Why Biomedical Relation Extraction Results are Incomparable and What to do about it

Sampo Pyysalo,[1] **Rune Sætre,**[2] **Jun'ichi Tsujii**[2] and **Tapio Salakoski**[1]
[1]Turku Centre for Computer Science (TUCS) and Dept. of IT, University of Turku
20014, Turku, Finland
{sampo.pyysalo,tapio.salakoski}@it.utu.fi
[2]Department of Computer Science, the University of Tokyo, Japan
Bunkyo-ku, Hongo 7-3-1, Tokyo, Japan
{rune.saetre,tsujii}@is.s.u-tokyo.ac.jp

Abstract

A large number of biomedical relation extraction methods, targeting for example protein-protein interactions (PPI), have been introduced in the preceding decade. However, the performance figures reported for these methods vary enormously, and results are largely incomparable across different studies. In this paper we study reasons leading to this situation and propose a solution to resolving them.

1 Introduction

Evaluation results for biomedical relation extraction methods vary greatly and are largely incomparable across different studies. This makes it difficult to assess what are the best tools, methods, techniques and general approaches to the task. A number of recent studies have brought to light several issues leading to this incomparability. In this paper we collect together these findings and discuss several other aspects of relation extraction experiments that may introduce unwanted variance into evaluation results. After reviewing the problems, we propose a solution to the known issues.

We assume throughout the paper the common task setting where relations are to be extracted by identifying entity pairs for which the relation holds, e.g. two proteins that are stated to interact. While a machine-learning perspective is involved in some parts of the discussion, most of the problems can occur for any extraction approach. We assume that evaluation aims to be able to establish differences in the performance of compared methods on the order of a few percentage units or less, a level of accuracy at least implicitly assumed in many comparisons of domain extraction methods but, as we shall discuss next, far from systematically achieved at present.

2 The problems

2.1 Different corpora

In a recent study of biomedical relation extraction performance across five corpora, Pyysalo et al. (2008) demonstrated that evaluation results for a single method on different corpora may vary up to 30%, and found a 19% average performance difference on the corpora. These differences stem in part from different definitions of what should or should not be extracted as a protein-protein interaction, which leads to differing positive/negative distributions of candidate relations: for example, the LLL corpus (Nédellec, 2005) contains 164 "true" (positive) relations out of 330 possible entity pairs, giving an "all-true" baseline performance of 66% F-score[1], while for the AIMed corpus (Bunescu et al., 2005) these figures are approx. 1000 positive out of 5800 candidate pairs for a baseline performance of 29% F-score.

While differing extraction targets are, in general, a benefit for evaluation—extraction approaches should be able to learn different targets—these differences render (unqualified) evaluation results from different corpora incomparable. Below, we will only consider factors complicating evaluation on a shared corpus.

[1]Assigning all candidates into the positive class gives a r(ecall) of 100% and a p(recision) of $\frac{c_p}{c_p+c_n}$, where c_p and c_n are the number of positive and negative candidates (resp.); F-score is $\frac{2pr}{p+r}$.

2.2 Corpus processing

Biomedical corpus annotation is rarely, if ever, distributed in a form that would explicitly specify the set of candidate relations. Instead, candidates must be generated, often from annotation that only specifies entities and positive relations. Negative relations are typically generated under the closed-world assumption. Along with various other details of annotation schemes, this opens the door to varying interpretations of single corpora.

2.2.1 Number of generated examples

With complex annotations including for example nested or noncontinuous entities, corpus annotation can allow for strikingly different numbers of positive and negative relations: Sætre et al. (2008) note that the AIMed corpus has been variously interpreted as containing between 951 and 1071 positive relations with 4026–5631 negative ones. For the most favorable combination (1071 positive, 4026 negative) the all-true baseline would stand at 35% and for the least favorable (951/5631) at 25% F-score. Thus, different preprocessings of the corpus can give a very large absolute difference even for a trivial baseline, rendering results for different preprocessings of the corpus incomparable.

A particular difficulty is presented by the existence of self-interactions, where an annotated (positive) relation involves only a single entity. While the AIMed corpus contains 54 such interactions, most studies on AIMed simply ignore their existence, since generating candidate relations involving only single entities would increase the number of negative candidates by thousands and lead to a considerably more difficult positive-negative ratio. A similar situation occurs when extracting directed relations: if each pair of entities is used to generate two directed candidate relations, the number of negative examples will more than double.

2.2.2 Entity name blinding

Biomedical corpora often focus on limited subdomains, either by design or due to bias introduced from document selection procedure (e.g. documents cited as evidence in an interaction database). Consequently, corpora can contain a disproportionate amount of relations between particular entities, which can be "memorized" by a learner if it is allowed to see their names. For example, in an experiment on the AIMed corpus we got an F-score of 33% when *only* the names of the candidates were used as features. As the all-true baseline is 30% for our version of the corpus, this suggests that memorizing names can provide a small but non-negligible benefit, again leading to diverging results. Extraction methods should be able to detect relations between entities whose names have not occurred in their training data—indeed, such novel interactions are more interesting than those already annotated. Thus, performance increments based on knowing the names of the entities involved do not reflect real benefits of extraction methods.

A related issue arises on corpora involving nested entities. For example on the AIMed corpus, the dataset applied in (Giuliano et al., 2006) appears to have been preprocessed so that nested entity names were treated differently depending on whether the inside entity was part of a true relation or not. For example, in the sentence *Cloning and functional analysis of* $[_1$*BAG-1] : a novel* $[_2[_3$*Bcl-2]-binding protein] with anti-cell death activity* there are three potential pairs (1,2), (1,3) and (2,3), but in the Giuliano dataset only two pairs for this sentence are given, one false pair, (1,3), and one true pair, (1,2), where the representation of the latter does not involve marking the tokens *-binding protein* as belonging to a protein name (and thus blinding). The negative candidate pair (2,3) is excluded in this case. Removing negative nested protein names raises evaluated performance in terms of F-score by increasing the positive/negative ratio. However, this way of preprocessing the data should not be performed unless there is a way to know in advance whether a nested entity is involved in a relation or not before running the extraction method. Comparison of evaluations where one employs such information and the other does not may not yield meaningfully comparable results: Airola et al. (2008) ran the method published by Giuliano et al. (2006) on a differently blinded version of AIMed and reported a 52.4% F-score, over 6% points lower than the 59.0% reported by Giuliano et al.

2.3 Experimental setup

There are numerous potential pitfalls in setting up a relation extraction experiment, in particular when it involves machine learning. Two frequently encountered issues relate to the role of training and test sets in evaluation.

2.3.1 Isolating training and test data

To establish a meaningful estimate of generalization performance, the training and test sets must represent independent samples: test data that resembles the training data more than the overall distribution benefits overfit learners and leads to overestimation of performance.

Sætre et al. (2008) observed that a number of biomedical relation extraction studies performed cross-validation by first preprocessing the data to form all the possible candidate pairs of related entities, which were then randomly split into different sets for training and evaluation. In this procedure, pairs from the same sentence ended up being used both for training and testing within a single fold. Since the features from two neighboring pairs in a sentence are practically identical, this was shown to lead to an 18% points overestimation of the F-score performance compared to a more realistic setting. In the realistic test setting, all the data from a single abstract is kept together through the whole processing pipeline, to avoid using it both for training and testing in the same fold.

2.3.2 Parameter selection

The data on which methods are tested should, ideally, represent completely new, unseen data. While this ideal is rarely achieved, a small number of tests on the whole dataset is unlikely to cause much bias. However, experiments are often set up to include repeated, systematic tests on the entire dataset, of which the best result is reported. Perhaps the most frequent such setup arises from parameter selection, e.g. using cross-validation on the entire corpus. Especially when the parameter space is multi-dimensional and the data set is small, this approach can find considerable benefit from identifying "spikes" in the parameter space. Evaluation necessarily involves some random variation for different parameter settings, and a parameter selection protocol that allows the test set to be seen will yield an overestimate of performance relative to the magnitude of that variation. On smaller corpora (e.g. LLL), random effects changing the assignment of just a few examples can already make a percentage unit difference in results.

A related issue arises from picking the best point (e.g. in terms of F-score) from a precision-recall curve generated for a single extraction method with fixed overt parameters. This corresponds to implicitly optimizing a classification threshold parameter, again with reference to the whole dataset. When comparing methods with otherwise similar performance, these differences can cause misleading results: Using the method of Airola et al. (2008) on AIMed, picking the optimum threshold was estimated to provide at least a 2% overestimate over the more realistic setting of selecting the threshold on the training data.[2]

2.4 Metrics

Even when the same corpus, preprocessing, experimental setup, and metric are applied, differences arising from the details of how the metric is calculated can cause results to deviate.

2.4.1 Extracting Identical Relations

A relation is typically taken to be correctly extracted if the (unordered) pair of related entities is identified. However, this definition leaves open a question relating to entity identity: are two mentions of the same name one or two entities, and consequently, should two relations annotated between the same two names both be extracted, or does it suffice to find either one?

Giuliano et al. (2006) termed two answers to these questions One Answer per Occurrence in a Document (OAOD) and One Answer per Relation in a Document (OARD): here the OAOD criterion requires each mention to be extracted, while OARD only demands that each unique pair of names is identified. They found that an otherwise identical evaluation yielded an F-score of 59% under the OAOD criterion and 64% under OARD, indicating that results evaluated using different criteria cannot be directly compared.

The two alternatives studied by Giuliano et al. are not the only ones possible: we might propose One Answer per Sentence, One Answer per Corpus, One Answer per (cross-validation) Fold, or One Answer per Journal. While one might argue that extracting each relation from the corpus once suffices for some practical applications, we take the view that from the evaluation perspective the specific names (between which relations are stated) are of secondary importance and suggest that each relation be considered. That is, One Answer per Occurrence; from this perspective, the "D" in "OAOD" is superfluous.

[2] Thanks to Antti Airola for running this number for us.

2.4.2 Averages

How averages are calculated is a lesser, but not negligible, issue. This question often arises from cross-validation, where two basic alternatives are available: either calculate performance for each fold separately and average the results (macroaveraging), or pool the answers and calculate one result for the entire dataset (microaveraging). Different choices might cause non-trivial differences in otherwise identical setups for small corpora: for example, when examples are carefully divided into cross-validation folds on the document level, some test sets can contain documents with unusually high numbers of entities and thus of candidate relations. With macroaveraging, folds with a large number of relations will contribute equally to the final result as folds with fewer, whereas if results are pooled the contributions of folds will be inequal, but each relation will contribute equally. As the number of candidate relations grows quadratically with the number of entities in a given context and the growth of positive relations is likely to be slower, we would expect folds with more relations to represent more difficult problems in terms of metrics sensitive to the positive/negative distribution (e.g. F-score) and thus macroaveraged results to be higher.

3 A proposal for a solution

The problems discussed above highlight a need for standardization to establish meaningful comparisons between different relation extraction method evaluations. Before these issues are addressed to some extent, the only direct comparisons between methods that can be meaningfully performed are those done within a single study (or at least by the same authors) and those from shared tasks. The incomparability comes at a great cost to the community, as reimplementation is often the only way to reliably determine the relative merits of proposed methods.

We do not expect that specific choices to the many alternatives discussed could be enforced by fiat. Instead, we propose a positive solution: we have constructed a standard dataset containing data derived from different corpora, building on the unification of five corpora under a common format by Pyysalo et al. (2008). We have extended this work by including explicit candidate pairs with blinded protein names, thus addressing the issues in corpus processing. Further, predefined train/test splits are provided, and the distribution of the dataset is accompanied with evaluation scripts that implement the basic metrics in a standardized way, thus eliminating possible differences arising from metric application. The data and software is freely available from `http://mars.cs.utu.fi/PPICorpora`.

4 Conclusion

We have discussed a number of issues in biomedical relation extraction system evaluation that complicate, or even prevent, meaningful comparison of reported results, and we proposed a solution to address these issues. We believe that the proposed dataset and evaluation approach can serve as a step toward stable, reliable evaluation of biomedical relation extraction methods.

Acknowledgments

The work was partially supported by the Academy of Finland, Grant-in-Aid for Specially Promoted Research (MEXT, Japan) and Genome Network Project (MEXT, Japan).

References

Antti Airola, Sampo Pyysalo, Jari Björne, , Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the BioNLP'08*, pages 1–9.

Razvan C. Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun Kumar Ramani, and Yuk Wah Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of EACL'06*.

Claire Nédellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of LLL'05*.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Suppl 3):S6.

Rune Sætre, Kenji Sagae, and Jun'ichi Tsujii. 2008. Syntactic features for protein-protein interaction extraction. In *Proceedings of LBM'07*, volume 319, pages 6.1–6.14.