Tikk et al. BMC Bioinformatics 2010, 11(Suppl 5):P8
http://www.biomedcentral.com/1471-2105/11/S5/P8

BMC Bioinformatics

POSTER PRESENTATION Open Access

A fast and effective dependency graph kernel for PPI relation extraction

Domonkos Tikk[1,2*], Peter Palaga[1], Ulf Leser[1]

From Workshop on Advances in Bio Text Mining
Ghent, Belgium. 10-11 May 2010

Background

Extraction of protein-protein interactions (PPIs) reported in scientific publications is a core topic of biomedical text mining. The ultimate goal is to devise a PPI extraction method that performs well on large amount of unseen text independently from the training corpus. One popular, machine-learning based approach to PPI extraction builds on the convolution kernels, i.e., similarity functions defined on the parse-based representation of sentences and interactions. Kernel functions differ in (1) the underlying sentence representation (bag-of-words, syntax tree parse, dependency graphs), (2) the substructures retrieved from the sentence representation to define interactions, and (3) calculation of the similarity function.

Method

We present a novel kernel method called k-band shortest path spectrum kernel (kBSPS), an extension of the spectrum tree kernel (SpT) [1]. It combines three ideas: First, interactions are represented as vertex-walks as in SpT but adapted to dependency graphs. The kBSPS kernel includes also edge labels into vertex-walks, thus also exploiting the dependency type of a relationship. Second, it uses a novel similarity function on vertex-walks permitting certain mismatches, thus allowing for linguistic variations. The tolerant matching distinguishes three types of nodes: dependency types (D), candidate entities (E), and other surface tokens (L). Mismatches / matches are scored differently depending on the type of nodes. Third, apart from the shortest path between the proteins of the candidate interaction, kBSPS also adds all nodes within distance k from this path to the vertex-walk representation.

Results

We evaluated kBSPS kernel on the 5 standard PPI benchmark corpora (AIMed, BioInfer, HPRD50, IEPA, LLL)

Table 1 Comparison of kBSPS in terms of AUC, F_1-measure and classification time with other state-of-the-art kernels using the CV and CL evaluation scenarios

Kernels	AIMed			BioInfer			IEPA		
	AUC	F_1		AUC	F_1		AUC	F_1	
	CV/CL	CV/CL	time(s)	CV/CL	CV/CL	time(s)	CV/CL	CV/CL	time(s)
SL	83.5/**77.5**	54.5/42.6	10.8	**81.1/74.9**	**60.0**/46.2	24.0	81.1/75.6	69.3/60.4	1.8
SpT	66.1/56.8	27.3/28.6	44.5	74.1/64.2	53.4/43.0	79.8	75.9/54.2	64.7/15.5	0.6
APG	**84.6/77.6**	**56.2/43.8**	3.7	**81.5**/69.6	**60.7**/39.1	6.2	**83.9/82.4**	**73.1**/59.6	0.4
kBSPS	75.1/72.1	44.6/40.3	0.4	75.2/73.3	55.1/**47.6**	1.5	**83.2**/81.0	70.5/**70.7**	< 0.1

Legend: For brevity, only the results on the three largest corpora are shown. Best results are highlighted with bold typesetting (differences under 1 base point are ignored). We reran all experiments since CL results are hardly available. SL - shallow linguistic kernel [1], SpT - spectrum tree kernel [2], APG - all path graph kernel [3].

* Correspondence: tikk@informatik.hu-berlin.de
[1]Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
Full list of author information is available at the end of the article

using document-level 10-fold cross-validation (CV) and cross-learning (CL; 4-vs-1) evaluation. CV evaluation is somewhat biased, because the training and the test data have very similar corpus characteristics and machine learners tend to learn that, therefore CL evaluation, where the training and test data sets are drawn from different distributions, provides a more unbiased picture. Our results are compared with three state-of-the-art kernel approaches to PPI extraction (see Table 1).

Conclusion

We have shown that kBSPS kernel is on par with state-of-the-art kernels at the more general CL evaluation. Furthermore, its performance is more stable (drops the least from CV) than other methods. Notably, kBSPS is also much faster than any other kernel, making it applicable to very large corpora.

Acknowledgements

Domonkos Tikk was supported by the Alexander-von-Humboldt Foundation.

Author details

[1]Knowledge Management in Bioinformatics, Institute for Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. [2]Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar Tudósok krt 2., Hungary.

Published: 6 October 2010

References

1. Kuboyama T, Hirata K, Kashima H, Aoki-Kinoshita KF, Yasuda H: **A spectrum tree kernel.** *Information and Media Technologies* 2007, **2**:292-299.
2. Giuliano C, Lavelli A, Romano L: **Exploiting shallow linguistic information for relation extraction from biomedical literature.** *Proc. of the 11st Conf. of the European Chapter of the Association for Computational Linguistics (EACL'06)* Trento, Italy: The Association for Computer Linguistics 2006, 401-408 [http://acl.ldc.upenn.edu/E/E06/E06-1051.pdf].
3. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, *et al*: **All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning.** *BMC Bioinformatics* 2008, **9(Suppl 11)**:S2.

doi:10.1186/1471-2105-11-S5-P8
Cite this article as: Tikk *et al.*: A fast and effective dependency graph kernel for PPI relation extraction. *BMC Bioinformatics* 2010 **11**(Suppl 5):P8.