Identification of concepts bridging diverse biomedical domains

Matjaž Juršič[1*], Igor Mozetič[1], Miha Grčar[1], Nada Lavrač[1,2]
[1]Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
[2]University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

*matjaz.jursic@ijs.si

Background

In biology and medicine, experts are challenged daily with linking information from various highly specialized subfields. The individual subfields can be considered habitually different domains since experts usually master only one of them. However, many novel discoveries are achieved by gaining new insights and knowledge via fusing two or more diverse fields. In this work we propose a method that reveals key concepts which are the most informative and promising to pursue when bridging diverse domains. We evaluate the results against manually selected bridging concepts studied in papers [1] and [2].

Materials and methods

This work focuses on identifying bridging concepts (*bridging terms* or *b-terms*) in two datasets, each consisting of a pair of domains. The training dataset consists of titles of articles about migraine (first domain) and magnesium (second domain) with b-terms identified in [1]. In the testing dataset are abstracts about autism and calcineurin with b-terms presented in [2]. In these two pairs of domains (retrieved from PubMed) b-terms are known and verified by the expert to provide potential new discoveries in the field.

Our methodology of b-term detection is the following: 1. Employ text mining to pre-process the texts and encode them in the bag-of-words representation; 2. Calculate the heuristics which favour b-terms over other terms; 3. Sort terms by the best heuristic measure and present the top terms (hopefully representing b-terms) to the expert during interactive exploration of the two domains.

The search for the most promising heuristic is based on two phases: 1. Training – we propose over 40 heuristics, from very simple term-frequency statistics to very elaborate combined measures. We evaluate their quality on the first dataset and select the best one, the so-called *b-potential measure* calculated as a multiplication of the term's tf-idf weights in the two centroids of the two domains. 2. Testing - we evaluate the b-potential measure on the second dataset to confirm its domain independence and quality of b-term identification.

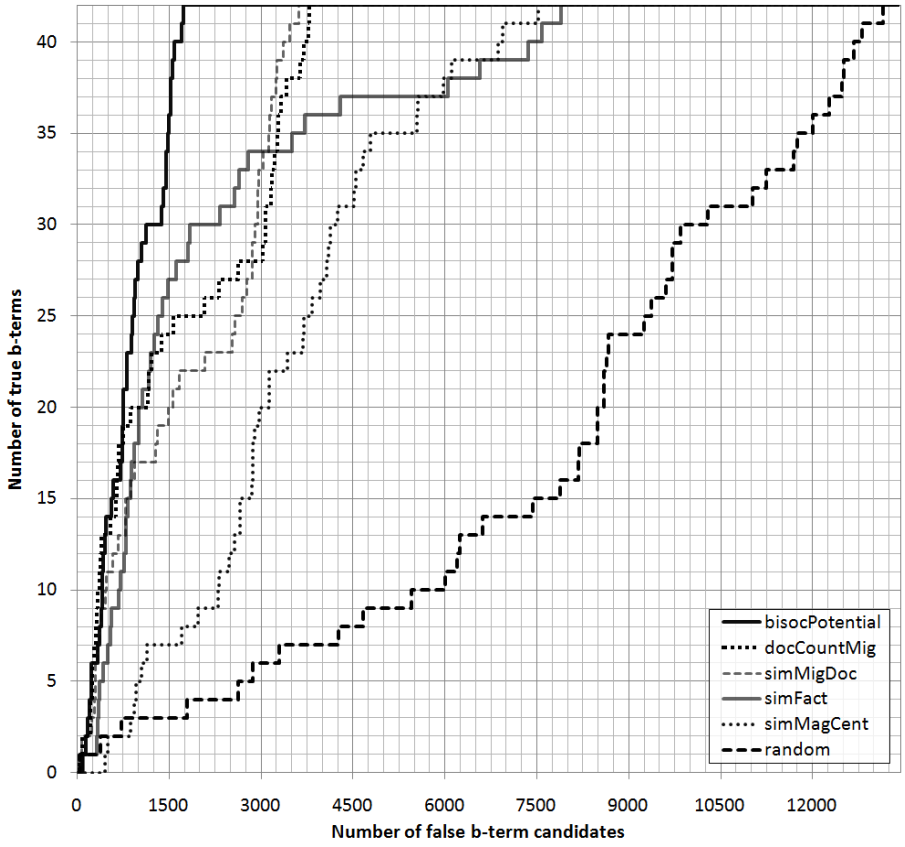
Results and conclusion

We experimentally confirmed that the method for identification of concepts bridging diverse biomedical domains using the proposed *b-potential* measure is the best heuristic for b-term detection and is able to retrieve b-terms approximately 7 times faster compared to a random approach (see Figure 1). Consequently, the b-term identification from the papers [1] and [2] would be considerably simplified by using the b-potential sorted list of terms presented to the experts for a manual selection (as the top of such sorted list is 7 times more probable to contain a b-term in comparison to a random list).

Acknowledgement

This work was partially supported by the Slovenian national project Knowledge Technologies and by the EU project FP7-211898 BISON.

Figure 1. ROC curves to evaluate different heuristics for ranking of b-terms on the migraine-magnesium training dataset. A curve is constructed by drawing a vertical line when a term is indeed a b-term, and a horizontal line when a term is not a b-term. Therefore, the y-axis shows the number of b-terms and the x-axis shows the number of non-b-terms. The figure presents a selection of the best heuristics, where a comparison of the best, b-potential (leftmost solid line) with the random heuristics (rightmost dashed line) is indicative.



References

1. Swanson, DR: **Migraine and magnesium: eleven neglected connections**. Perspectives in Biology and Medicine 1988, 31(4):526-557.

2. Petrič I, Urbančič T, Cestnik B, Macedoni-Lukšič M: **Literature mining method RaJoLink for uncovering relations between biomedical concepts**. J. Biomed. Inform. 2009, 42(2):219-227.