An Analysis of Gene/Protein Associations at PubMed Scale

Sampo Pyysalo* **Tomoko Ohta*** **Jun'ichi Tsujii*†‡**
*Department of Computer Science, University of Tokyo, Tokyo, Japan
†School of Computer Science, University of Manchester, Manchester, UK
‡National Centre for Text Mining, University of Manchester, Manchester, UK
{smp,okap,tsujii}@is.s.u-tokyo.ac.jp

Abstract

Event extraction following the GENIA Event corpus and BioNLP'09 shared task models has been a considerable focus of recent work in biomedical information extraction. This work includes efforts applying event extraction methods to the entire PubMed, far beyond the narrow subdomains of biomedicine for which annotated resources are available. We aim to estimate the coverage of all statements of gene/protein associations in PubMed that existing resources for event extraction can provide. We base our analysis on a recently released corpus automatically annotated for gene/protein entities and syntactic analyses covering the entire PubMed, and use named entity co-occurrence, shortest dependency paths and an unlexicalized classifier to identify likely statements of gene/protein associations. A set of high-frequency/high-likelihood association statements are then manually analyzed with reference to the GENIA ontology. We provide a first estimate of the overall coverage of existing resources for event extraction and identify several classes of biologically significant associations of genes and proteins that are not addressed by these resources.

1 Introduction

In recent years, there has been a significant shift in focus in biomedical information extraction from simple pairwise relations representing associations such as protein-protein interactions (PPI) toward representations that capture typed, structured associations of arbitrary numbers of entities in specific roles, frequently termed *event extraction* (Ananiadou et al., 2010). Much of this work draws on the GENIA Event corpus (Kim et al., 2008), a resource of 1500 PubMed abstracts in the domain of *transcription factors in human blood cells* annotated for genes, proteins and related entities, events and syntax. This resource served also as the source for the annotations in the BioNLP'09 shared task on event extraction (BioNLP ST), the first collaborative evaluation of event extraction methods (Kim et al., 2009).

Another recent trend in the domain is a move toward application of extraction methods to the full scale of the existing literature, with results for various targets covering the entire PubMed literature database of nearly 20 million citations being made available (McIntosh and Curran, 2009; Björne et al., 2010b; Gerner et al., 2010a; Gerner et al., 2010b). As event extraction methods initially developed to target the set of events defined in the GENIA / BioNLP ST corpora are now being applied at PubMed scale, it makes sense to ask how much of the full spectrum of gene/protein associations found there they can maximally cover, a question separate from issues relating to their performance in extracting the targeted event types.

In this study, we seek to characterize the full range of gene/protein associations described in the literature and estimate what coverage of these associations state-of-the-art event extraction systems can maximally achieve. We approach these questions by assuming that associations are stated through specific words, analogously to the widely applied concepts of *interaction words* in protein-protein interaction extraction and *text binding* words in event extraction. We follow a statistical approach to identifying such candidate words using an automatically tagged corpus covering the entire PubMed literature database.

2 Task Definition

We term our extraction target *gene/protein associations*. So as not to limit the applicabil-

ity of our results, we define our target entities ("genes/proteins") broadly. The specific definition of this entity type is provided by the GENETAG corpus annotation (Tanabe et al., 2005) on which the applied automatic tagger is trained. GENETAG annotates a single class of gene/protein entities that encompasses genes and gene products as well as related entities such as domains, promoters, and complexes. This inclusiveness permits the identification of associations between more than only the gene and gene product entities included in the GENIA / BioNLP ST annotation (Ohta et al., 2009).

We also intend "associations" broadly, understanding it to encompass direct PPI-type interactions as well as experimental findings suggesting them, as pursued in the BioCreative PPI tasks (Krallinger et al., 2007), BioNLP-style events ("things that happen") such as *expression* and *localization*, as well as *static relations* in the sense of (Pyysalo et al., 2009), associations such as *part-of* that hold between entities without necessarily implying change. Indeed, while we take "association" to exclude properties and states that involve only a single entity, we do not set other specific constraints, following instead a loose biologically motivated definition that can be characterized informally as "any association between genes, gene products, or related entities that is of biological interest."

We note that while our aims and approach share a number of features with tasks such as protein-protein interaction extraction, they differ in focus on statements of association (as opposed to the entities stated to be associated) and in that we do not aim to find *instances* of the expressions of interest with high recall, but rather identify association *types*. Due to the large scale of the PubMed corpus it is possible to pursue an approach that only considers a small, high-reliability portion of the available data (discarding most instances) and still finds associations of interest. Thus, instead of instance-level recall, we pay particular attention to not introducing overt bias e.g. toward particular forms of expression so as to be able to use the result to estimate relative frequencies of the associations in the full corpus.

3 Corpus resources

This study is based on the 2009 distribution of the full PubMed literature database, encompassing approximately 18 million citations of biomedical domain scientific articles. For the analysis of this data, we make use of the Turku PubMed Scale (TPS) corpus (Björne et al., 2010b), an automatically annotated corpus covering the entire PubMed. Note that while the original focus of the corpus is on BioNLP-style events, we do not use these annotations. Instead, we make use of the automatically identified sentence boundaries, named entities, and the syntactic analyses, briefly presented in the following.

All PubMed documents in the TPS corpus were initially processed with the GENIA sentence splitter with simple heuristic post-processing to correct some errors from the machine learning-based splitter.[1] The sentence splitter is estimated to achieve an F-score of 99.7% on the GENIA corpus. Gene/protein named entities were tagged in all sentences using the BANNER named entity recognition system (Leaman and Gonzalez, 2008) trained on the GENETAG corpus and thus reflect its inclusive definition of gene/protein. The release of BANNER applied to tag the TPS corpus was reported to achieve 86.4% F-score on the GENETAG corpus, and an evaluation on a random sample of tagged entities in TPS data found 87% precision (Björne et al., 2010a), suggesting that the tagger generalizes well to the whole PubMed.

Finally, the TPS corpus distribution includes syntactic analyses for all sentences in which at least one named entity has been tagged.[2] Parses were produced using the McClosky-Charniak parser, a version of the Charniak-Johnson parser (Charniak and Johnson, 2005) adapted to the biomedical domain. The parser has shown state-of-the-art performance in recent intrinsic (McClosky and Charniak, 2008) and extrinsic (Miwa et al., 2010) evaluations. The McClosky-Charniak parser produces constituency (phrase structure) analyses in the Penn Treebank scheme, with Penn part-of-speech tags. In addition to the these analyses, dependency analyses in the Stanford Dependency (SD) scheme (de Marneffe and Manning, 2008), created from the constituency analyses by automatic conversion using the using the Stanford parser tools[3] are provided in the TPS corpus.

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/∼y-matsu/geniass/

[2] Sentences not containing entities are not parsed as parsing was the most computationally intensive part of the automatic corpus annotation and the system could only extract events from sentences with entities.

[3] http://nlp.stanford.edu/software/lex-parser.shtml

Word	Frequency
patients	8728330
cells	5384960
results	4175016
study	4149760
treatment	3436331
cell	3230831
activity	2763031
group	2635275
protein	2553732
effect	2457417

Table 1: Most frequent words in PubMed.

Word	Frequency
cells	1455897
protein	1057920
expression	923002
activity	753521
cell	750293
gene	704434
receptor	641766
human	635468
levels	603117
factor	518676

Table 2: Most frequent words in sentences containing two or more gene/protein entity mentions in PubMed.

4 Identification of Gene/Protein Associations

In this section, we present our approach to identifying statements of gene/protein associations through an extended analysis of word statistics in PubMed.

4.1 Overall Statistics

As expected for a corpus of English, the most frequent words in PubMed are prepositions, determiners, conjunctions and forms of the copula ("is") and, if non-word tokens are included, punctuation. In this work, we focus on content words, filtering closed class words and non-words and applying a basic stopword list including the PubMed stopword list. Table 1 shows the most frequent such words in PubMed.[4] The distribution suggests that medical topics dominate biomolecular ones overall, with e.g. the word "patients" occurring more than three times as often as the word "protein". Although general expressions such as "activity" and "effect" can be used to describe protein associations, the most frequent words contains no word specific to protein associations.

4.2 Gene/Protein Mentions

The automatic tagging for mentions of gene/protein-related named entities in the TPS corpus covers a total of 36.4 million gene/protein mentions in 5.4 million documents, approximately 30% of all PubMed citations. These annotations allow focus on texts likely relevant to gene/protein associations. Here, as we are interested in particular in texts describing associations between two or more gene/protein related entities, we apply a focused selection, picking only those individual sentences in which two or more mentions co-occur. While this excludes associations in which the entities occur in different sentences, their relative frequency is expected to be low: for example, in the BioNLP ST data, all event participants occurred within a single sentence in 95% of the targeted biomolecular event statements. In the TPS data, there are 9.0 million sentences with at least two tagged entities. These sentences contain 25.4 million entity mentions; approximately 70% of the total number.

Table 2 shows the most frequent words in sentences with at least two tagged protein mentions. The list suggests that this simple selection is sufficient to identify a subset of PubMed where biomolecular topics are prominent: both "protein" and "expression" appear ranked near the top.

4.3 Dependency Paths

The TPS corpus contains both constituency and dependency analyses of sentence syntax. While both forms of representation arguably capture largely the same information, dependency representations have been argued to make the relevant syntactic relations more immediately accessible and have been successfully employed in many recent domain information extraction approaches, frequently in conjunction with the use of the *shortest dependency path* between two entities to discover stated associations (see e.g. (Bunescu and Mooney, 2005; Fundel et al., 2007; Miwa et al., 2009; Björne et al., 2009)).

Here, we follow the assumption that when two entities are stated to be associated in some way, the most important words expressing their association will typically be found on the shortest dependency path connecting the two entities (cf. the *shortest path hypothesis* of (Bunescu and Mooney, 2005)) The specific dependency representation ap-

[4] For this and other word statistics in this section, basic tokenization separating punctuation from words and lowercasing has been applied but stemming or lemmatization is not performed.

Word	Frequency
expression	590810
activity	470393
levels	386130
cells	349648
activation	240942
induced	221177
binding	153806
mediated	129620
effect	124948
increased	124564

Table 3: Most frequent words on shortest dependency paths connecting two gene/protein entity mentions in PubMed.

Word	E_w
expression	68803.3
activity	56372.9
activation	43987.9
binding	28989.3
induced	24132.8
phosphorylation	22971.9
binds	17757.0
production	16893.2
inhibited	15972.9
inhibition	14546.0

Table 4: Words ranked highest by E_w, the expected number of times they occur on shortest paths likely to express a gene/protein association.

plied here is the collapsed, coordination-processed variant of the Stanford representation, which is expressly oriented toward use in this type of information extraction approaches (de Marneffe and Manning, 2008). When extracting the shortest paths, we further avoid traversing coordinating conjunction dependencies (`conj*`) to assure that relevant words are not excluded in sentences involving coordination and that similar paths are extracted for all coordinated words (Figure 1).

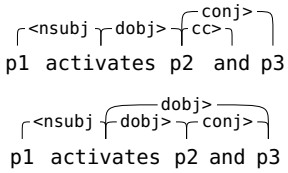

Figure 1: Variants of the SD representation. Basic SD above, collapsed, coordination-processed SD below.

The corpus contains 31.8 million pairs of gene/protein mentions co-occurring in a sentence, and a connecting shortest path could be extracted for 97% of these.[5] Table 3 shows the words most frequently occurring on these paths. This list again suggests an increased focus on words relating to gene/protein associations: *expression* is the most frequent word on the paths, and *binding* appears in the top-ranked words.

4.4 Path probabilities

Entities often co-occur in text without any association being stated between them, but some shortest dependency path can be found connecting (nearly) all co-occurring entities. Distinguishing paths that state associations from those that do not could thus help identify words that are key to expressing those associations.

A wealth of approaches for distinguishing relevant paths from irrelevant ones have been proposed in the protein-protein interaction extraction literature, including rule-based, pattern-based (hand-written and learned) and supervised classification-based methods (e.g. (Ding et al., 2003; Yakushiji et al., 2005; Rinaldi et al., 2006; Fundel et al., 2007; Sætre et al., 2007; Airola et al., 2008; Miwa et al., 2009)). However, writing explicit rules conflicts with our aim of discovering associations (and statements of associations) that we do not already know about, and application of standard supervised learning methods would similarly limit the scope of what can be extracted by the (known) training data.

Here, drawing on ideas from Open Information Extraction (Etzioni et al., 2008), we adopt a probabilistic approach using an "unlexicalized" machine learning method. We defer detailed description of the method to Section 5, now simply assuming a way to assign to each path p an (estimated) probability $P(p)$ that the path expresses an association between the entities it connects. We make use of $P(p)$ in two obvious ways to refine the pure frequency-based word rankings presented above: first, only count words when they occur on paths that have an estimated probability higher than a given threshold of being relevant, and second, replacing the "raw" word count with the expected number of times that word appears in a relevant path, informally $E_w = \sum_{p:w \in p} P(p)$.

Table 4 shows the top-ranked words by E_w as calculated using the method described below. The listing contains only words that are regularly used to express gene/protein associations, suggesting

[5] Failures to extract a path were primarily due to clause-level coordination (e.g. "we study P_1 and we find that P_1 is ...") and, rarely, failures from the parser or the dependency conversion.

that probabilistic ranking can allow clear focus on the targeted statements.

5 Machine Learning

We applied supervised machine learning to estimate the probability that a dependency path connecting gene/protein mentions expresses an association of these entities, training with "unlexicalized" features (Banko and Etzioni, 2008) to force the learning method to generalize and to learn based on the patterns of expression only.

5.1 Training Data

For training data, we could potentially draw from a wealth of corpus resources annotated for some form of association between genes/proteins, such as PPI corpora (see e.g. (Pyysalo et al., 2008)). However, as we are in particular interested in event extraction approaches, we chose to use the BioNLP ST data. This dataset also identifies the expressions stating the annotated events ("trigger words"), providing test material for the method.

As the BioNLP ST data does not explicitly identify *pairs* of entities that are stated to be associated, it was first necessary to derive a pairwise representation from the event representation. We applied a mapping similar to that introduced by (Heimonen et al., 2008) for deriving pairwise relations from the event-style annotations of the BioInfer corpus (Pyysalo et al., 2007): for each co-occurring entity pair, we identified all paths through event structures connecting the two entities. If these paths included at least one where the direction of causality was not reversed on the path, the pair was marked as a positive example of an association; otherwise it was marked negative. Finally, we interpreted the Equiv annotations identifying equivalent entity references in the data: any pair where entities are equivalent to those of at least one positive pair was marked positive (see Figure 2).

Finally, to make this pair data consistent with the TPS event spans, tokenization and other features, we aligned the entity annotations of the two corpora, mapping a BioNLP ST entity to a TPS entity if their spans matched or the source entity was entirely contained within the span of the candidate target entity. Unmatched entities were removed from the data. This processing was applied to the BioNLP ST training set, creating a corpus of 6889 entity pairs of which 1119 (16%) were marked as expressing an association (positive).

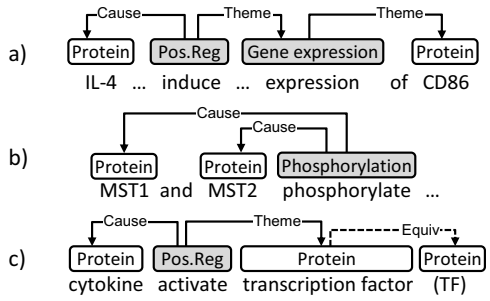

Figure 2: Reinterpreting BioNLP ST event structures as associated entity pairs. A positive pair is extracted for the proteins in a) but not in b) as they are not causally connected. In c), two positive pairs are extracted due to the equivalence relation.

5.2 Learning Method

We applied the libSVM Support Vector Machine implementation using probabilistic outputs (Chang and Lin, 2001). For training the classifier, we applied features derived only from the words and dependencies along the shortest path between any two entities. We first replaced each word marked as a gene/protein mention with a placeholder string and each other word with its part of speech tag, using the Penn tags included in TPS. We then derived a set of frequently used dependency path features from this representation (see e.g. (Airola et al., 2008; Van Landeghem et al., 2008; Miwa et al., 2009)): path length, path "tokens" (PoS/placeholder), dependency types on the path, and "token"/dependency 2-grams and 3-grams. Preliminary experiments using cross-validation on the training data suggested performance was not sensitive to the details of the feature representation. The SVM regularization parameter was selected similarly, testing parameter values on the scale $\ldots, 2^{-1}, 2^{0}, 2^{1}, \ldots$ and selecting $c = 2^{-3}$ for the final experiment.

The resulting classifier is intentionally weak, being trained to recognize not the specific properties of positive statements in its training set but rather their general characteristics. Development testing indicated an F-score and AUC of approximately 50% and 70%, substantially below the state of the art for the comparable PPI pair extraction task (Miwa et al., 2009).

5.3 Calculating E_w

E_w, informally characterized as the expected number of times a word w occurs on a dependency path which is estimated to be likely to express a

gene/protein association, is central to the applied probabilistic ranking. In technical detail, we derived E_w as follows.

We first extracted all instances of shortest dependency paths connecting two genes/proteins. We then combined all paths sharing the same "unlexicalized" representation, giving a total of 6.8 million unique paths. To make storage and processing more feasible, we removed paths occurring only once in the entire corpus. This filtered out 6.0 million paths – 88% of the total number of unique paths – but due to the Zipfian properties of the distribution, the remaining 0.8 million unique paths account for 16.7 million occurrences, or 74% of the total occurrences. We thus do not expect this practically motivated filtering to fundamentally alter the basic statistical properties of the data.

Each path was then assigned the estimated probability $P(p)$ using the probabilistic outputs of the SVM trained as described above. At this stage, we could potentially introduce a threshold parameter into the method defining a tradeoff between path quality and inclusiveness. However, as initial testing suggested the method to be relatively robust to the choice of cutoff, we simply take the obvious choice of defining as "likely positive" path any for which $P(p) > 0.5$. We then removed any path that did not meet this condition as not (likely) expressing an association, leaving 46437 unique unlexicalized paths (5.7% of the total) predicted to express gene/protein associations. Finally, each occurrence of a word w on one of these paths is assigned the path probability $P(p)$. In cases where words appear on multiple paths, they are simply assigned the maximum of the path probabilities. E_w is then the sum of these probabilities over the entire corpus.

6 Evaluation

We first evaluated each of the word rankings discussed in Section 4 by comparing the ranked lists of words against the set of single words marked as trigger expressions in the BioNLP ST development data. These single-word triggers account for 92% of all trigger expressions marked in the data, and there are 343 unique triggers. Figure 3 shows precision/recall curves for each of the four rankings generated by the word frequency/expected value. The result supports the informal observations made through the top-ranked words in Tables 1-4: the later approaches provide a much more relevant ranking for identifying words expressing associations.

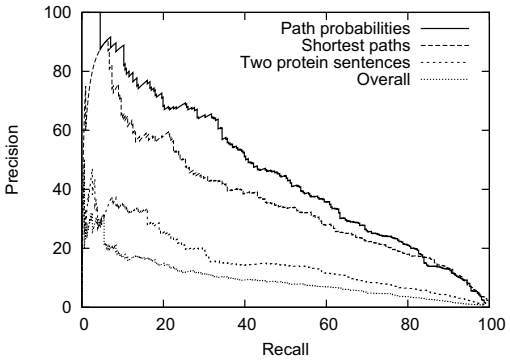

Figure 3: Precision/recall curves of the four word rankings.

We next performed a manual study of candidate words for stating gene/protein associations using the E_w ranking. Here, we take as *known* any word for which the normalized, lemmatized form[6] matches that of any word appearing as a trigger expression in the BioNLP ST training or development test data. We then selected the words ranked highest by E_w that were not known, grouped by normalized and lemmatized form, and added for reference examples of frequent shortest dependency paths on which any of these words appear. These groups were evaluated by a PhD biologist with expertise in event annotation and basic understanding of the Stanford Dependency representation of syntax (TO), with instructions to mark as positive words that in contexts like those provided can be understood to express a gene/protein association, defined broadly as described in Section 2.

In total, 1200 candidate expressions were manually evaluated, of which 660 were judged to express an association. We then proceeded to manually cluster them by the type of association they would typically express. Following preliminary analysis, we performed a top-level division into three categories: events ("things that happen") involving gene/protein entities in their natural environment (55% of associations), "static" relations holding between the entities (28%), and experimental observations and manipulations that do not occur naturally (17%). We further grouped the new event statements into event classes using the

[6]Using the NLM LVG `norm` normalizer, available at `http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/`

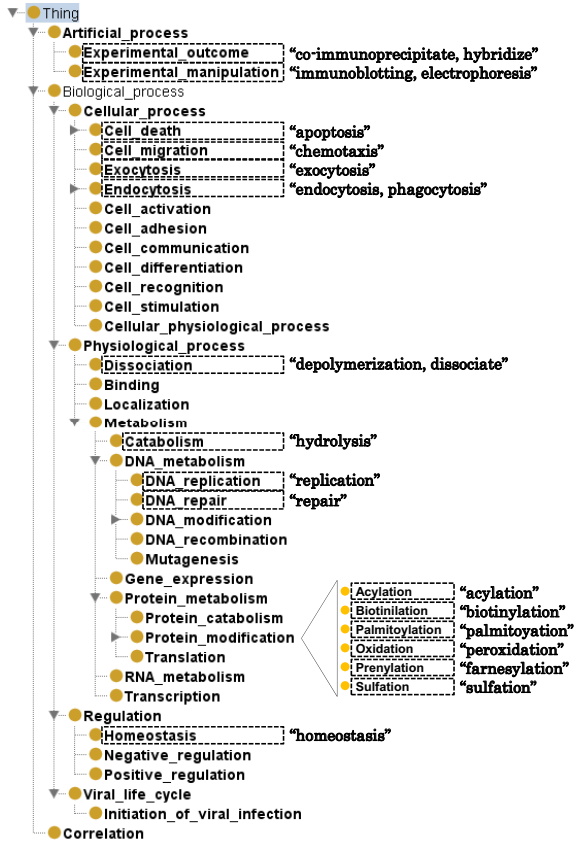

Figure 4: Organization of proposed new event classes into the GENIA ontology, with examples of expressions stating each type. New classes shown in dotted rectangles.

Gene Ontology (The Gene Ontology Consortium, 2000) for reference and identified event classes that were not previously included in the GENIA event ontology. This process suggested 18 event classes that were not previously considered in the GENIA ontology, shown in Figure 4 with a tentative proposal on how these classes could be organized into the GENIA ontology, with examples of identified words expressing each new event type.

Finally, to estimate the relative prominence of the known (i.e. BioNLP ST) expressions of associations in PubMed compared to those that were newly identified, we compared the E values of the unique lemmas, counted as the sum of E_w for words sharing the lemma. Figure 5 shows a plot of the values ranked from high to low E. The result was unexpected: the estimate suggests that even though the newly identified association words are drawn from PubMed without subdomain restrictions and include more than only event expressions, expressions of event-type associations us-

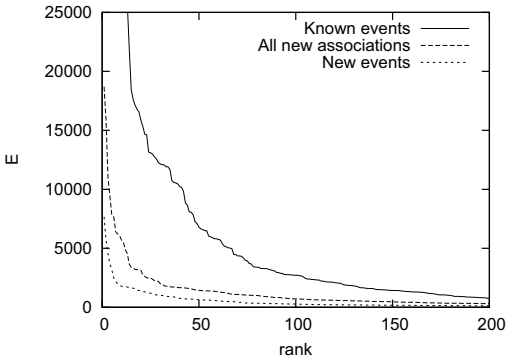

Figure 5: Comparison of estimated coverage of previously known and newly identified words expressing gene/protein associations. Note truncated ranges.

ing the previously known words are overall much more prominent in PubMed. Specifically, the total E value mass of all the newly identified associations (the area under the curve in Figure 5) is just 22% of that of the known events, and the mass of the newly identified events 37% of all the new associations; only 8% of that of the known events.

7 Discussion

We found that currently existing resources for event extraction are lacking in coverage of e.g. relatively rare but biologically important protein post-translational modifications and experimental outcomes that suggest (but do not state) causal connections. However, the statistical analysis suggests that resources already cover the clear majority of gene/protein events in PubMed, indicating that an annotation-based approach to extending coverage of event types (e.g. (Ohta et al., 2010)) may offer a realistic path to near-complete coverage of all major gene/protein events in the near future. With resources for static relation extraction (Pyysalo et al., 2009) this coverage could be further extended beyond event-type associations.

However, the approach to identifying gene/protein associations considered here is limited in a number of ways: it excludes associations stated across sentence boundaries, does not treat multi-word expressions as wholes, and only directly includes associations stated between exactly two entities. The approach is also fundamentally limited to associations expressed through specific words and thus blind to e.g. part-of relations implied by statements such as *CD14 Sp1-binding site*. Further, our estimate of

overall association statement frequency ignored the "long tail" of the distribution, thus excluding rare expressions which may nevertheless add up to a not insignificant fraction of the total. These factors limit the reliability of the presented coverage estimates. Finally, it should be noted that while we have taken any expression of association for which even a single annotation exists as "known", the performance at which many of these association can be extracted in practice may be limited.

8 Conclusions

We have presented an approach to discovering expressions of gene/protein associations from PubMed based on named entity co-occurrences, shortest dependency paths and an unlexicalized classifier to identify likely statements of gene/protein associations. Drawing on the automatically created full-PubMed annotations of the Turku PubMed-Scale (TPS) corpus and using the BioNLP'09 shared task data to define positive and negative examples of association statements, we distilled an initial set of over 30 million protein mentions into a set of 46,000 unique unlexicalized paths estimated likely to express gene/protein associations. These paths were then used to rank all words in PubMed by the expected number of times they are predicted to express such associations, and 1200 candidate association-expressing words not appearing in the BioNLP'09 shared task data evaluated manually. The study of these candidates suggested 18 new event classes for the GENIA ontology and indicated that the majority of statements of gene/protein associations not covered by currently available resources are not statements of biomolecular events but rather statements of static relations or experimental manipulation.

The event annotation of the GENIA corpus was originally designed to cover events discussed in publications on a limited subdomain of biomolecular science. It could thus be assumed that the event types and the specific statements annotated in GENIA would have only modest coverage of all gene/protein association types and statements in PubMed. However, our results suggest that even the BioNLP'09 shared task data, a subset of GENIA, may represent a clear majority of all gene/protein associations. However, this estimate of coverage is a first attempt and involves many uncertain factors and potential sources of error, calling for more research.

The data derived from TPS created in this study, including the shortest paths, their estimated probabilities, and the word lists ranked by probability of stating a gene/protein association are available for research purposes from from the GENIA project homepage `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA`.

Acknowledgments

This work was supported by Grant-in-Aid for Specially Promoted Research (MEXT, Japan).

References

Antti Airola, Sampo Pyysalo, Jari Bjorne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics*, 9(Suppl 11):S2.

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B. Kell. 2010. Event extraction for systems biology by text mining the literature. *Trends in Biotechnology*.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. In *Proceedings of ACL-08: HLT*, pages 28–36.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Shared Task*, pages 10–18.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010a. Complex event extraction at PubMed scale. *Bioinformatics*, 26(12):i382–390.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010b. Scaling up biomedical event extraction to the entire pubmed. In *Proceedings of BioNLP'10*, pages 28–36.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 724–731.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL'05*, pages 173–180.

Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. 2003. Extracting biochemical interactions from MEDLINE using a link grammar parser. In *Proceedings of ICTAI'03*, pages 467–471.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM*, 51(12):68–74.

Katrin Fundel, Robert Kuffner, and Ralf Zimmer. 2007. RelEx–Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

Martin Gerner, Goran Nenadic, and Casey Bergman. 2010a. Linnaeus: A species name identification system for biomedical literature. *BMC Bioinformatics*, 11(1):85.

Martin Gerner, Goran Nenadic, and Casey M. Bergman. 2010b. An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In *Proceedings of BioNLP'10*, pages 72–80.

Juho Heimonen, Sampo Pyysalo, Filip Ginter, and Tapio Salakoski. 2008. Complex-to-pairwise mapping of biological relationships using a semantic network representation. In *Proceedings of SMBM'08*.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(10).

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP 2009 Shared Task*.

Martin Krallinger, Florian Leitner, and Alfonso Valencia. 2007. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of BioCreative II*, pages 41–54.

Robert Leaman and Garciela Gonzalez. 2008. Banner: An executable survey of advances in biomedical named entity recognition. In *Proceedings of PSB'08*, pages 652–663.

Marie-Catherine de Marneffe and Christopher Manning. 2008. The stanford typed dependencies representation. In *COLING Workshop on Cross-framework and Cross-domain Parser Evaluation*.

David McClosky and Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of ACL-HLT'08*, pages 101–104.

Tara McIntosh and James R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of ACL/IJCNLP'09*, pages 396–404.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12):e39–e46.

Makoto Miwa, Sampo Pyysalo, Tadayoshi Hara, and Jun'ichi Tsujii. 2010. A comparative study of syntactic parsers for event extraction. In *Proceedings of BioNLP'10*, pages 37–45.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009. Incorporating GENETAG-style annotation to GENIA corpus. In *Proceedings of BioNLP'09*, pages 106–107.

Tomoko Ohta, Sampo Pyysalo, Makoto Miwa, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event extraction for post-translational modifications. In *Proceedings of BioNLP'10*, pages 19–27.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

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

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of BioNLP'09*, pages 1–9.

Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. 2006. An environment for relation mining over richly annotated corpora: The case of GENIA. In *Proceedings of SMBM'06*, pages 68–75.

Rune Sætre, Kenji Sagae, and Jun'ichi Tsujii. 2007. Syntactic features for protein-protein interaction extraction. In *Proceedings of LBM'07*, pages 6.1–6.14.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl. 1):S3.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29.

Sofie Van Landeghem, Yvan Saeys, Bernard De Baets, and Yves Van de Peer. 2008. Extracting protein-protein interactions from text using rich feature vectors and feature selection. In *Proceedings of SMBM'08*.

Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, and Jun'ichi Tsujii. 2005. Biomedical information extraction with predicate-argument structure patterns. In *Proceedings of SMBM'05*, pages 60–69.