (total number of sentences or document/number of sentence or document with a given term in it) plus one.

IDF (Learning) = 1+ loge (3/3) =1
TF*IDF is equal to 1/5
TF and IDF for other terms are calculated.

The cosine similarity metrics measures the correlation between two sentences. This helps to identify the similarity between words, verbs, adjectives, nouns, and bigrams.

Different word sequences provide a different meaning. The ordering of the word determines the semantic meaning in the sentence. Thus, there is a need for bigram similarity. This is used to identify the semantic meaning of the sentences.

B. Overlap Ratio of Words between the First Sentence and the Second Sentence and Vice Versa

The overlap ratio helps to identify whether words in the second sentence also appear in the first sentence.

Algorithm:

1. Number of words in sentences are identified (WS1, WS2)
2. Calculate the number of common words in two sentences (CoWs)
3. Identify the overlapping word ratio (OWRS) by
   OWRS1 = (CoWs*2)/WS1
   OWRS2 = (CoWs*2)/WS2

OWRS ranges from zero to one where the former represents the least overlapping and latter represents higher overlapping.

This calculation determines how much the sentence matches with each other. This helps us to identify the citations, indirect speech, subsumption.

C. The longest matching word sequence against the first sentence

Identifying the longest matching word sequence will benefit the identification of overlapping sentences or subsumption.

Algorithm:

1. Number of words in the sentence are identified and it is represented as WS1 and WS2
2. List of matching word sequence (MWS) from the sentences is identified.
3. Longest Matching Word Sequence (LMWS) is identified and its length is identified (LLMWS) For example, if two sentences have 15 words each and if they have 8 recurring consecutive word pattern then the longest word occurrence would be 8
4. Longest common word sequence ration is identified by the formulae
   LCWS1 = (LLMWS) / (WS1)
   LCWS2 = (LLMWS) / (WS2)

D. The Grammatical Relationship between the Words in the Paragraphs.

The grammatical relationship between the words provides information about the context of the sentences. This approach helps to identify the change of perspective, descriptive sentence pair. To achieve this, algorithms like SVM are used.

Algorithm:

1. Number of subjects in First Sentence(NSS1) and Second Sentence(NSS2) are identified
2. Number of common subjects are identified (NCS)
3. Number of Objects in the first sentence(NOS1) and the second sentence(NOS2) are identified
4. Number of common objects(NCO) are identified
5. Subject Rate(SR) is calculated using the formulae
   SR1 = NCS/NSS1
   SR2 = NCS/NSS2
6. Object Rate (OR) are Identified by the formulae
   OR1 = NCO/NOS1
   OR2 = NCO/NOS2
7. Number of Noun in a sentence (NNS1 and NNS2) are identified
8. Descriptive sentence ration(DSR) is calculated using the formulae
   DSR1= (NCS*NNS1)/NSS1
   DSR2= (NCS*NNS2)/NSS2

CONCLUSION

Many applications require the affiliation of sentences (which includes text summarization, answering questions, producing natural language, analyzing natural language, and text clustering). Estimating the relationship among phrases at some point of the NLP data interpretation remains a notably critical problem.. This paper suggested the concept of accomplice similar sentences. This paper suggested a way by finding the similarity of sentence may be improved by the suggested steps. Structure of the sentences like property of the sentence, the length of the sentence, number of conjunctions, part of speech, and NER types was identified. Semantic relationship of the sentence was identified by finding the cosine similarity, term frequency,grammatical relations based on these criteria the similar sentence from a text chunk is clustered.

FUTURE OF WORK

This research is in prototype phase. Suitable algorithm to detect the grammatical relationship between the sentences need to be identified.
Lightweight question identification framework, and word net need to be added to improve the context based text clustering.

REFERENCES

[1] Exploiting Discourse Relations between Sentences for Text Clustering Nik AdilahHaninZahri* Fumiyo Fukumoto SuguruMatsuyoshi Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan g09dhl03*,fukumoto,sugurum@yamanashi.ac.jp
[2] Calculating the similarity between words and sentences using a lexical database and corpus statistics AtishPawar, Vijay Mago
[3] M. C. Lee, J. W. Chang, and T. C. Hsieh, "A grammar-based semantic similarity algorithm for natural language sentences," The Scientific World Journal, vol. 2014, 2014.
[4] A. Islam and D. Inkpen, "Semantic text similarity using corpus based word similarity and string similarity," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 2, no. 2, p. 10, 2008.
[5] A. Freitas, J. Oliveira, S. ORiain, E. Curry, and J. Pereira da Silva, "Querying linked data using semantic relatedness: a vocabulary independent approach," Natural Language Processing and Information Systems, pp. 40–51, 2011.
[6] Radev, D.R., Ottherbacher, J. and Zhang, Z., "CSTBank: A Corpus for the Study of Cross-document Structural Relationship", In Proc. of Language Resource and Evaluation Conference (LREC), 2004.
[7] Wolf, F., Gibson, E., Fisher, A. and Knight, M.," Discourse Graphbank", Linguistic Data Consortium, Philadelphia, 2005.
[8] Vapnik, V., "The Nature of Statistical Learning Theory", Springer, New York, 1995.
[9] Marcu, D., "From Discourse Structures to Text Summaries", In Proc. of the Association for Computational Linguistics (ACL) on Intelligent Scalable Text Summarization, pp. 82-88, 1997.
[10] Radev, D.R., Jing, H., Sty,M., and Tam, D.,"Centroid-based Summarization of Multiple Documents", Inf. Process.Manage.(40), pp. 919-938, 2004.
[11] Zhang, Z., Blair-Goldensohn, S. and Radev, D.R., "Towards CST-enhanced Summarization", In Proc. of the 18th National Conference on Artificial Intelligence (AAAI) , 2002.
[12] Uzêda, V.R., Pardo, T.A.S., Nunes, M.G.V.,"A Comprehensive Summary Informativeness Evaluation for RST-based Summarization Methods", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM) ISSN: 2150- 7988 Vol.1, pp.188-196, 2009.
[13] Louis, A., Joshi, A., and Nenkova, A., "Discourse Indicators for Content Selection in Summarization", In Proc. of 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pp. 147-156, 2010.
[14] Litkowski, K., "CL Research Experiments in TREC-10 Question Answering", The 10th Text Retrieval Conference (TREC 2001). NIST Special Publication, pp. 200-250, 2002.
[15] Verberne, S., Boves, L., and Oostdijk, N., "Discourse-based Answering of WhyQuestions", TraitementAutomatique des Langues, special issue on Computational Approaches to Discourse and Document Processing, pp. 21-41, 2007.
[16] Theune, M., "Contrast in Concept-to-speech Generation", Computer Speech & Language,16(3-4), ISSN 0885-2308, pp. 491-530, 2002.
[17] Piwek, P. and Stoyanchev, S., " Generating Expository Dialogue from Monologue Motivation, Corpus and Preliminary Rules", In Proc. of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2010.