Turkish Journal of Computer and Mathematics Education *Vol.11 No.03 (2020), 1607 - 1611*

Research Article

Similarity Metrics for Aspect-based Text Classification

Triveni Lal Pal [a], Kamlesh Dutta [b]

[a,b] Department of Computer Science and Engineering, National Institute of Technology Hamirpur, India.

Abstract: Cosine similarity compares two units of text to get the semantic relation between them. This comparison is based on the numerical value (features) represented by semantic vectors. Orthogonality between the feature vectors makes them inefficient for semantic comparisons. Modifying the metrics to handle orthogonality perform better taking the advantage of representations. This article, proposed modified cosine similarity metrics for comparing sentences based on multi-feature embedding vectors. Our approach relies on the assumption that linguistic units may have multiple aspects of semantics which should be considered while calculating the similarity between the two units.

Keywords: Cosine similarity, semantic measures, similarity metrics, vector space model, word embeddings

1.Introduction

A huge amount of text data are generated daily which requires to be managed automatically. This needs understanding the semantics of the text data which, in turn, require machine learning tools to have deep understanding of the features of the text. Text classification is one of the most challenging NLP and text mining task that entice researchers to come with new solutions for real applications. The objective of text classification is to place new text/documents into the appropriate class. These classes are created with knowledge of the document structure or with knowledge of expected topics. Traditionally, text classification has been performed by a classifier using labeled data. Research in fully automatic categorization using a machine is lagging far behind. So, representation of text which also conveys document structure contributing towards semantic information is highly desirable and challenging task in the field of text classification/categorization.

Traditionally, text categorization has been performed by a classifier using labeled data. However, people can categorize documents into named categories without any explicit training because we know the meaning of category names. Research in fully automatic categorization using a machine is lagging far behind. So, representation of text which also conveys semantic information is highly desirable and challenging task in the field of text categorization. Text classification can be performed by comparing two text units using semantic measures (SMs) based on their semantics. These measures compare the abstract representations (their numerical values) for semantic proxies. The semantic proxies ultimately extracted as semantic evidence for comparison. These evidences are expected characterize the semantics, directly/indirectly, of the compared linguistic units (Pal et al., 2018).

For text categorization, we saw that the objective is to place new documents into the appropriate folders. These folders were created by someone with knowledge of the document structure, someone who knew the expected topics. The clustering process is equivalent to assigning the labels needed for text categorization. Because there are many ways to cluster documents, it is not quite as powerful a process as assigning answers (i.e., known correct labels) to documents. Still, clustering can be insightful. By studying key words that characterize a cluster, a company could learn about its customers.

Even though traditional approaches for semantic similarity calculations, such as human evaluation and other cognitive science approaches are better than automatic text similarity metrics, they are labor intensive (Pal & Kumari, 2017). Therefore to minimize the human intervention one needs to look at an alternative of automatic text processing driven by statistical data.

SMs have applications in a broad range of text mining and NLP applications, like text summarization (Schubert, 2015) (Yan et al., 2015), machine translation (Pal et al. 2012) (Rinaldi, 2008) document classification (Zhou et al., 2016) (Niraula et al., 2015) (Huang et al., 2016) (Navigli et al., 2011)(Chen et al., 2014) and retrieval (Schubert, 2015) (Yan et al., 2015) (Rinaldi, 2008), information extraction (Soderland, 1999), question-answering (Zhou et al., 2016), semantic similarity applications (Niraula et al., 2015) (Huang et al., 2016), word sense disambiguation (Navigli et al., 2011), (Chen et al., 2014) web search (Shen et al., 2014) and social media mining

(Ahsan et al., 2018) (Singh et al., 2017). They enable to take advantage of the knowledge encompassed in unstructured/semi-structured texts corpora and KRs to compare things. They are therefore essential tools for the design of numerous algorithms and treatments in which semantics matters.

In this paper, we proposed a semantic metrics based on multi-aspect sematic representations for comparing two linguistic units (here, sentences). The approach is a modification of cosine similarity which makes it more suitable for multi- aspect and multi-dimensional text. Next section elaborates the significance of the study. Section 3 is more imperative and contains the research done in the field by other researchers followed by section 4 in which we presented the proposed model. Section 5 contains results and discussion. Last section (Section 6) concludes, discussing feature of proposed model comparing it with other similarity metrics and forecasting the future research. The contributions of the study include:

1. Semantic metrics for multi-aspect embeddings.
2. Induce weight corresponding to every aspect.
3. The study shows that the proposed scheme performs better than baseline cosine similarity applied on VSM for text classification task.

2.Significance of the Study

Similarity calculations are highly dependent on representations. Cosine similarity calculates vector similarity between two vectors. Thus, two similar vectors may have maximum similarity index (Pal et al. 2018), however it might be possible that these vectors are a result of erroneous calculations. The basic model (VSM) for representation is based on hypothesis that all words are orthogonal in Euclidian space. Thus we have taken multi-aspect semantic representations (our previous work) for similarity calculations. Multi-aspect representation is inspired by bit plane slicing in image processing where instead of highlighting gray level images, one highlights the contribution made by specific bits to the total image appearance. Multi-aspect representation is just reverse of bit-plane slicing. The representations incorporating different aspects are combined to get more general compact high dimensional representations for linguistic unis.

3.Related Work

Goodness of semantic metrics significantly depends on the goodness of representation system. Therefore, in this section, we will discuss semantic metric giving introduction to few most important representation approaches. In conventional approaches for semantic representations like Vector Space Model (VSM) (Salton & McGill, 1986) (G. Salton, 1975), text has been represented as a bag of words (BOW). VSM, in its most basic form, use Boolean entries for each element in the vector to indicate presence or absence of the word in the document. Further, term-frequency (*tf*), term frequency-inverse document frequency (*tf-idf*), point-wise positive mutual information (*PPMI*), etc. were used as weighting factors to capture the notion that all words cannot be equally important in the document. The vector space model considers numerical feature vectors in a Euclidean space. Each word in VSM was treated as independent from other, thus losing the semantic relation between the words. BoW ignores word-order, thus missing important semantic relations between the words. Indeed, researchers, in the text mining community, proposed ingenious solutions to incorporate the semantic relations (word-order) in the vector space model. N-gram statistical language model and Language Modeling for Information Retrieval (Croft, B., & Lafferty, J., 2003), are one of such attempts. N-gram model intended to incorporate semantics by using context word in predicting the target word. The target word is predicted using conditional probability $P(w_n|w_{1:n-1})$. Where, wn is the target word and words $w_{1:n-1} = w_1, \dots, w_{n-1}$ are called the context. The similarity between two words is calculated based on similarity between their vectors.

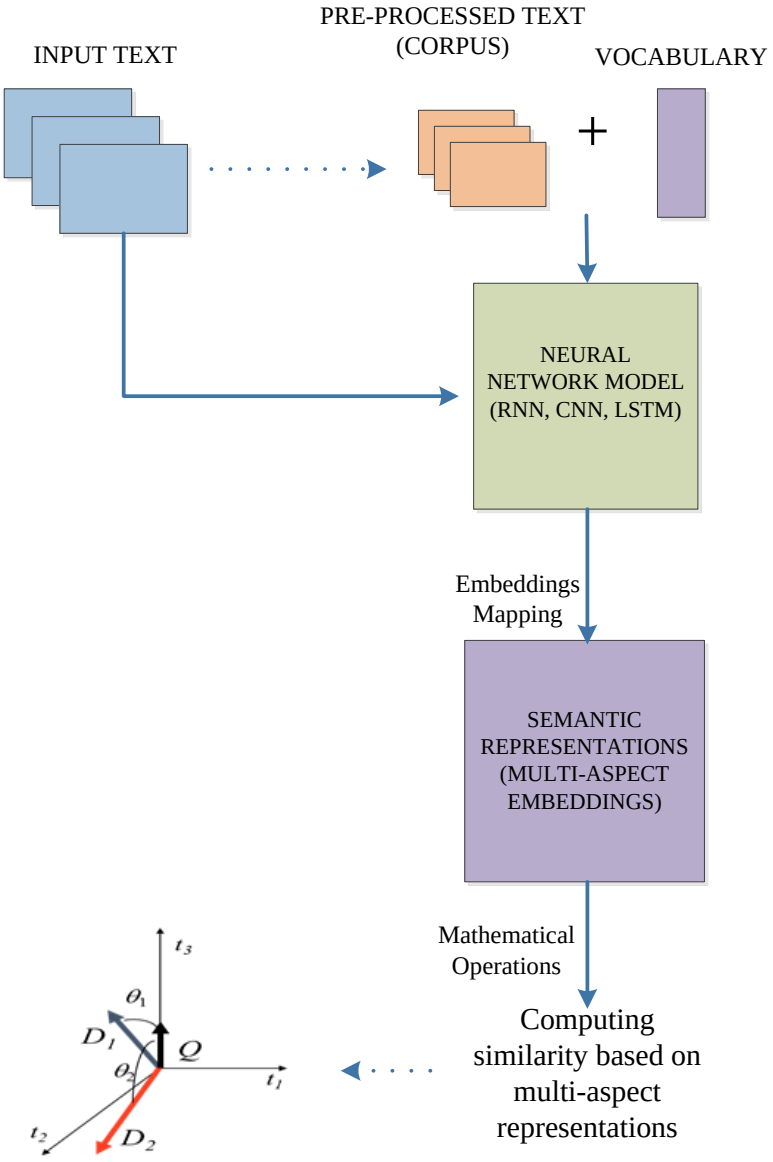
Another approach that has gain most attention of all semantic space models and known for its ability to incorporate hidden semantic relations between words/documents is Latent Semantic Analysis (LSA) (Deerwester et al., 1990). Though LSA too uses BoW approach, it proved to be better than basic VSM because of its unique dimensionality reduction algorithm. LSA first forms a term-document matrix using a document collection and then finds its low-rank approximation using singular value decomposition (SVD). LSA has the capability of finding out hidden semantic relations (that's why the name 'latent') even if the two words never co-occurred in the document. The basic idea behind LSA's meaning induction of a word is the aggregate contexts (in which a word does or does not occur), that produce a set of constraints which generates the meaning of the word (Landauer et al., 1998). Firth (Firth, 1957) has put this idea as, "you shall know a word by the company it keeps."

In recent years, low dimensional, dense vectors called "Word embeddings" based on neural networks learning, gaining attention for semantic learning. These methods are quite successful in learning the semantic representation of words. The skip-gram model and continuous bag-of-word model (CBOW) (Mikolov, Corrado, et al., 2013), (Mikolov, Chen, et al., 2013) are popular machine learning approaches for learning word representations that can be efficiently trained on large amounts of text data (Mikolov, Le, et al., 2013). CBOW has the training objective

to combine the representations of surrounding (context) words to predict the word in the middle (target word). Whereas, the skip-gram model trained to predict the source context words based on the target word. Recently, (Xu et al., 2016) uses word embeddings in another approach that uses spatial distance to show word relatedness known as "Semantic word cloud". This approach better visualizes the semantic relatedness between the words by improving the aesthetic on word layout.

4.Semantic Metrics for Multi-Aspect Embeddings

In this section, we proposed a semantic metrics for multi-aspect representations (Fig. 1) (for calculating semantic similarity and related both between two linguistic units). Our model is based on the assumption that multiple embeddings are required to completely represent a linguistic unit (with its all semantic features intact) as every linguistic units have multiple aspects (dimensions) that contribute significantly to overall meaning of the unit. Therefore, cosine similarity in its original form cannot be applied to multiple embeddings corresponding to a linguistic unit. Here, in our proposed metrics, we have taken embeddings of word-sequences (ξ_s), word co-occurrences (ξ_c) and word-hierarchies (ξ_h). Then, semantic similarity between two units is calculated based on similarity between these embeddings applying cosine similarity with some modifications. Task specific induced weights are learned to get overall similarity.



$$sim(D_j, Q_q) = \frac{W_s * sim(\xi_{D_{j,s}}, \xi_{Q_s}) + W_c * sim(\xi_{D_{j,c}}, \xi_{Q_c}) + W_h * sim(\xi_{D_{j,h}}, \xi_{Q_h})}{W_s + W_c + W_h}$$

Figure 1. Depiction of proposed scheme for similarity measure

$$sim(D_j, Q_q) = \frac{W_s * sim(\xi_{D_{j,s}}, \xi_{Q_s}) + W_c * sim(\xi_{D_{j,c}}, \xi_{Q_c}) + W_h * sim(\xi_{D_{j,h}}, \xi_{Q_h})}{W_s + W_c + W_h} \quad (1)$$

where,

$$sim(\xi_{D_{j,s}}, \xi_{Q_s}) = \frac{\sum_{i=1}^{n} \xi_{i,j,s} \xi_{i,Q,s}}{\sqrt{\sum_{i=1}^{n} \xi_{i,j,s}^2} \sqrt{\sum_{i=1}^{n} \xi_{i,Q,s}^2}} \quad (2)$$

$\xi_{D_{j,s}}$, ξ_{Q_s} are sequence embeddings; $\xi_{D_{j,c}}$, ξ_{Q_c} are co-occurrence embeddings and $\xi_{D_{j,h}}$, ξ_{Q_h} are embeddings with hierarchical information of document D_j and query Q respectively. W_s, W_c, and W_h are weights learned corresponding to three different aspects.

5.Results and Discussion

The proposed metrics is tested on classification and clustering tasks and clearly suggest that proposed scheme perform better in terms of accuracy of the both sentence classification and clustering applications.

The proposed metrics is a generalization of cosine similarity. Varying the values W_s, W_c and W_h one can adopt the metrics for their task. For example if an application require sequence information predominantly then learn the weights to adjust in such a way that vector contribution is dominated by sequence embeddings. Similarly in an application like hyponymy-hyperonymy or synonymy-antonymy pairs detection, W_c can be set close to 1. Proposed model is generalization of cosine similarity. As a specific case setting $W_s = W_c = W_h = 1$, we get cosine similarity.

6.Conclusion

Cosine similarity is easy to understand and conceptualize. Therefore, most often it is opted over other; furthermore it is mathematically easy because of vector algebra. Cosine similarity in its basic form can be easily implemented for vectors obtained in vector space model. However these vectors are incapable of incorporating semantic features as VSM vectors are orthogonal, which consider only frequency information and ignore word-order and contextual information. Proposed metrics when applied to multiple vectors (multi-aspect embeddings), gives better results with other state-of-the-art metrics on text classification task. Moreover, it is generalization of cosine similarity metrics which can be applied to any task by learning the weights (corresponding to sequence, co-occurrence and hierarchical embeddings) specific to the task. As a future research direction, this metrics is yet to be seen how it works on other text mining tasks.

References

Ahsan, M., Kumari, M., Singh, T., & Pal, T. L. (2018). Sentiment based information diffusion in online social networks. International Journal of Knowledge Discovery in Bioinformatics (IJKDB), 8(1), 60-74.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, 161175.

Chen, X., Liu, Z., & Sun, M. (2014, October). A unified model for word sense representation and disambiguation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1025-1035.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American society for information science, 41(6), 391-407.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. In Studies in linguistic analysis, 1–32.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Huang, Y., Xiong, D., Shi, X., Chen, Y., Wu, C., & Huang, G. (2016). Adapted competitive learning on continuous semantic space for word sense induction. Neurocomputing, 171, 1475-1485.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Croft, B., & Lafferty, J. (Eds.). (2003). Language modeling for information retrieval. Springer Science & Business Media, 13.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, 3111-3119.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

Navigli, R., Faralli, S., Soroa, A., de Lacalle, O., & Agirre, E. (2011). Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In Proceedings of the 20th ACM international conference on Information and knowledge management, 2317-2320.

Niraula, N. B., Gautam, D., Banjade, R., Maharjan, N., & Rus, V. (2015). Combining word representations for measuring word relatedness and similarity. In The twenty-eighth international flairs conference.

Pal, T. L., Dutta, K., & Singh, P. (2012). Anaphora resolution in Hindi: Issues and challenges. International Journal of Computer Applications, 42(18), 7-13.

Pal, T. L. & Kumari, M. (2017). Semantic Similarity Metrics for Analysing Semantic Representations. In Proc. of International Conference on Emerging Trends in Engineering Innovations and Technology Management, ISBN- 978-93-86724-30-4, 30-35.

Pal, T. L., Kumari, M., Singh, T., & Ahsan, M. (2018). Semantic Representations in Text Data. International Journal of Grid and Distributed Computing, 11(9), 65-80.

Rinaldi, A. M. (2008). A content-based approach for document representation and retrieval. In Proceedings of the eighth ACM symposium on Document engineering, 106-109.

Salton, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill, Inc. New York, NY, USA.

Schubert, L. (2015). Computational Linguistics. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. In Proceedings of the 23rd international conference on world wide web, 373-374.

Singh, T., Kumari, M., Pal, T. L., & Chauhan, A. (2017). Current trends in text mining for social media. International Journal of Grid and Distributed Computing, 10(6), 11-28.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. Machine learning, 34(1), 233-272.

Xu, J., Tao, Y., & Lin, H. (2016, April). Semantic word cloud generation based on word embeddings. In 2016 IEEE Pacific Visualization Symposium (PacificVis), 239-243.

Yan, Y., Yin, X. C., Li, S., Yang, M., & Hao, H. W. (2015). Learning document semantic representation with hybrid deep belief network. Computational intelligence and neuroscience.

Zhou, G., Zhou, Y., He, T., & Wu, W. (2016). Learning semantic representation with neural networks for community question answering retrieval. Knowledge-Based Systems, 93, 75-83.