Learning Thematic Similarity Metric Using Triplet Networks

Liat Ein Dor*, Yosi Mass*, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov and Noam Slonim
IBM Research, Haifa, Israel
{liate,yosimass,alonhal,eladv,ilyashn,ranita,noams}@il.ibm.com

Abstract

In this paper we suggest to leverage the partition of articles into sections, in order to learn thematic similarity metric between sentences. We assume that a sentence is thematically closer to sentences within its section than to sentences from other sections. Based on this assumption, we use Wikipedia articles to automatically create a large dataset of weakly labeled sentence triplets, composed of a pivot sentence, one sentence from the same section and one from another section. We train a triplet network to embed sentences from the same section closer. To test the performance of the learned embeddings, we create and release a sentence clustering benchmark. We show that the triplet network learns useful thematic metrics, that significantly outperform state-of-the-art semantic similarity methods and multipurpose embeddings on the task of thematic clustering of sentences. We also show that the learned embeddings perform well on the task of sentence *semantic* similarity prediction.

1 Introduction

Text clustering is a widely studied NLP problem, with numerous applications including collaborative filtering, document organization and indexing (Aggarwal and Zhai, 2012). Clustering can be applied to texts at different levels, from single words to full documents, and can vary with respect to the clustering goal. In this paper, we focus on the problem of clustering sentences based on thematic similarity, aiming to group together sentences that discuss the same theme, as opposed to the related task of clustering sentences that represent paraphrases of the same core statement.

Thematic clustering is important for various use cases. For example, in multi-document summarization, one often extracts sentences from multiple documents that have to be organized into meaningful sections and paragraphs. Similarly, within the emerging field of computational argumentation (Lippi and Torroni, 2016), arguments may be found in a widespread set of articles (Levy et al., 2017), which further require thematic organization to generate a compelling argumentative narrative.

We approach the problem of thematic clustering by developing a dedicated sentence similarity measure, targeted at a comparative task – Thematic Distance Comparison (TDC): given a pivot sentence, and two other sentences, the task is to determine which of the two sentences is thematically closer to the pivot. By training a deep neural network (DNN) to perform TDC, we are able to learn a thematic similarity measure.

Obtaining annotated data for training the DNN is quite demanding. Hence, we exploit the natural structure of text articles to obtain weakly-labeled data. Specifically, our underlying assumption is that sentences belonging to the same section are typically more thematically related than sentences appearing in different sections. Armed with this observation, we use the partition of Wikipedia articles into sections to automatically generate sentence triplets, where two of the sentences are from the same section, and one is from a different section. This results in a sizable training set of weakly labeled triplets, used to train a triplet neural network (Hoffer and Ailon, 2015), aiming to predict which sentence is from the same section as the pivot in each triplet. Table 1 shows an example of a triplet.

To test the performance of our network on the-

* These authors contributed equally to this work.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 1–6
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

matic clustering of sentences, we create a new clustering benchmark based on Wikipedia sections. We show that our methods, combined with existing clustering algorithms, outperform state-of-the-art general-purpose sentence embedding models in the task of reconstructing the original section structure. Moreover, the embeddings obtained from the triplet DNN perform well also on standard semantic relatedness tasks. The main contribution of this work is therefore in proposing a new approach for learning thematic relatedness between sentences, formulating the related TDC task and creating a thematic clustering benchmark. To further enhance research in these directions, we publish the clustering benchmark on the IBM Debater Datasets webpage [1].

2 Related Work

Deep learning via triplet networks was first introduced in (Hoffer and Ailon, 2015), and has since become a popular technique in metric learning(Zieba and Wang, 2017; Yao et al., 2016; Zhuang et al., 2016). However, previous usages of triplet networks were based on supervised data and were applied mainly to computer vision applications such as face verification. Here, for the first time, this architecture is used with *weakly-supervised* data for solving an *NLP* related task. In (Mueller and Thyagarajan, 2016), a supervised approach was used to learn *semantic* sentence similarity by a Siamese network, that operates on pairs of sentences. In contrast, here the triplet network is trained with weak supervision, aiming to learn *thematic* relations. By learning from triplets, rather than pairs, we provide the DNN with a context, that is crucial for the notion of similarity. (Hoffer and Ailon, 2015) show that triplet networks perform better in metric learning than Siamese networks, probably due to this valuable context. Finally, (Palangi et al., 2016) used click-through data to learn sentence similarity on top of web search engine results. Here we propose a different type of weak supervision, targeted at learning *thematic* relatedness between sentences.

3 Data Construction

We present two weakly-supervised triplet datasets. The first is based on sentences appearing in same vs. different sections, and the second is based on section titles. The datasets are extracted from the Wikipedia version of May 2017.

3.1 Sentence Triplets

For generating the sentence triplet dataset, we exploit the Wikipedia partitioning into sections and paragraphs, using OpenNLP[2] for sentence extraction. We then apply the following rules and filters, in order to reduce noise and to create a high-quality dataset, ‘triplets-sen’: i) The maximal distance between the intra-section sentences is limited to three paragraphs. ii) Sentences with less than 5, or more than 50 tokens are filtered out. iii) The first and the ”Background” sections are removed due to their general nature. iv) The following sections are removed: ”External links”, ”Further reading”, ”References”, ”See also”, ”Notes”, ”Citations” and ”Authored books”. These sections usually list a set of items rather than discuss a specific subtopic of the article’s title. v) Only articles with at least five remaining sections are considered, to ensure focusing on articles with rich enough content. An example of a triplet is shown in Table 1.

1. McDonnell resigned from Martin in 1938 and founded McDonnell Aircraft Corporation in 1939
2. In 1967, McDonnell Aircraft merged with the Douglas Aircraft Company to create McDonnell Douglas
3. Born in Denver, Colorado, McDonnell was raised in Little Rock, Arkansas, and graduated from Little Rock High School in 1917

Table 1: Example of a section-sen triplet from the article ‘James Smith McDonnell’. The first two sentences are from the section ’Career’ and the third is from ‘Early life’

In use-cases such as multi-document summarization(Goldstein et al., 2000), one often needs to organize sentences originating from different documents. Such sentences tend to be stand-alone sentences, that do not contain the syntactic cues that often exist between adjacent sentences (e.g. co-references, discourse markers etc.). Correspondingly, to focus our weakly labeled data on sentences that are typically stand-alone in nature, we consider only paragraph opening sentences.

An essential part of learning using triplets, is the mining of difficult examples, that prevent quick stagnation of the network (Hermans et al., 2017). Since sentences in the same article essentially discuss the same topic, a deep understanding of se-

[1] http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

[2] https://opennlp.apache.org/

mantic nuances is necessary for the network to correctly classify the triplets. In an attempt to obtain even more challenging triplets, the third sentence is selected from an adjacent section. Thus, for a pair of intra-section sentences, we create a maximum of two triplets, where the third sentence is randomly selected from the previous/next section (if exists). The selection of the third sentence from both previous and next sections is intended to ensure the network will not pick up a signal related to the order of the sentences. In Section 5 we compare our third-sentence-selection method to two alternatives, and examine the effect of the selection method on the model performance.

Out of the $5.37M$ Wikipedia articles, $809K$ yield at least one triplet. We divide these articles into three sets, training (80%), validation and test (10% each). In terms of number of triplets, the training set is composed of $1.78M$ triplets, whereas the validation and test are composed of $220K$ and $223K$ triplets respectively.

3.2 Triplets with Section Titles

Incorporating the section titles into the training data can potentially enhance the network performance. Correspondingly, we created another triplets data, 'triplets-titles', where in each triplet the first sentence in the section (the 'pivot') is paired with the section title[3], as well as with the title of the previous/next sections (if exists), where the former pair is assumed to have greater thematic similarity. After applying the filters described above we end up with $1.38M$, $172K$ and $173K$ triplets for the training, validation and test set respectively. An example of a triplet is shown in Table 2.

Note, that for this variation of the triplets data, the network is expected to find a sentence embedding which is closer to the embedding of the true section title, than to the embedding of the title of the previous/next section. The learned representation is expected to encode information about the themes of the different sections to which the sentence can potentially belong. Thus, thematically related sentences are expected to have similar representations.

1. Bishop was appointed Minister for Ageing in 2003.
2. Julie Bishop Political career
3. Julie Bishop Early life and career

Table 2: Example of a triplet from the triplet-titles dataset, generated from the article 'Julie Bishop'.

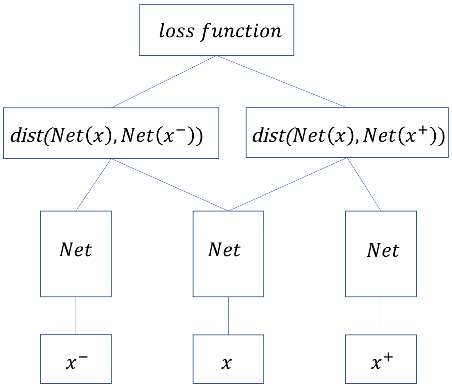

Figure 1: Triplet Network

3.3 Sentence Clustering Benchmark (SCB)

Our main goal is to successfully partition sentences into subtopics. Unfortunately, there is still no standard evaluation method for sentence clustering, which is considered a very difficult task for humans (Geiss, 2009). Correspondingly, we leverage again the partition of Wikipedia articles into sections. We assume that this partition, as performed by the Wikipedia editors, can serve as ground truth for the clustering of the article sentences. Based on this assumption we create a sentence clustering benchmark (SCB). SCB includes 692 articles that were not used in the training and validation sets of 'triplet-sen' and 'triplet-titles'. The number of sections (and correspondingly clusters) per article ranges from 5 to 12. The number of clustered sentences ranges from 17 to 1614, with an average of 67 sentences per article.

4 Model Architecture

We adopt the triplet network architecture (Hoffer and Ailon, 2015) (Figure 1) for obtaining sentence embeddings via metric learning as follows.

Assume a training data of sentences, arranged into triplets (x,x^+,x^-), where the pair (x,x^+) is presumably more similar than the pair (x,x^-). To train the model, each of the three sentences of each triplet, is fed into the same network (`Net`), as a sequence of word embeddings. The layer outputs their representations `Net`(x), `Net`(x^+) and

[3] We define the section title to be the article title concatenated to the section title. For example, the title of the section "Pricing" in the article "Black Market" is "Black Market Pricing".

Net(x$^-$) respectively. Our objective is to make the representations of x and x$^+$ closer than the representations of x and x$^-$. Thus the next layer uses a distance function, denoted by 'dist', to compute two distances

$$d^+ = \text{dist}(\texttt{Net}(x), \texttt{Net}(x^+))$$
$$d^- = \text{dist}(\texttt{Net}(x), \texttt{Net}(x^-))$$

The final layer applies softmax on (d^+,d^-) that results in p(d^+) and p(d^-). Finally, the loss function is given by:

$$\text{loss} = |p(d^+)| + |1 - (p(d^-)|$$

Net is composed of a Bi-directional LSTM with hidden size 300 and 0.8 dropout followed by an attention (Yang et al., 2016) layer of size 200. The input to Net are the pre-trained glove word embeddings of 300d trained on 840B tokens (Pennington et al., 2014). For *dist* and the loss function we use the $L1$ distance, which we found to yield better results than $L2$ and cosine-similarity. The selected loss function outperformed the popular triplet loss suggested in (Schroff et al., 2015). Finally, we use Adam optimizer with initial learning rate of 0.001. Given a sentence s, Net(s) provides a sentence embedding of dimension 600.

5 Experiments

5.1 Reconstructing Article Sections

As mentioned, our main objective task is clustering sentences into subtopics. As a preliminary step, we first evaluate our method on the triplet-sen test set. We compare the model trained on triplet-sen to two well known methods. The first, mean-vectors, is simply the mean of the GloVe embeddings of the sentence words (Tai et al., 2015), which is considered a strong unsupervised baseline. The second, skip-thoughts (Ryan Kiros, 2015), is among the state-of-the-art unsupervised models for semantic similarity, and the most popular multi-purpose embedding method. We address two versions of skip-thoughts: one is based on the original 4800-dimensional vectors (skip-thoughts-cs), and the other, skip-thoughts-SICK, is based on the similarity function learned from the SICK semantic similarity dataset, as described in (Ryan Kiros, 2015). The aim of assessing skip-thoughts-SICK is to examine how well a state-of-the-art semantic similarity function performs on the thematic clustering task. In the case of mean-vectors and skip-thoughts-CS, the similarity between the sentences is computed using the cosine similarity (CS) between the embedding vectors.

Table 3 indicates that our method, denoted by triplet-sen, clearly outperforms the other tested methods. Surprizingly, skip-thoughts-SICK is inferior to skip-thoughts-CS. Note that an additional interesting comparison is to a skip-thought version obtained by learning a linear transformation of the original vectors using the triplet datasets. However, no off-the-shelf algorithm is available for learning such transformation, and we leave this experiment for future work.

Method	accuracy
mean-vectors	0.65
skip-thoughts-CS	0.615
skip-thoughts-SICK	0.547
triplets-sen	**0.74**

Table 3: Results on the triplets data

Next we report results on the clustering benchmark, SCB (Section 3.3). We evaluate three triplet-based models. Triplets-sen and triplets-titles are the models trained on triplets-sen and triplets-titles datasets respectively. Triplets-sen-titles is a concatenation of the representations of our two models. In addition we compare to mean-vectors and skip-thoughts-CS.

The evaluation procedure is performed as follows: for each method, we first compute for the sentences of each article, a similarity matrix, by calculating the CS between the embedding vectors of all pairs of sentences. We then use Iclust (Yom-Tov and Slonim, 2009; Slonim et al., 2005) and k-means to cluster the sentences, where the number of clusters is set to the number of sections in SCB[4]. Since the clustering algorithms themselves are not the focus of this study, we choose the classical, simple k-means, and one more advanced algorithm, Iclust. For the same reason, we also set the number of clusters to the correct number. Finally, we use standard agreement measures, MI, Adjusted MI (AMI) (Vinh et al., 2009), Rand Index (RI) and Adjusted Rand Index (ARI) (Rand, 1971), to quantify the agreement between the ground truth and the clustering results.

As exhibited in Table 4, our models significantly outperform the two other methods for both clustering algorithms, where the best performance is achieved by the concatenated representations (triplets-sen-titles), suggesting the two models,

[4] For k-means, using L1 as the distance metric gave similar results

triplets-sen and triplets-titles, learned complementary features. The performance of skip-thoughts-SICK on this task (not shown) was again inferior to skip-thoughts-CS.

As mentioned in Section 3.1, the third sentence in triplet-sen was selected from the sections adjacent to the pivot section, aiming to obtain more difficult triplets. We use the clustering task to examine the effect of the selection method on the model performance. We compare to two alternative methods: one that chooses the third sentence from a random section within the *same* article, and another (triplets-sen-rand-art), that chooses it randomly from a random *different* article. Results show that the first method leads to the same performance as our method, whereas triplets-sen-rand-art yields inferior results (see Table 4). A possible explanation is that the within-article triplets are difficult enough to prevent stagnation of the learning process without the need for further hardening of the task. However, the cross-article triplets are too easy to classify, and do not provide the network with the challenge and difficulty required for obtaining high quality representations.

	iclust			
Method	MI	AMI	RI	ARI
mean-vectors	0.811	0.222	0.774	0.154
skip-thoughts-CS	0.656	0.125	0.747	0.087
triplets-sen-rand-art	0.885	0.266	0.787	0.192
triplets-sen	0.935	0.296	0.801	0.224
triplets-titles	0.904	0.273	0.799	0.206
triplets-sen-titles	**0.945**	**0.303**	**0.803**	**0.230**
	kmeans			
mean-vectors	0.706	0.153	0.7760	0.103
skip-thoughts-CS	0.624	0.099	0.745	0.067
triplets-sen-rand-art	0.793	0.205	0.775	0.145
triplets-sen	0.873	0.257	0.791	0.195
triplets-titles	0.836	0.231	0.786	0.172
triplets-sen-titles	**0.873**	**0.258**	**0.791**	**0.194**

Table 4: Results on the clustering task

5.2 Semantic Relatedness

As evident from the clustering results, our models learned well to capture *thematic* similarity between sentences. Here we investigate the performance of our model in the more classical task of semantic relatedness of sentences. Specifically, we examine the SemEval 2014 Task 1: semantic relatedness SICK dataset (Marelli et al., 2014). We adopt the experimental setup of (Ryan Kiros, 2015) and learn logistic regression classifiers on top of the absolute difference and the component-wise product for all sentence pairs in the training data. The evaluation measures are Pearson r, Spearman ρ, and mean square error (MSE). Table 5 shows that like in the clustering task, best results are achieved by the concatenated embedding triplets-sen-titles, which performs in the range between mean-vector and skip-thoughts-SICK.

Method	r	ρ	MSE
mean-vectors	0.757	0.673	0.4557
skip-thoughts-SICK	**0.858**	**0.791**	**0.287**
triplets-sen	0.797	0.704	0.372
triplets-titles	0.786	0.685	0.393
triplets-sen-titles	0.818	0.724	0.339

Table 5: Results on the SICK semantic relatedness subtask.

Table 6 presents some examples of predictions of triplets-sen-titles compared to the ground truth and to skip-thoughts-SICK predictions. The first pair is semantically equivalent as both methods detect. In the second pair, the first sentence is a negation of the second, but from the thematic point of view they are rather similar, thus assigned a relatively high score by our model.

sentences	GT	Tr	Sk
1. A sea turtle is hunting for fish 2. A sea turtle is hunting for food	4.5	4.2	4.5
1. A sea turtle is not hunting for fish 2. A sea turtle is hunting for fish	3.4	4.1	3.8

Table 6: Example predictions on the SICK data. GT = groundtruth, Tr=triplets-sen, Sk=skip-thoughts-SICK

6 Summary

In this paper we suggest a new approach for learning thematic similarity between sentences. We exploit the Wikipedia section structure to generate a large dataset of weakly labeled triplets of sentences with no human involvement. Using a triplet network, we learn a high quality sentence embeddings, tailored to reveal thematic relations between sentences. Furthermore, we take a first step towards exploring the versatility of these embeddings, by showing their good performance on the semantic similarity task. An interesting direction for future work is further exploring this versatility, by examining the performance of the embeddings on a variety of other NLP tasks.

Acknowledgments

We would like to thank Yoav Goldberg for helpful advise.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

Johanna Geiss. 2009. Creating a gold standard for sentence clustering in multi-document summarization. In *Proceedings of the ACL-IJCNLP Student Research Workshop*.

Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pages 40–48. Association for Computational Linguistics.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. *ArXiv e-prints*.

Elad Hoffer and Nir Ailon. 2015. DEEP METRIC LEARNING USING TRIPLET NETWORK. *ArXiv e-prints*.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Hamid Palangi, Li Deng, Yelong Shen, Jianfeng Gao, Xiaodong He, Jianshu Chen, Xinying Song, and Rabab Ward. 2016. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *ArXiv e-prints*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Ruslan Salakhutdinov Richard S. Zemel Antonio Torralba Raquel Urtasun Sanja Fidler Ryan Kiros, Yukun Zhu. 2015. Skip-Thought Vectors. *ArXiv e-prints arXiv:1506.06726*.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Noam Slonim, Gurinder Singh Atwal, Gaper Tkaik, and William Bialek. 2005. Information-based clustering. *PNAS*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

N. X. Vinh, J. Epps, and J. Bailey. 2009. Information theoretic measures for clusterings comparison. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*, pages 1480–1489.

Ting Yao, Fuchen Long, Tao Mei, and Yong Rui. 2016. Deep semantic-preserving and ranking-based hashing for image retrieval. In *IJCAI*, pages 3931–3937.

Elad Yom-Tov and Noam Slonim. 2009. Parallel pairwise clustering. In *SIAM International Conference on Data Mining*.

Bohan Zhuang, Guosheng Lin, Chunhua Shen, and Ian Reid. 2016. Fast training of triplet-based deep binary embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5955–5964.

Maciej Zieba and Lei Wang. 2017. Training triplet networks with gan. *arXiv preprint arXiv:1704.02227*.