Errata for All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning

Antti Airola

February 27, 2009

This is an errata for the article All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning [1, 2].

The behavior of the implementation of the graph kernel used in our experiments deviates from how the method is described in the article. The difference is in the construction of the matrix G, which is introduced towards the end of the section "The all-paths graph kernel".

Let $L \in \mathbb{R}^{m \times n}$ be the label allocation matrix and $W \in \mathbb{R}^{n \times n}$ the final adjacency matrix of the graph. Then, the definition given for G in the article is

$$G = LWL^{\mathrm{T}}. \tag{1}$$

When written open as a sum this becomes

$$G_{i,j} = \sum_{k=1}^{n} \sum_{l=1}^{n} L_{i,k} W_{k,l} L_{j,l}.$$

However, when computing the values of the entries in matrix G, the implementation used in the experiments erroneously had a reassignment operation in place of the sum operation. The resulting behavior can be very closely approximated by the following definition for G.

$$G_{i,j} = \max_{1 \leq k,l \leq n} \{L_{i,k} W_{k,l} L_{j,l}\}. \tag{2}$$

The resulting difference is that instead of summing together the weights of all paths connecting two labels, we take the maximum over these.

The results reported in [1, 2] are based on using (2). Further experiments seem to indicate, that using the latter definition of G can lead to better

performance, than using the former definition (roughly, by order of 2–3 F-score units on the five corpora). The evidence is however not conclusive, as the differences fall within variance in performance estimation, and contrary results have also been observed in a re-implementation of the method. We note that both (1) and (2) lead to valid kernels, and thus we encourage anyone using the graph kernel to try both variants of the method.

The graph kernel implementation that was made available at `http://mars.cs.utu.fi/PPICorpora/GraphKernel.html` also originally used definition (2). An implementation of definition (1) has been added to the software to facilitate replication efforts.

Acknowledgments

We would like to thank Erik Fässler for bringing this matter to our attention.

References

[1] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics, special issue*, 9(Suppl 11):S2, 2008.

[2] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 1–9, 2008.