Link Prediction with Mutual Attention for Text-Attributed Networks

Robin Brochier [1,2], Adrien Guille [1], Julien Velcin [1]

[1] Université de Lyon, Lyon 2, ERIC EA3083
[2] Peerus, DSRT F-69003

Introduction

We present an algorithm that learns a semantic similarity measure from a network of documents. We leverage the *Scaled Dot-Product Attention* (*SDPA*), a recently proposed attention mechanism, to design a mutual attention mechanism [1] between pairs of documents. To train its parameters, we use the network links as supervision.

Objective

From a network of document, we want to learn a semantic similarity measure that confront words representations from pairs of connected documents.

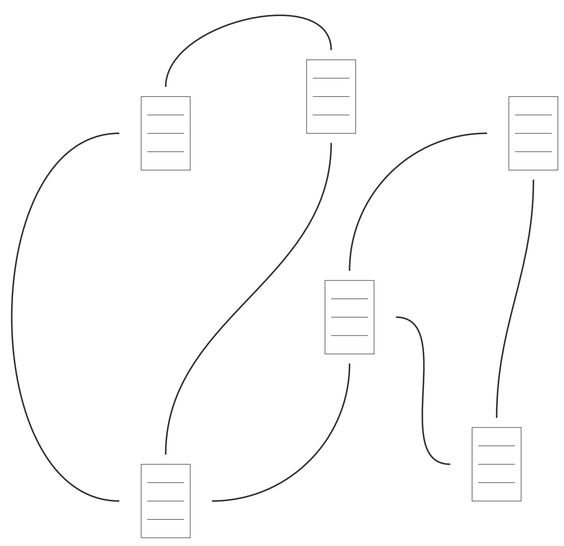

Figure 1: A hypothetical network of documents. Each node in the network is associated with a set of pre-trained word embeddings.

Table 1: Notations.

Sign	Description
N	Number of nodes/documents in the network.
C	Multiset of pairs of connected documents in the network.
W	Word embeddings of dimension D.
W^{t_u}	Matrix view of the L_u words embeddings corresponding to the document $u \in [1, N]$.
P^Q, P^K, P^V	Matrices of dimensions $D \times D$ that we seek to learn and which transform the word embeddings into queries, keys and values.
M_u^v	Mutual attention weights between the words of document u and the words of document v. M_u^v is of dimension $L_u \times L_v$. Note that $M_u^v \neq M_v^u$
e_u^v	D-dimensional representation of u given v.

Overall Optimization

For each pair of nodes (u, v) in the network, *MATAN* produces mutual representations e_u^v for u and e_v^u for v. We aim at optimizing three projection matrices, using noise-contrastive estimation following [2]:

$$\underset{P^Q,P^K,P^V}{\operatorname{argmin}} \sum_{(u,v)\in C} \Bigg(\log \sigma(e_u^v \cdot e_v^u) + \sum_{i=1}^{k} \mathbb{E}_{z\sim q}\big[\log \sigma(-e_u^z \cdot e_z^u) \big] \Bigg)$$

P^Q, P^K and P^V are meant to produce the mutual representations from the pre-trained word embeddings of documents u and v.

Projection matrices

From bags of word embeddings W^{t_u} and W^{t_v} associated to documents u and v, we generate three sets of word representations:

- queries $Q_u = W^{t_u} P^Q$;
- keys $K_v = W^{t_v} P^K$;
- values $V_v = W^{t_v} P^V$.

Mutual attention weights

Following the *SDPA* attention mechanism, each query of document u is confronted to each key of document v by dot-product, scaled by $\sqrt{D}$. To obtain probability laws over the words of v, we take the softmax of each row:

$$M_u^v = \text{softmax}\left(\frac{Q_u K_v^T}{\sqrt{D}}\right)$$

M_u^v is a matrix where each entry (i, j) is a compatibility weight between the i^{th} word of u and the j^{th} word of v.

Mutual attention vectors

Finally, a representation of u given v is constructed by (1) averaging the values of document v for each word of u given the attention weights in M_u^v and then (2) averaging these new attention word vectors:

$$e_u^v = \frac{\sum_{i=0}^{L} (M_u^v V_v)_i}{L_u}$$

Explainability

Here we show a heat map of M_u^v. For each word of u, we see the distribution of weights associated to the words of v. We see that few keys of v generate high compatibility weights whereas most queries do.

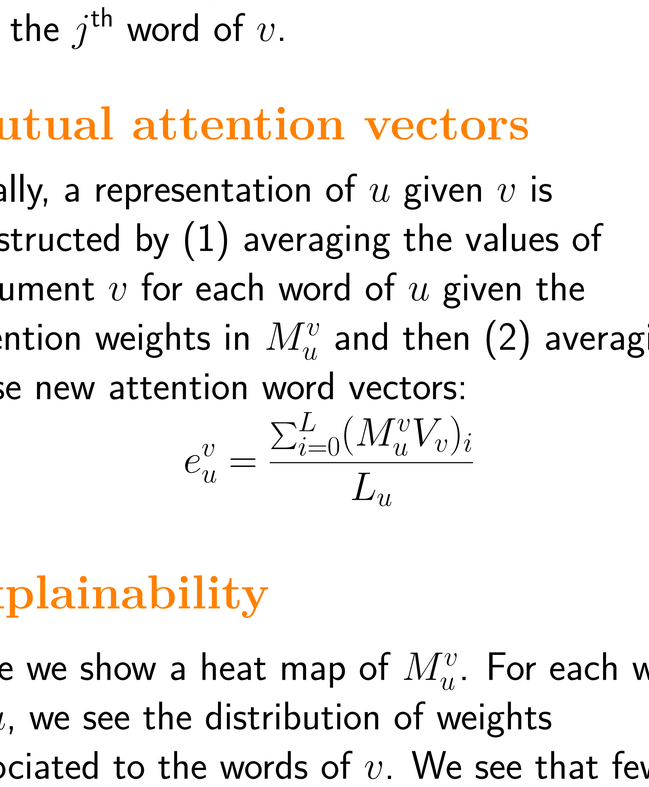

Figure 2: Heat map of the mutual attention weight matrix of the words of a document u given the words of a document v. Note that each row sums to one. The higher a weight is, the stronger is the contribution of the pair of word to the mutual representation.

By averaging the attention weight in the columns of M_u^v (respectively of M_v^u), we can capture the overall contribution of each of word of v (respectively of u). In the next figure, we show these weights by coloring accordingly the words in the documents.

support vector machines and kernel methods the new generation of learning machines kernel methods new generation of learning algorithms utilize techniques from optimization statistics and functional analysis to achieve maximal generality flexibility and performance these algorithms are different from earlier techniques used in machine learning in many respects for example they are explicitly based on theoretical model of learning rather than on loose analogies with natural learning systems or other heuristics they come with theoretical guarantees about their performance and have modular design that makes it possible to separately implement and analyze their components they are not affected by the problem of local minima because their training amounts to convex optimization in the last decade sizable community of theoreticians and practitioners has formed around these methods and number of practical applications have been realized although the research is not concluded already now kernel methods are considered the state of the art in several machine learning tasks their ease of use theoretical appeal and remarkable performance have made them the system of choice for many learning problems successful applications range from text categorization to handwriting recognition to classification of gene expression data

an introduction to support vector machines and other kernel based learning methods from the publisher this is the first comprehensive introduction to support vector machines svms new generation learning system based on recent advances in statistical learning theory svms deliver state of the art performance in real world applications such as text categorisation hand written character recognition image classification biosequences analysis etc and are now established as one of the standard tools for machine learning and data mining students will find the book both stimulating and accessible while practitioners will be guided smoothly through the material required for good grasp of the theory and its applications the concepts are introduced gradually in accessible and self contained stages while the presentation is rigorous and thorough pointers to relevant literature and web sites containing software ensure that it forms an ideal starting point for further study equally the book and its associated web site will guide practitioners to updated literature new applications and on line software

Figure 3: Average contribution of the mutual attention weights of the words of two connected documents in a citation network.

Experiments Results

We conducted two types of evaluations, the first where links are hidden and the second where nodes are hidden.

Table 2: Edges-hidden link prediction ROC AUC

% of training data	10%	20%	30%	40%	50%
NeMF	59.0	67.2	77.5	83.2	87.2
TADW	68.0	82.0	87.1	**93.2**	**94.5**
MATAN	**82.3**	**87.1**	**88.6**	90.9	91.0

Table 3: Nodes-hidden link prediction ROC AUC

% of training data	10%	20%	30%	40%	50%
TADW	64.2	**75.8**	**80.3**	**81.9**	**82.3**
MATAN	**69.4**	73.0	75.4	77.9	78.6

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.
Attention is all you need.
In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[2] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, and Emmanuel Müller.
Verse: Versatile graph embeddings from similarity measures.
In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 539–548. International World Wide Web Conferences Steering Committee, 2018.

[3] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Y Chang.
Network representation learning with rich text information.
In *IJCAI*, pages 2111–2117, 2015.

[4] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang.
Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec.
In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 459–467. ACM, 2018.

Contact

- Web: http://www.robinbrochier.com
- Email: robin.brochier@univ-lyon2.fr