Inductive Document Network Embedding with Topic-Word Attention

42nd European Conference on Information Retrieval
ECIR 2020

Robin Brochier, Adrien Guille, Julien Velcin

15.04.2020

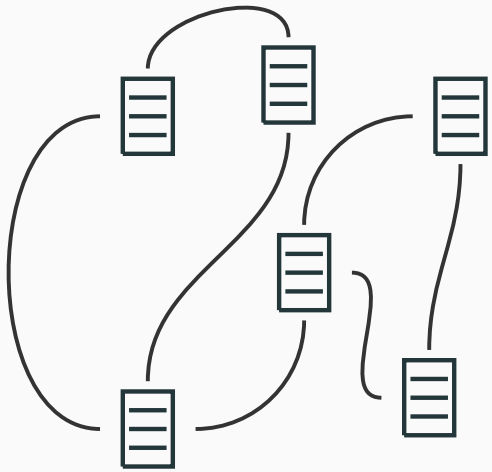
Université de Lyon (Lyon 2, ERIC)

Table of Contents

Introduction

Document Network Embedding (DNE)

- Idea: learning low-dimensional representations $D \in \mathbb{R}^{n_d \times p}$ for a structured text corpus *i.e.* when documents are linked to each other.
- Examples: question-and-answer websites, the scientific literature, social media...
- Applications: classification, link prediction, clustering...



n_d documents
n_w words
$A \in \mathbb{N}^{n_d \times n_d}$ adjacency matrix
$X \in \mathbb{N}^{n_d \times n_w}$ document-term matrix

Figure 1: Hypothetical network of documents.

Contributions

- A novel attention mechanism, **Topic-Word Attention (TWA)**, that produces **representations of text** where latent topic vectors attend to the word vectors of a document.
- A model, **Inductive Document Network Embedding (IDNE)** [1], that trains the parameters of TWA by leveraging the links of the document network.
- An evaluation which assesses the capacity of DNE algorithms to **induce** representations for unseen documents.
- A qualitative insight on the **interpretability** of the word and topic vectors learned by IDNE.

[1] https://github.com/brochier/idne

Related Works

DNE algotithms

- Text-Associated DeepWalk (**TADW**) [Yang et al., 2015] extends DeepWalk [Perozzi et al., 2014] to deal with textual attributes by constraining the factorization problem with pre-computed document representattions via LSA [Deerwester et al., 1990].
- Graph2Gauss (**G2G**) [Bojchevski and Günnemann, 2018] embeds each node as a Gaussian distribution instead of a vector. The algorithm is trained by passing node attributes through a non-linear transformation via a deep neural network (encoder).
- **GVNR-t** [Brochier et al., 2019] is a matrix factorization approach for document network embedding, inspired by GloVe [Pennington et al., 2014], that simultaneously learns word, node and document representations.

Attention Mechanism

- The Transformer [Vaswani et al., 2017] introduces a formalism of attention mechanisms for NMT. Given a query vector q, a set of key vectors K and a set of value vectors V, an attention vector is produced with the following formula:

$$a = \omega(qK^T)V.$$

- CANE [Tu et al., 2017] is an example of DNE algorithm that applies word-to-word attention, using convolutional neural nets and a mutual attention mechanism.
- In this paper, we propose a topic-word attention mechanism, where a set of shared topic vectors play the role of queries. **We want to contextualize some global topic representations given the words in each document.**

Inductive Document Network Embedding (IDNE)

Overview of IDNE

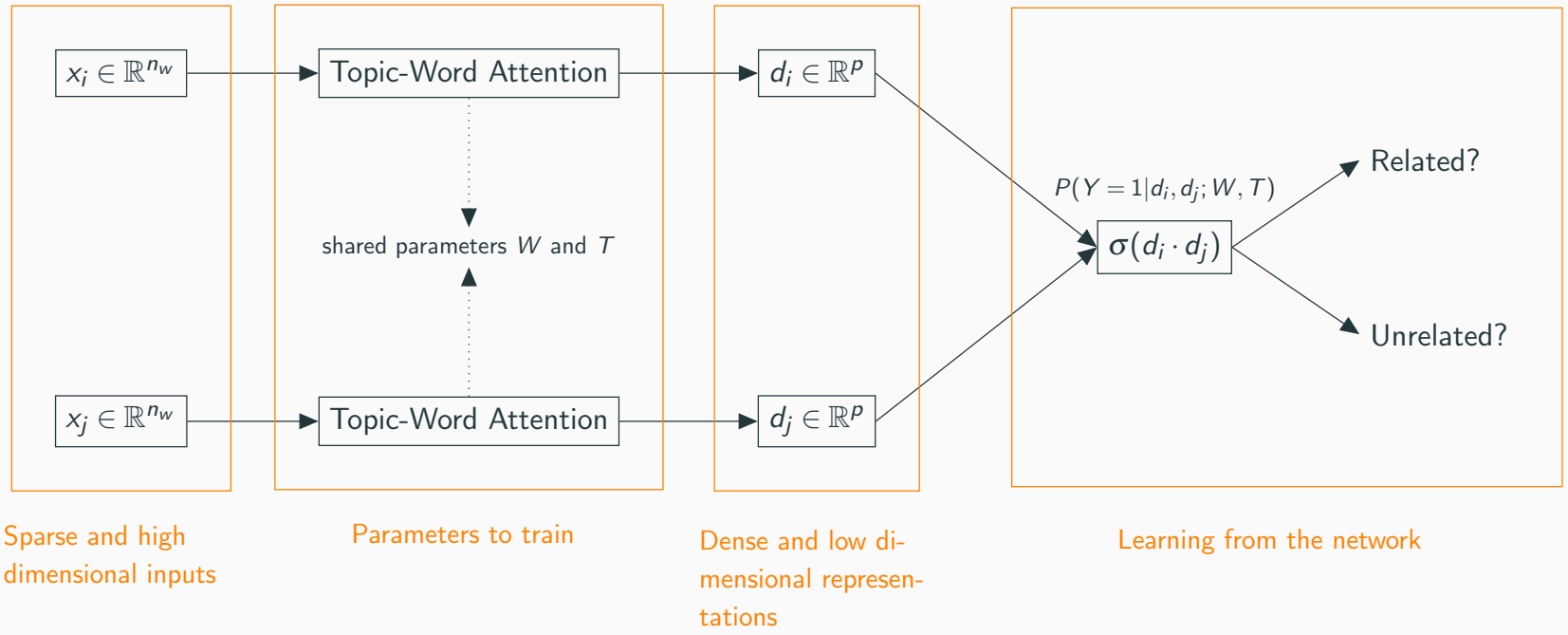


Figure 2: Overview of IDNE. Pairs of input document attributes x_i and x_j are sampled and transformed into dense representations d_i and d_j by the Topic-Word Attention (TWA) mechanism. The parameters of TWA are learned using the network structure by defining a probability of two documents to be related.

Topic-Word Attention (1)

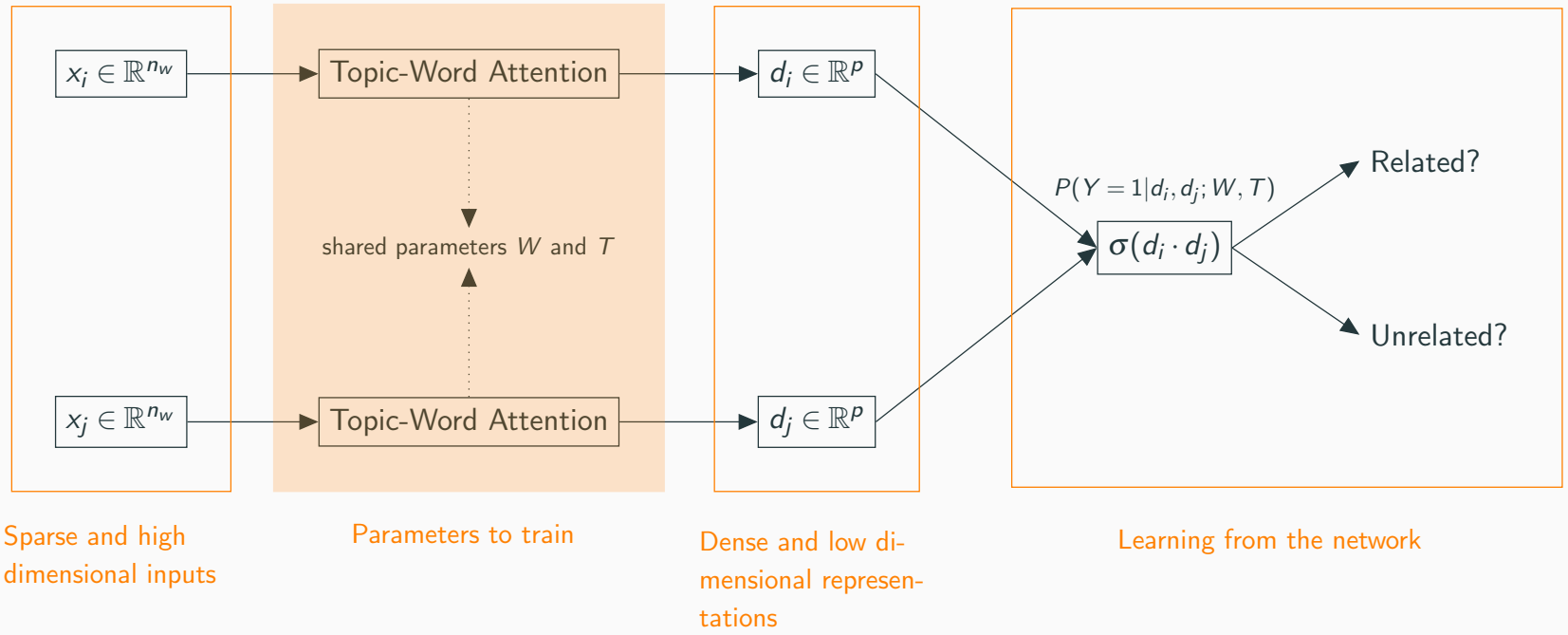


Figure 3: Overview of IDNE. Pairs of input document attributes x_i and x_j are sampled and tranformed into dense representations d_i and d_j by the Topic-Word Attention (TWA) mechanism. The parameters of TWA are learned using the network structure by defining a probability of two documents to be related.

Topic-Word Attention (2)

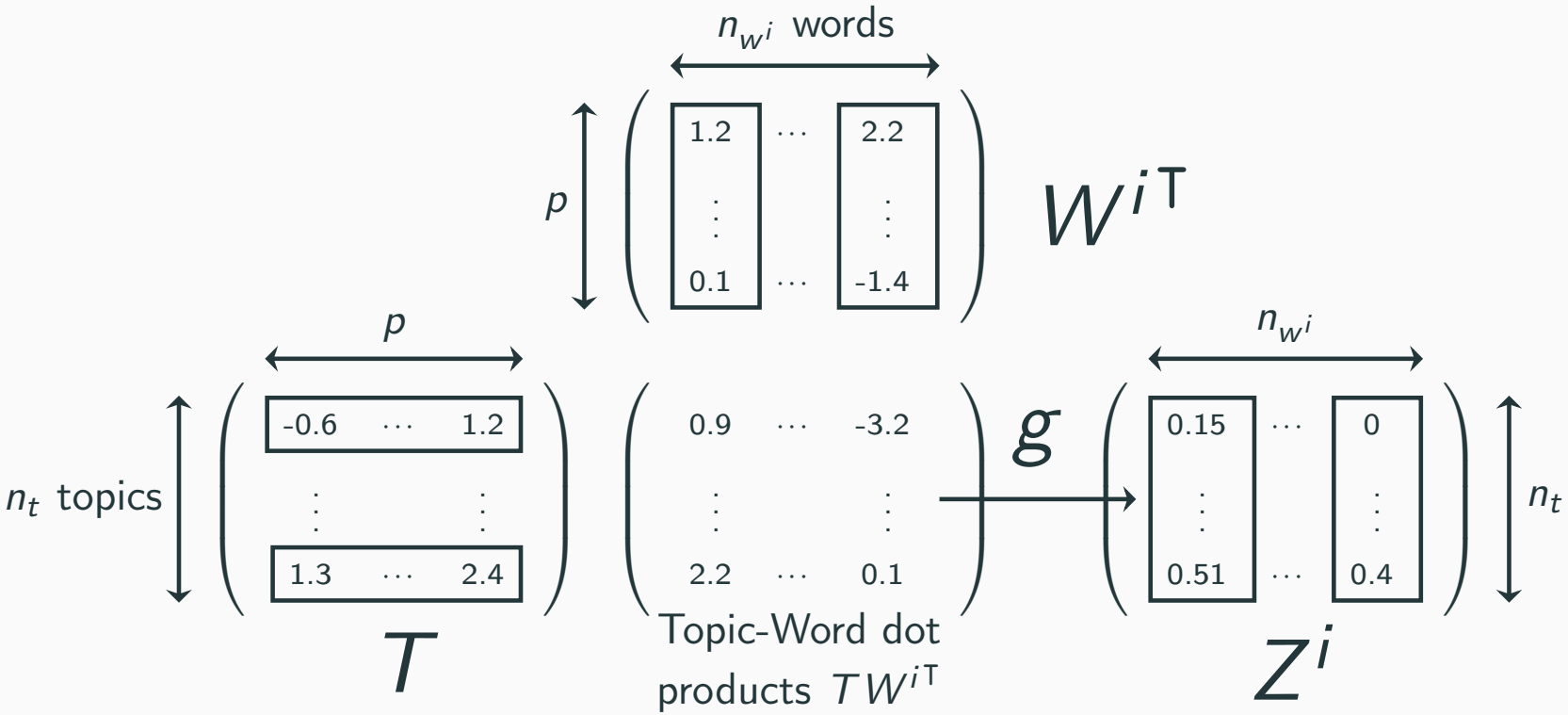


Figure 4: Matrix computation of the attention weights. W^i the lookup of the word vectors in W based on x_i. n_{w^i} denotes the number of distinct words in document i.

Topic-Word Attention (3)

- Attention weights: $Z^i = \mathrm{g}(TW^i)$.
- g is composed by a rectified linear unit $\mathrm{ReLU}(y) = \max(0, y)$ followed by a column-wise normalization.
- Topic-specific representations of the documents: $D_k^i = \frac{Z_k^i W^i}{|x_i|_1}$.
- Document final representation: $d_i = \sum_k D_k^i$.

Training (1)

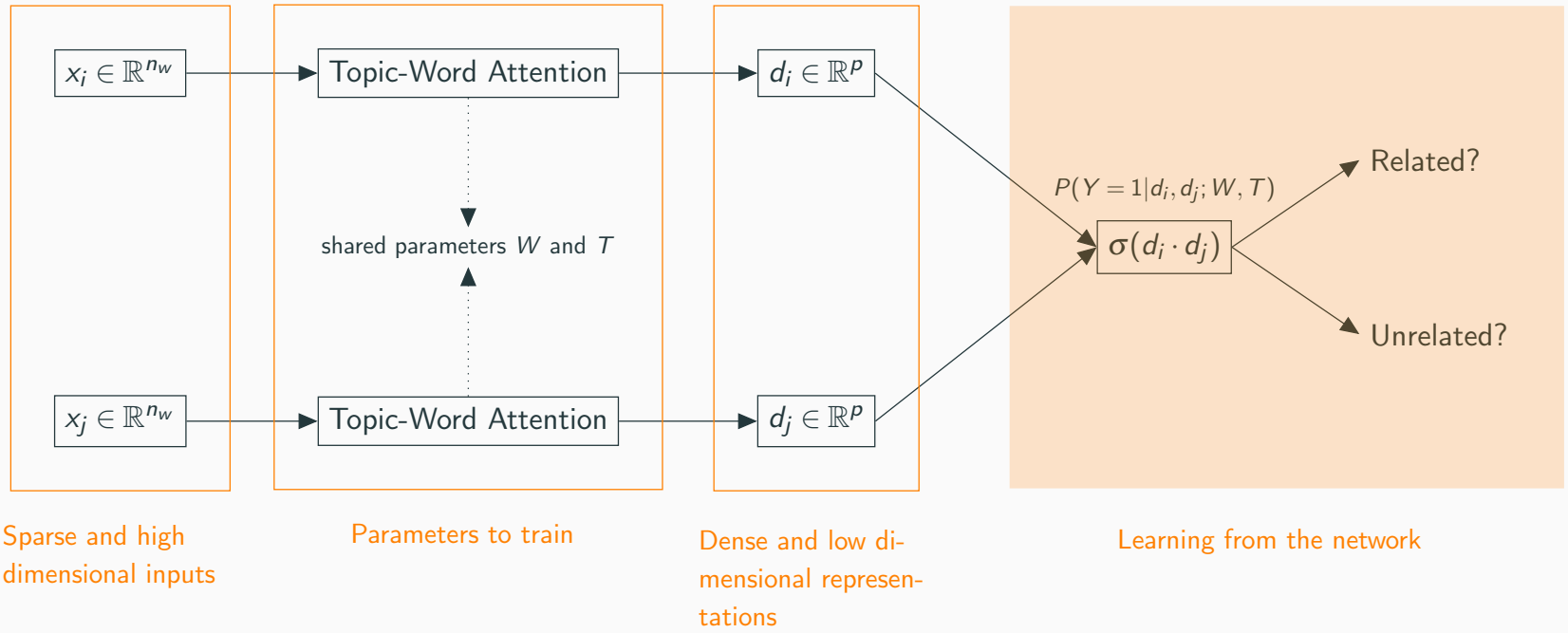


Figure 5: Overview of IDNE. Pairs of input document attributes x_i and x_j are sampled and tranformed into dense representations d_i and d_j by the Topic-Word Attention (TWA) mechanism. The parameters of TWA are learned using the network structure by defining a probability of two documents to be related.

Training (2)

- we define a binary matrix Δ between the documents from the network adjacency matrix where:

$$\delta_{ij} = \begin{cases} 1 & \text{if } (A + A^2)_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- probability of two documents to be related:

$$P(Y = 1 | d_i, d_j; W, T) = \sigma(d_i \cdot d_j).$$

- log-likelihood of Δ given W and T:

$$\begin{aligned} \ell(W, T) &= \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \log P(Y = \delta_{ij} | d_i, d_j; W, T) \\ &= \sum_{i=1}^{n_d} \sum_{j=1}^{n_d} \delta_{ij} \log \sigma(d_i \cdot d_j) + (1 - \delta_{ij}) \log \sigma(-d_i \cdot d_j). \end{aligned}$$

Training (3)

- Mini-batch SGD with the ADAM [Kingma and Ba, 2014] update rule.
- Rather than uniformly sampling entries of Δ, we sample 5000 balanced mini-batches in order to favor convergence.
- 16 positive examples ($\delta_{ij} = 1$) and 16 negative ones ($\delta_{ij} = 0$) per mini-batch.
- Positive pairs of documents are drawn according to the number of paths of length 1 or 2 linking them (given by the values of $A + A^2$).
- Negative samples are uniformly drawn.

Evaluation

“Traditional” transductive evaluation

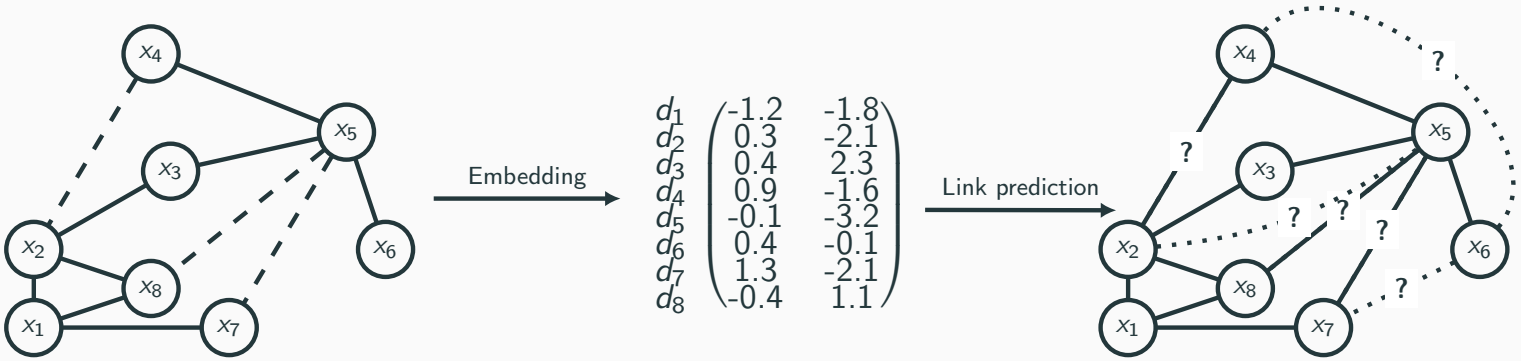


Figure 6: Transductive link prediction with network embedding.

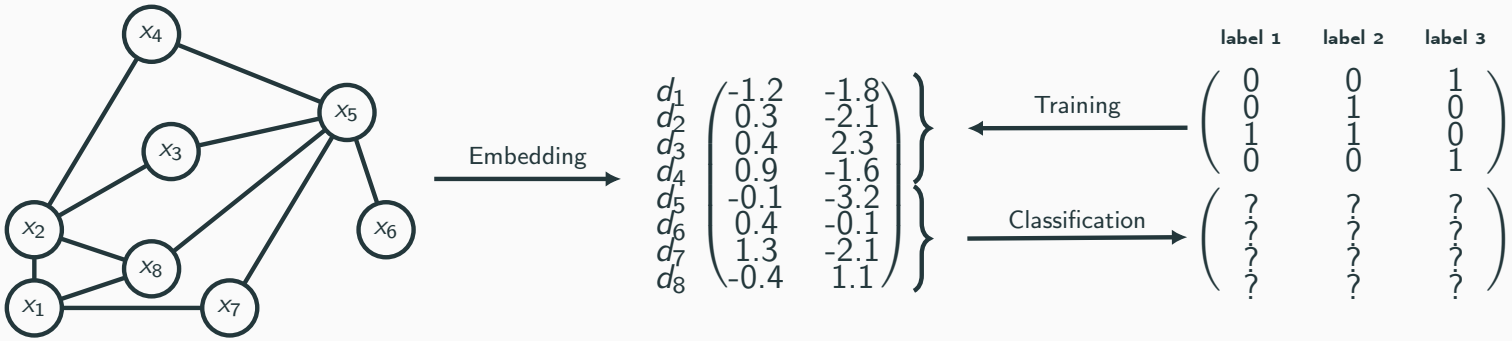


Figure 7: Transductive node classification with network embedding.

Inductive evaluation

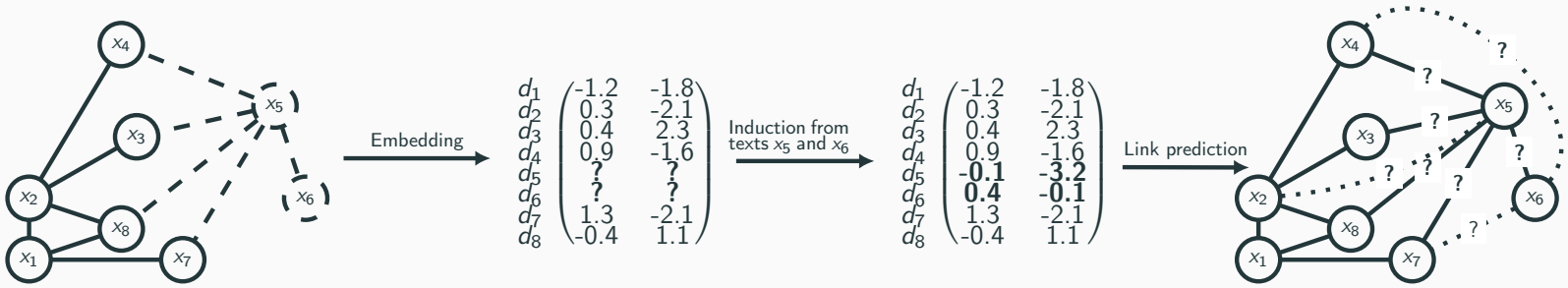


Figure 8: Inductive link prediction with network embedding.

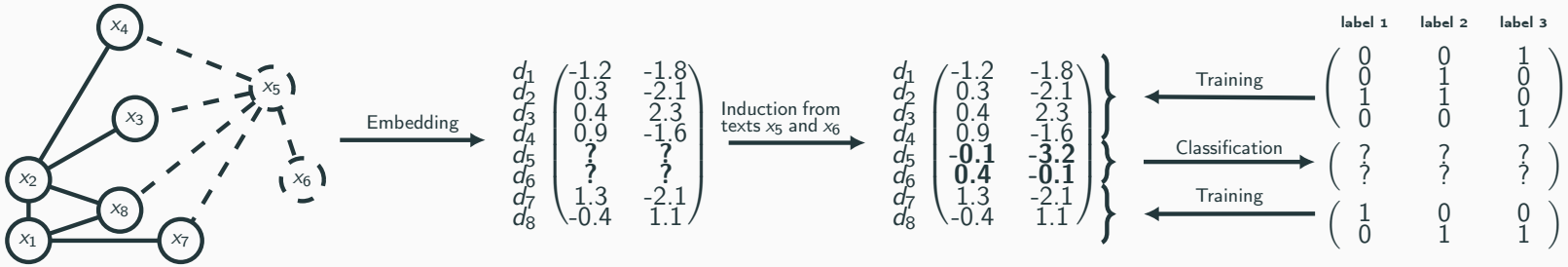


Figure 9: Inductive node classification with network embedding.

Induction with IDNE

Figure 10: To induce an embedding of a single document, IDNE simply applies TWA to the document-term representation of that document.

Datasets

- Cora: well-known citation network of scientific publications.
- New York Times (NYT) titles of articles from January 2007 linked according to common tags (*e.g.*, business, arts, technology) and labeled with the section they appear in (*e.g.*, opinion, news).
- Gaming and Travel Stack Exchange (Q&A websites): we keep questions and answers with more than 10 user votes. We use tags as labels.

Table 1: General properties of the studied networks.

	# docs	# links	# labels	vocab size	# words per doc	density	multi-label
Cora	2,211	4,771	7	4,333	67 ± 32	0.20%	no
NYT	5,135	3,050,513	4	5,748	24 ± 17	23.14%	no
Gaming	22,872	400,664	40	15,760	53 ± 74	0.15%	yes
Travel	15,087	465,696	60	14,539	70 ± 73	0.41%	yes

Experiment Results

Experiment Results: transductive classification (1)

Table 2: Micro AUC scores on Cora and NYT

	Cora					NYT				
	2%	4%	6%	8%	10%	2%	4%	6%	8%	10%
LSA	67.54	81.76	88.63	89.68	91.43	79.90	82.06	81.18	83.99	86.06
TADW	65.17	74.11	80.27	83.04	86.56	85.28	**88.91**	87.49	89.39	88.72
G2G	91.12	92.38	91.98	93.79	94.09	79.74	81.41	80.91	82.37	81.42
CANE	**94.40**	**95.86**	95.90	96.37	95.88	NA	NA	NA	NA	NA
GVNR-t	87.13	92.54	94.37	95.21	95.83	**85.83**	87.67	**88.76**	**90.39**	**89.90**
IDNE	93.34	94.93	**95.98**	**96.77**	**96.68**	82.40	84.60	86.16	86.72	87.98

Experiment Results: transductive classification (2)

Table 3: Micro AUC scores on Stack Exchange Networks

	Gaming SE					Travel SE				
	10%	20%	30%	40%	50%	10%	20%	30%	40%	50%
LSA	86.73	88.51	89.51	90.25	90.18	80.18	83.77	83.40	84.12	84.60
TADW	88.05	90.34	91.64	93.18	93.29	78.69	84.33	85.05	83.60	84.62
G2G	82.12	84.42	85.14	86.10	87.84	66.04	67.48	69.67	70.94	71.58
GVNR-t	89.09	92.60	94.14	**94.79**	95.24	79.47	83.47	85.06	85.85	86.58
IDNE	**92.75**	**93.53**	**94.72**	94.61	**95.57**	**86.83**	**88.86**	**89.24**	**89.31**	**89.26**

Experiment Results: inductive classification and link prediction

Table 4: Micro AUC scores by keeping 80% of the documents for learning.

	Inductive classification				Inductive link prediction			
	Cora	NYT	Gaming SE	Travel SE	Cora	NYT	Gaming SE	Travel SE
LSA	97.02	**89.45**	90.70	85.88	88.10	60.71	58.99	58.97
TADW	96.23	86.06	93.16	91.35	84.82	69.10	57.00	57.91
G2G	94.04	85.44	89.81	80.71	81.58	74.22	58.18	**59.50**
GVNR-t	**97.60**	88.47	**96.09**	**91.54**	82.27	71.15	59.71	58.39
IDNE	96.58	88.21	95.22	90.78	**91.66**	**77.90**	**62.82**	58.43

Experiment Results: qualitative insight

Table 5: Topics with their closest words produced by IDNE on Cora and words whose vector L_2 norms are the largest (resp. the smallest) reported in parenthesis. The labels in this dataset are: Case Based, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning and Theory.

Topic 1	casebased, reasoning, reinforcement, knowledge, system, learning, decision
Topic 2	chain, belief, probabilistic, length, inference, distributions, markov
Topic 3	search, ilp, problem, optimal, algorithms, heuristic, decision
Topic 4	genetic, algorithm, fitness, evolutionary, population, algorithms, trees
Topic 5	bayesian, statistical, error, data, linear, accuracy, distribution
Topic 6	accuracy, induction, classification, features, feature, domains, inductive
Largest	genetic (8.80), network(8.07), neural(7.43), networks (6.94), reasoning (6.16)
Smallest	calculus (0.34), instability (0.34), acquiring (0.34), tested (0.34), le (0.34)

Conclusion

Conclusion

- IDNE performs well and achieves near SOTA scores for all tasks.
- The learned topic and word vectors are interpretable.
- Studying the impact of the choice of Δ might be interesting. What topics appear if we use structural similarities such as SimRank [Jeh and Widom, 2002]?
- What about applying TWA to other NLP tasks?

Thank you!

Bibliography i


Bojchevski, A. and Günnemann, S. (2018).
Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking.
In *International Conference on Learning Representations*.


Brochier, R., Guille, A., and Velcin, J. (2019).
Global vectors for node representations.
In *The World Wide Web Conference*, pages 2587–2593. ACM.


Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41(6):391–407.

Bibliography ii


Jeh, G. and Widom, J. (2002).
Simrank: a measure of structural-context similarity.
In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM.


Kingma, D. and Ba, J. (2014).
Adam: A method for stochastic optimization.
International Conference on Learning Representations.


Pennington, J., Socher, R., and Manning, C. (2014).
Glove: Global vectors for word representation.
In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Bibliography iii


Perozzi, B., Al-Rfou, R., and Skiena, S. (2014).
Deepwalk: Online learning of social representations.
In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.


Tu, C., Liu, H., Liu, Z., and Sun, M. (2017).
Cane: Context-aware network embedding for relation modeling.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1722–1731.

Bibliography iv


Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Advances in neural information processing systems*, pages 5998–6008.


Yang, C., Liu, Z., Zhao, D., Sun, M., and Chang, E. (2015).
Network representation learning with rich text information.
In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.