ROLE OF NAMED ENTITIES IN UNDERSTANDING SEMANTIC SIMILARITY OF ENGLISH TEXT

Sumit Kumar and Shubhamoy Dey
Indian Institute of Management Indore, Rau-Pithampur Road, Indore – 453556, India

ABSTRACT

Understanding semantic similarities between documents is challenging but have enormous benefits, like plagiarism detection and information retrieval. Various techniques are available in Natural language processing, which help in understanding similarities between text documents. Every approach aims to find a unique set of features that help differentiate between two or more documents.

Names of persons, organizations, locations, medical codes, acronyms, technical terms, date & time expressions, quantities, monetary values, and percentages (collectively known as Named Entities) and the order in which they appear in a document contribute a great deal to the uniqueness of the document (Li et al., 2020). If two documents share them, they must present the same information or discuss the same concept. Another advantage of Named Entities (NE) in the context of plagiarism detection is that they do not have synonyms – replacing words with their synonyms to avoid detection is, therefore, not an option. Thus, NEs have a high potential for detecting similarities between documents. Yet, going by the availability of literature, it is an under-researched concept.

In this article, we discuss and explore the concept of NEs and their meta characteristics, and propose a way of using that information to find similarities between documents.

Our initial experimental results, discussed in this article, demonstrate the efficacy of the approach intuitively argued above. This article is unique in its methodology, thus comparing the results with other available methods on textual similarity is inappropriate. We have compared the results of the proposed NE based approach with existing approaches based on Term Frequency and TF-IDF.

The future goal of the ongoing research work is to combine NEs and their meta characteristics with other characteristics to develop a robust and comprehensive framework for finding semantic similarities between documents.

KEYWORDS

Text Similarity, Semantic Similarity, Named Entity

1. INTRODUCTION

We use the concept of semantic text similarities in day-to-day life without being aware of their existence. Human civilization will progress further if machines can understand human language. Finding similarity is one step in the direction where machines can understand human language. For example, applications like human-machine interfaces (such as Seri, Alexa or Cortana) can become more flexible and useful. They depend on how machines find similarities between human language and their trained modules. Finding semantic similarities has various other applications in information retrieval, automatic question answering, machine translation, dialogue systems, and document matching (Pradhan et al.,2015).

The general approach is to extract features unique to each document and convert these features into numerical representations that can be easily converted into machine language. Named Entities are among the unique features of documents. Named Entity Recognition (NER) are a set of techniques that help in finding these unique words present in documents. Grishman and Sundheim (1996) discussed named entities in the 6th Message Understanding Conference. They have used NER to identify persons, places, or organization names, dates, currency, and numerical values. After many advancements in the NE proposed subsequently, researchers have segregated NEs into two broad categories: Generic NEs and domain-specific NEs. NER uses one of these four methods: Rule-based method (on hand-crafted rules), unsupervised (based on unlabelled data), feature-based supervised learning (feature engineering later followed with supervised learning approach), and deep learning based approach (detect similar patterns of words on which a model trained) (Li et al.,2020).

2. NAMED ENTITY RECOGNITION

"a named entity is a word or a phrase that identifies one item from a set of other items with similar attributes." (Sharnagat, 2014, pp.1-27). Examples of named entities are organizations, persons, location names, gene, protein, drug and disease names in the biomedical domain—the fundamental of NER is to find the location of such words and tag the words in the document.

Formally defining NER: consider a document with a unique set of words S = { W1, W2, W3, ..., WN}, where a subset of those words are a list of NEs, i.e. { I1, I2,... IM}.

At the time of MUC-6 (Grishman, & Sundheim, 1996), each NE was assigned a single tag. This was popularly known as coarse-grained NER. Another kind of NER developed recently is known as fined-grained NER, which can allocate more than one tag to some of the NEs where required.

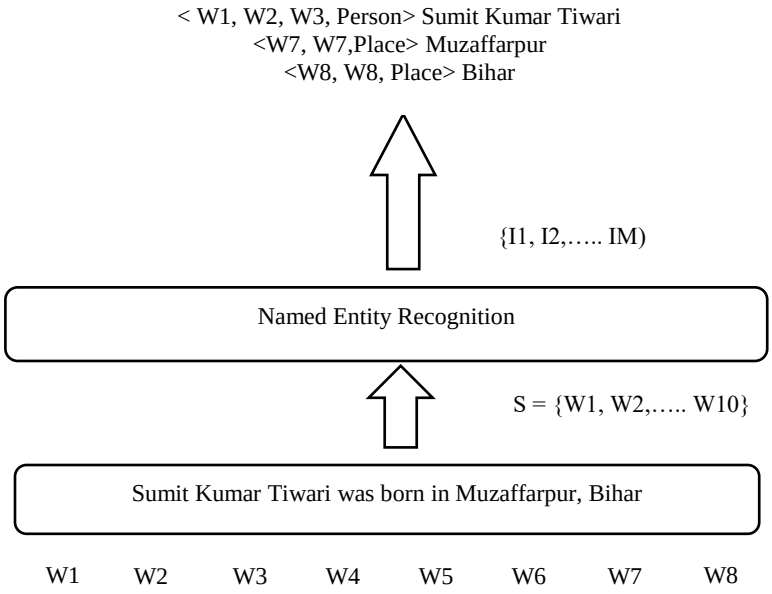

Figure 1. Illustration of named entity recognition

3. PROPOSED METHODOLOGY

3.1 Pre-Processing Prior to NER

Generally, while writing any document, if any proper noun or common noun repeats itself, the document's writer replaces it with a pronoun. The pronouns hide some important named entities and their location within the documents. It therefore, makes the task of NER more challenging. Fortunately, there are many tools and techniques available in Natural Language Processing (NLP) to resolve pronouns. Coreference resolution is a technique for finding all the expressions referring to the same entity in a document (Mitkov, 2001). Researchers have been working on the Co-reference resolution problem since the late 1970s when Hobb's naïve algorithm was proposed. It is a rule-based algorithm that focuses on syntactical characteristics of language to capture the NEs present in the form of pronouns. Advancement of statistical and machine learning techniques in the early 1990s led to the development of several new methods for coreference resolution. Probabilistic techniques such as the Bayesian rule, decision tree and genetic algorithm were some of the early approaches used for pronoun resolution (Aone & William, 1995; McCarthy & Lehnert, 1995). The next stage of development was the neural network and deep learning based approaches – eliminating the dependency on handcrafted rules. These models were able to understand the semantic structure. Clark and

Manning (2016) and Lee et al. (2017) have done some significant work in the area. Notably, Lee et al. (2017) have modelled end-to-end coreference resolution systems using character-level convolutional neural networks.

Our work uses a state-of-the-art method to resolve the coreferences. It is the spanBERT method based on the concept of span, where each span is numerically embedded. These spans are useful in reducing the size of the coreference cluster, and the leftover span uses the transitivity principle to find out the co-referent (Lee et al., 2018). The SpanBERT Model is available in allenNLP pre-trained package.

3.2 Finding Named Entities

Coreference resolution ensures that each coreference (i.e. pronoun) of an NE in a document is replaced by the actual NE. Various methods are available for tagging each document word with its proper named entities tag. In recent years, deep learning-based methods gained quite a lot of popularity due to their high accuracy over various document contexts. The strength of deep learning algorithms is that they are pre-trained and use transfer learning methodologies to achieve higher accuracy and low biasedness (Kitaev & Klein, 2018).

This current work uses the AllenNLP coreference resolution package for tagging NEs present in the documents. Internally AllenNLP uses ElMo, a sequence-to-sequence neural language model. Peters et al. (2018) have proposed ELMO representation with two layers of bi-directional language model.

The forward language model predicts the future token based on past available tokens. The backward language model runs over the sequence in reverse order, predicting the previous token given its future context. Given the token sequence $t_1, t_2, \ldots . t_N$ the forward and backward model compute the sequence of probability of token t_k.

This, somewhat complex representation of the document has capability to capture both syntactical and semantic characteristics. Also, ELMo is a pre-trained deep bidirectional Transformer that works by jointly conditioning on both left and right context in all layers. It provides a lot of strength to the model for finding NEs accurately.

3.3 Named Entities and Similarity

It is therefore clear from the foregoing discussion that NEs contribute to the determination of a document's uniqueness. While there has been substantial research on finding textual similarity (Friburger et al., 2002), the specific role played by NEs has received only scant attention of researchers.

3.3.1 NE Frequency Similarity

Basic Concept: The existence of unique NEs and their frequency is a characteristic contributing to the uniqueness of a document. If two documents share a list of named entities it increases the similarity of the documents. Adding frequency to the context means that the documents not only share the same NEs but the NEs also occur in each of them in the same proportion. This increases the possibility of the documents being similar to a great extent. It will be clear with the help of a small example:

Text 1: Mahatma Gandhi is the father of our nation. Gurudev Rabindranath Tagore had given him the title of Mahatama.

Text 2: At one point in time, Gurudev Rabindranath Tagore gave the Mahatma title to Mohandas Karamchand Gandhi, who came to be popularly known as Mahatma Gandhi after that.

It is clear from the above two pieces of text that they are very similar and carry almost the same information. In this example, Mahatma Gandhi, Gurudev Rabindranath Tagore, and Mohandas Karamchand Gandhi, that is two out of three NEs are common between the two pieces of text, with the exact frequency count.

This example shows that documents sharing the same NEs with equal frequency counts must be similar or would at least be discussing the same topic.

Table 1. NE Frequency

NE words	Text-1 Frequencies	Text-2 Frequencies
Mahatama	1	1
Mahatma Gandhi	1	1
Gurudev Rabindranath Tagore	1	1
Mohanddas Karamchand Gandhi	0	1
Gurudev	1	1

Euclidean distance suffers from high sensitivity to magnitudes (Xia et al., 2015). Due to this, information retrieval and related studies widely use cosine similarity (Rahutomo et al., 2015) to compare similarity of document vectors. Cosine similarity is the angle between two document vectors. It does not depend on magnitude.

The cosine similarity between the abovementioned texts is 0.895, which is understandable because they share many NEs with compare frequencies. This illustrative example clearly shows that above mentioned concept will play a significant role in understanding the similarity between documents.

3.3.2 NE Order Similarity

Basic Concept: the last subsection shows the importance of NE frequencies in the determining similarity. Wang et al. (2016) discussed ranking the NEs in the document based on their importance. However, that previous work is only confined to web documents and uses a tripartite graph for understanding the NE rank based on "importance". Want et al. (2016) did not leverage the order of occurrence of the NEs in the document.

This work focuses on the Concept of the order of occurrence of the NEs within the documents. It will capture the order information of each NEs in a document and compare it with their order of occurrence in another document. The argument being that if two documents share the same NEs in the same order of occurrence, the possibility of them being similar is very high. It signifies both documents present context and concepts in the same order, and even the unique building blocks of the documents (i.e. NEs) are similar. It will be clear with the help of a small example:

Text 1: Mahatma Gandhi is the father of our nation; Gurudev Rabindranath Tagore had given him the title of Mahatama.

Text 2: At one point in time, Gurudev Rabindranath Tagore gave the Mahatma title to Mohandas Karamchand Gandhi, who came to be popularly known as Mahatma Gandhi after that.

Table 2. NEs and their location in the documents

NE words	Location in Text-1	Location in Text-2
Mahatama	15	11
Mahatma Gandhi	1	24
Gurudev Rabindranath Tagore	8	6
Mohanddas Karamchand Gandhi	0	14
Gurudev	8	6

Table 3. NEs and their consecutive order in the documents

NE words	Order in Text-1	Order in Text-2
Mahatama	4	3
Mahatma Gandhi	1	5
Gurudev Rabindranath Tagore	3	2
Mohanddas Karamchand Gandhi	0	4
Gurudev	2	1

Once again, for calculation of NE order similarity between two documents, cosine similarity is useful. The cosine similarity between the abovementioned examples is 0.714, which is understandable because they share most of the NEs in nearly the same order. This illustrative example clearly shows that the above mentioned concept will play a significant role in understanding the similarity between the two documents.

4. EXPERIMENTS & RESULTS

Dataset: For experimentation purposes, this work uses a set of four documents. Two are on Mother Teresa (Mohita, 2014; Panda, 2022), and the other two are news reports on the Djokovic vaccine controversy (Kershaw, 2022; Rajan, 2022). In order to avoid bias, all four document sizes are in the range of 1000-1200 words.

Experimentation: In this work we have compared NEs between similar documents, i.e. between the two Djokovic news reports and between the two Mother Teresa documents, and dissimilar documents: one of the Djokovic reports with one of the Mother Teresa documents.

Obtained Results: The results obtained on both similar and dissimilar sets reflect the cosine similarity of NE frequencies and NE order. The results shown in the tables below justify our hypothesis. The assumption was that similar documents would have a high value of NE frequency and order similarity, and dissimilar documents would have a low value of NE frequency and order similarity, which is what is observed from the entries in Table 4. Table 5 shows the corresponding cosine similarity between the same pairs of documents when the conventional measures Term Frequency (TF) and Term Frequency – Inverse Document Frequency (TF-IDF) are used. While, the cosine similarity values between the "similar" pairs are higher than those between dissimilar pairs, the differences between the similar and dissimilar pairs are significantly lower. Also, notably TF-IDF performs worse than even TF based similarity values. This is because TF-IDF, having been developed to discriminate between documents, loses all the common words, including the common NEs between them, (the IDF part evaluates to zero) which makes TF-IDF weak in measuring similarity.

Table 4. NE Frequency and Order Cosine Similarity

	Similar Documents		**Dissimilar Documents**	
	Mother Teresa 1 Vs Mother Teresa 2	**Djokovic (BBC) Vs Djokovic (iNEWS)**	**Mother Teresa 1 Vs Djokovic (BBC)**	**Mother Teresa 2 Vs Djokovic (iNEWS)**
Frequency Similarity	0.964	0.675	0	0.0003
Order Similarity	0.445	0.255	0	0.004

Table 5. TF and TF-IDR similarity

	Similar Documents		**Dissimilar Documents**	
	Mother Teresa 1 Vs Mother Teresa 2	**Djokovic (BBC) Vs Djokovic (iNEWS)**	**Mother Teresa 1 Vs Djokovic (BBC)**	**Mother Teresa 2 Vs Djokovic (iNEWS)**
TF Similarity	0.413023	0.382043	0	0.005
TF-IDF Similarity	0.27693711	0.26206084	0.03841347	0.08569649

5. DISCUSSION

NEs have the potential to uncover latent semantic characteristics present in the documents. The frequency and order meta characteristics of NEs can enormously impact information retrieval, textual similarity, text summarization, voice command identification, and many other application areas. Semantic similarity of texts using named entities is an under-explored area; NE frequency and order information help in the semantic understanding of the text. Available plagiarism detection software is based only on syntactical or lexical comparisons. A proper understanding of NEs and their meta-characteristics would help improve the effectiveness of plagiarism detection software.

This current work demonstrates the advantages of NEs over conventional TF or TF-IDF based similarity measures in understanding text. In this research experiments were conducted with the help of a limited set of experiments. A more extensive analysis is required over a large set of documents to analyze the proposed method's full potential and limitations. Natural language processing is an evolving area with interesting developments due to the advancement of deep learning techniques among others. Some recently proposed techniques for co-reference resolution and extraction of NEs have the potential to produce more accurate results.

6. CONCLUSION

The work presented in this article has demonstrated the potential of meta characteristics of NEs in understanding semantic similarity between documents. The work in its current form is restricted by the effort required due to manual calculations. The entire algorithm will be automated in the near future to run more comprehensive tests over a large set of documents of various sizes and types to establish the strength and robustness of our proposed method.

REFERENCES

Aone, C., and William, S., (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In 33rd Annual Meeting of the Association for Computational Linguistics, pp. 122-129.

Clark, K., and Manning, C. D., (2016). Improving coreference resolution by learning entity-level distributed representations. arXiv preprint arXiv:1606.01323.

Friburger, N. et al, (2002). Textual similarity based on proper names. In Proc. of the workshop Mathematical/Formal Methods in Information Retrieval, pp. 155-167.

Grishman, R., and Sundheim, B. M., (1996). Message understanding conference-6: A brief history. In COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics.

Kershaw, T., (2022), January 20. Novak Djokovic's controversial beliefs and why he is opposed to the vaccine. The Independent. https://www.independent.co.uk/sport/tennis/novak-djokovic-vaccine-australian-open-2022-b1995236.html

Lee, K. et al, (2017). End-to-end neural coreference resolution. arXiv preprint arXiv:1707.07045.

Lee, K. et al, (2018). Higher-order coreference resolution with coarse-to-fine inference. arXiv preprint arXiv:1804.05392.

Li, J. et al, (2020). A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering, Vol. 34, No. 1, pp 50-70.

McCarthy, J. F., and Lehnert, W. G., (1995). Using decision trees for coreference resolution. arXiv preprint cmp-lg/9505043.

Mitkov, R., (2001). Outstanding issues in anaphora resolution. In International Conference on Intelligent Text Processing and Computational Linguistics Springer, Berlin, Heidelberg, pp. 110–125.

Mohita, N., (2014), February 27. Mother Teresa: Essay on Mother Teresa. Your Article Library. https://www.yourarticlelibrary.com/essay/mother-teresa-essay-on-mother-teresa/28438

Panda, I., (2022), June 2. The Life and Work of Mother Teresa - 1163 Words | Essay Example. Free Essays. https://ivypanda.com/essays/the-life-and-work-of-mother-teresa/

Pradhan, N. et al, (2015). A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications, Vol. 120, No. 9, pp. 29-34.

Rahutomo, F. et al, (2012). Semantic cosine similarity. In The 7th international student conference on advanced science and technology ICAST, Vol. 4, No. 1, p. 1.

Rajan, B.A., (2022), February 15. Novak Djokovic willing to miss tournaments over vaccine. BBC News. https://www.bbc.com/news/world-60354068

Wang, C. et al, (2016). Nerank: Bringing order to named entities from texts. In Asia-Pacific Web Conference, Springer, Cham, pp. 15-27.

Xia, P. et al, (2015). Learning similarity with cosine similarity ensemble. Information Sciences, Vol. 307, pp. 39-52.