Static Relations: a Piece in the Biomedical Information Extraction Puzzle

Sampo Pyysalo* **Tomoko Ohta*** **Jin-Dong Kim*** **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,jdkim,tsujii}@is.s.u-tokyo.ac.jp

Abstract

We propose a static relation extraction task to complement biomedical information extraction approaches. We argue that static relations such as part-whole are implicitly involved in many common extraction settings, define a task setting making them explicit, and discuss their integration into previously proposed tasks and extraction methods. We further identify a specific static relation extraction task motivated by the BioNLP'09 shared task on event extraction, introduce an annotated corpus for the task, and demonstrate the feasibility of the task by experiments showing that the defined relations can be reliably extracted. The task setting and corpus can serve to support several forms of domain information extraction.

1 Introduction

Relation Extraction (RE) is a key task in biomedical Information Extraction (IE). The automatic detection of relevant types of relations — for various definitions of relevant — between entities has been one of the primary focus points for significant domain research efforts over the past decade, and a substantial number of biomedical RE methods and annotated corpora have been published (Zweigenbaum et al., 2007). Motivated by the needs of biologists and e.g. database curation efforts, most domain RE efforts target relations involving biologically relevant changes in the involved entities, commonly to the complete exclusion of static relations. However, static relations such as entity membership in a family and one entity being a part of another are not only relevant IE targets in themselves but can also play an important supporting role in IE systems not primarily targeting them.

In this paper, we investigate the role of static relations in causal RE and event extraction. Here, we use *relation extraction* in the MUC and ACE (Sundheim, 1995; Doddington et al., 2004) sense to refer to the task of extracting binary relations, ordered pairs of entities, where both participating entities must be specified and their roles (agent, patient, etc.) are fixed by the relation. By contrast, *event extraction* is understood to involve events (things that happen) and representations where the number and roles of participants may vary more freely. We refer to relations where one one entity causes another to change as *causal relations*; typical domain examples are *phosphorylation* and *activation*. Static relations, by contrast, hold between two entities without implication of change or causality: examples from the ACE IE task include Physical.Located and Part-Whole.Artifact.

2 Task definition

In the following, we argue that static relations are relevant to much of current biomedical IE work, present a task setting making these relations explicit, and discuss applications of static relation annotation and extraction methods.

2.1 Named entity-driven IE and static relations

Named entities (NEs) provide a simple anchor connecting text to entities in the real world and thus a natural starting point for IE. Named entity recognition (NER) is well studied and several biomed-

Proceedings of the Workshop on BioNLP, pages 1–9,
Boulder, Colorado, June 2009.

ical NER systems are available (see e.g. (Wilbur et al., 2007; Leaman and Gonzalez, 2008)), and most domain IE approaches are NE-driven: a typical way to cast the RE task is as deciding for each pair of co-occurring NEs whether a relevant relation is stated for them in context. Like the previous LLL and BioCreative2-PPI relation extraction tasks (Nédellec, 2005; Krallinger et al., 2007), the BioNLP'09 shared task on event extraction (Kim et al., 2009) similarly proceeds from NEs, requiring participants to detect events and determine the roles given NEs play in them.

Any domain IE approach targeting nontrivial causal NE relations or events necessarily involves decisions relating to static relations. Consider, for example, the decision whether to extract a relation between NE_1 and NE_2 in the following cases (*affects* should here be understood as a placeholder for any relevant statement of causal relation):

1) NE_1 *affects* NE_2 *gene*
2) NE_1 *affects* NE_2 *promoter*
3) NE_1 *affects* NE_2 *mutant*
4) NE_1 *affects* NE_2 *antibody*
5) NE_1 *affects* NE_2 *activator*

The decision here depends on the interpretation of the noun compounds (NCs) NE_2 *gene*, NE_2 *promoter*, etc. Depending on the IE setting, one might, for example, judge that statements (1)–(3) justify the extraction of an (NE_1, NE_2) relation, while (4) and (5) do not. This question is rarely formalized as a separate (sub)task in domain studies, and methods targeting e.g. the LLL, BioCreative2-PPI and BioNLP'09 shared task relations and events must learn to resolve this question together with the separate issue of which words and syntactic structures express relevant causal relations.

2.2 Task setting

The relation extraction problems represented by examples (1)–(5) above are closely related to the well-studied issue of NC semantics. However, the problem extends past simple binary NCs to include judgments on the relations of arbitrary base NPs (nouns with premodifiers) to contained NEs,

NE_1 *affects truncated* NE_2

NE_1 *affects* NE_2/NE_3 *complexes*

NE_1 *affects* NE_2*-dependent phosphatase*

and further to relations of NPs with NEs that are syntactically less immediately attached:

NE_1 *affects first exon of* NE_2

NE_1 *affects an element in the* NE_2 *promoter*

NE_1 *affects members of the immediate-early activation genes family such as* NE_2

The problem thus encompasses also more general relations between nominals.

While these different cases could also be studied as separate tasks, in the current IE context they can be seen as presenting a continuum of different syntactic realizations of similar relations that also carry the same implications for further processing. We propose to treat them together, formulating the specific task studied in this paper as follows:

> Given: named entity NE and another entity E with their context in text,
> Determine: whether there is a relevant static relation R(NE, E) and its type.

Here, relevant relations are defined as those that justify an inference of some role for the NE in causal relations/events involving E. Additionally, the level of granularity chosen for typing is chosen according to the need to determine the role of the NE in the relations/events. These choices are intentionally dependent on the IE context: we do not expect to be able to formulate a universally accepted set of relevance criteria or relations. Our choice of relation scope and types here follows the perspective of a currently highly relevant IE problem, the BioNLP'09 shared task on event extraction. We aim to recognize a set of relations sufficient to capture the relevant relationships of the NEs provided as given information in the shared task (all of protein/gene/RNA type) and the terms annotated in the GENIA Event corpus (Kim et al., 2008) as participants in events.

We note that this task setting excludes the recognition of candidate NEs and other entities. The assumption that they are given is analogous to the common NE-NE causal relation extraction setting. Further, requiring their recognition would, in our view, unnecessarily complicate the task with aspects of NER and NP chunking, well-studied separate tasks.

We next sketch a formulation of an causal relation/event extraction task incorporating static relations and briefly present one possible way in which

static relation extraction could be applied in IE settings not explicitly targeting such relations.

2.3 Applications of static relations

In the following, we assume that NEs are detected in a prior processing step. Consider, then, the task of extracting relevant information from the following sentence:

NE_1 *is a subunit of the complex that inhibits the expression of mutant forms of* NE_2

An example causal relation extraction target here could be

Inhibit(NE_1,NE_2)

while an event extraction task might aim to recognize the events

E1:Expression(NE_2)
E2:Inhibit(NE_1, E1)

An IE system directly targeting either representation will need to simultaneously address issues relating to the causal statements and static relations. Static relation annotation makes this explicit (square brackets are used to mark non-NE entities):

Part-Whole.Component-Object(NE_1, [complex])
Variant(NE_2, [mutant forms])

This type of static relation detection as prior step to causal relation or event extraction could be applied in at least two different ways: primarily *augmenting* the extracted information, or alternatively *assisting* in the extraction of the information considered above. Assuming the successful extraction of the above static relations, the input can be reformulated as

NE_1 *is a subunit of the [complex] that inhibits the expression of [mutant forms] of* NE_2

Then, under the augmented extraction model, the causal relation and event extraction targets would be, respectively,

Inhibit([complex],[mutant forms])

and

E1:Expression([mutant forms])
E2:Inhibit([complex], E1)

Taken together with the static relations, this provides a more detailed representation of the information stated in the example sentence. Further, simple rules would suffice to derive the simplified representations involving only the NEs, and such rules would have the further benefit of making explicit which intervening static relations are taken to support the inference that an NE is involved in a stated causal relation or event.

Alternatively, under the assisted extraction model, with the assumption that the static relations are taken to allow the inference that any relation or event holding of the other entities holds for the NEs, the input to the causal relation or event extraction system can be recast as

NE_1 *is a subunit of the* NE'_1 *that inhibits the expression of* NE'_2 *of* NE_2

where NE'_1 and NE'_2 should be understood as aliases for NE_1 and NE_2, respectively. Now, under the causal relation extraction model, each of the (NE_1,NE_2), (NE'_1, NE_2), (NE_1,NE'_2), (NE'_1,NE'_2) pairs can serve as an example of the desired relation, both for the purposes of training and actual extraction (the event extraction case can be treated analogously). By increasing the number of positive cases, this application of information on static relations would be expected to have a positive effect on the performance of the primary causal relation/event extraction method.

While these two alternatives are only rough sketches of possible uses of static relation annotation, we expect either could be developed into a practical implementation. Further, these examples by no means exhaust the possibilities of this class of annotation. As static relation extraction can thus be seen to have multiple potential benefits for both causal relation and event extraction, we believe the efforts to pursue static relations as a separate task and to develop resources specific to this task are justified.

3 Relations

Based on an analysis of the shared task data (see Section 4.1), we recognize the static relations illustrated in Table 1. In the following, we briefly discuss the types and their selection.

Name	Examples
Variant	Bcl-6 gene, IL-1 mRNA, wild-type SHP1, TRADD mutant, human IL-1beta, [cell-surface isoforms] of CD43, phosphorylated CREB protein
PW.Object-Component	IL-6 promoter, GR N-terminal transactivation domain, SAA promoter sequence, proximal IL-2 promoter-enhancer, [transcriptional enhancers] including IFNB
PW.Component-Object	NF-kappa_B1/RelA heterodimer, p65 homodimer, p50-p65 complex, STAT1-containing [DNA-binding complex], [heterodimer] of p50 and p65
PW.Member-Collection	CREB/ATF family, p21ras small GTP binding proteins, [non-heat shock genes] such as IL1B, [cellular genes] including GM-CSF
PW.Place-Area	beta-globin locus

Table 1: Relations. In examples, NEs are underlined and square brackets are used to mark the extent of non-NE entities that do not span the entire example text.

3.1 Selection criteria

Relations could be recognized and split into different types at a number of different granularities. Motivated by practical IE applications, we aimed to define a static relation extraction subtask that fits naturally into existing IE frameworks and to create annotation that supplements existing annotation and avoids overlap in annotated information. The practical goals also motivate our aim to recognize a minimal set of different relation types that can satisfy other goals, fewer distinctions implying an easier task and more reliable extraction.

To decide whether to use a single relation type or introduce several subtypes to annotate a given set of cases, we aimed to introduce coherent relation types, each implying consistent further processing. More specifically, we required that each relation R(*NE*, entity) must uniquely and consistently define the relation and roles of the participants, and that in the relevant IE context the relation alone is sufficient to decide how to interpret the role of the *NE* in other relations/events. Specific examples are given in the introduction of the chosen relation types below.

In the following, we follow in part the relation taxonomy and relation definitions of (Winston et al., 1987). However, we recognize that there is no clear agreement on how to subdivide these relations and do not suggest this to be the only appropriate choice.

3.2 Part-whole relations

Part-whole, or meronymic, relations are, not surprisingly, the most common class of static relations in our data: a single generic Part-Whole relation could capture more than half of the relevant relations in the corpus. However, although the relations between the NE and entity in, for example, *[complex] containing* NE and *[site] in* NE are both types of Part-Whole (below PW) relations, the roles of participants are not consistently defined: in PW(*NE*, [site]) the entity is a component of the *NE*, while in PW(*NE*, [complex]) the roles are reversed. We thus recognize separate PW.Object-Component and PW.Component-Object relations. By contrast, while the relation between a *NE* representing a gene and a site on that gene is is arguably different from the relation between a protein *NE* and a site on the protein, we do not distinguish these relations as the annotation would duplicate information available in as part of the entity typing in the corpus and would further imply a static relation extraction task that incorporates aspects of NE recognition.

Also frequent in the data are relations such as that between a protein and a protein family it belongs to. While many cases are clearly identifiable as PW.Member-Collection relations, others could alternatively be analysed as Class-Member. As in our context the relations in e.g. *P, a member of the [type F protein family]* and *P, a [type F protein]* imply the same processing, we will apply the PW.Member-Collection label to both, as well as to *ad hoc* collections such as *[cellular genes] such as* NE, even if this requires a somewhat relaxed interpretation of the relation label. Finally, there are a few cases in our data (e.g. NE *locus*) that we view as instances of the PW.Place-Area relation.

3.3 Variant relations

To avoid unnecessary division of relations that imply in our context similar interpretation and processing, we define a task-specific Variant relation that

encompasses a set of possible relation types holding between an NE and its variants along multiple different axes. One significant class of cases annotated as Variant includes expressions such as NE *gene* and NE *protein*, under the interpretation that *NE* refers to the abstract information that is "realized" as either DNA, RNA or protein form, and the entity to one of these realizations (for alternative interpretations, see e.g. (Rosario and Hearst, 2001; Heimonen et al., 2008)).

The Variant relation is also used to annotate *NE*-entity relations where the entity expresses a different state of the *NE*, such as a phosphorylated or mutated state. While each possible post-translational modification, for example, could alternatively be assigned a specific relation type, in the present IE context these would only increase the difficulty of the task without increasing the applicability of the resulting annotation.

3.4 Other/Out annotation

We apply a catch-all category, Other/Out, for annotating candidate (NE, entity) pairs between which there is no relevant static relation. This label is thus applied to a number of quite different cases: causal relations, both implied (e.g. NE *receptors*, NE *response element*) and explicitly stated (NE *binds the [site]*), relations where the entity is considered too far removed from the NE to support reliable inference of a role for the NE in causal relations/events involving the entity (e.g. *[antibodies] for* NE), and cases where no relation is stated (e.g. NE *and other [proteins]*). The diversity of this generic category of irrelevant cases is a necessary consequence of the aim to avoid annotation involving decisions directly relating to other tasks by creating distinctions between e.g. *causal* and *no relation*.

3.5 Sufficiency of the setting and relation types

We have cast the static relation extraction task as always involving an NE, which in the present context is further always of a protein, gene or RNA type. This restriction considerably simplifies the task conceptually and reduces annotation effort as well as expected extraction difficulty, as the type of only one of the entities involved in the relation can vary significantly. However, it is not obvious that the restriction allows coherent relations types to be defined. If the corpus contained frequent cases where the stated relationship of the NE to the entity involved different types of relevant relations (e.g. collections of parts of an NE), it would be necessary to either recognized "mixed" or combined relations or extend the task to include general entity-entity relations.

Interestingly, during annotation we encountered only two cases (less than 0.1% of those annotated) involving two of the recognized relation types at once: *mutant* NE *promoter* and *5' truncation mutants of the* NE *promoter*[1]. While this result is likely affected by a number of complex factors (annotation criteria, NE and entity types, granularity of relations, etc.), we find the outcome — which was neither planned for nor forced on the data — a very encouraging sign of the sufficiency of the task setting for this and related domain IE tasks.

4 Data

We created the data set by building on the annotation of the GENIA Event corpus (Kim et al., 2008), making use of the rich set of annotations already contained in the corpus: term annotation for NEs and other entities (Ohta et al., 2002), annotation of events between these terms, and treebank structure closely following the Penn Treebank scheme (Tateisi et al., 2005).

4.1 Annotation

The existing GENIA annotations served as the basis of the new annotation. We initially selected as candidates entities annotated as participating in events considered in the BioNLP'09 shared task.

As the term annotation includes nesting of entities, NEs contained within these relevant entities were used as the starting point for the annotation. We first performed a preliminary study of the relevant static relations occurring between the entities and NEs occurring within them to determine the set of relations to annotate. Next, all unique cases where a selected entity contained an NE were annotated with the appropriate relation based on the contained text of the entity, with the text of the contained NE normalized away. For the present study, we excluded from consideration cases where the annota-

[1] To resolve these cases, we simply ignored the implied Variant relation.

tion indicated simple aliasing (e.g. *[CREB/ATF]*), a relation irrelevant to our purpose and found in the selected data only due to the annotation specifying one entity but two NEs in these cases. In this step, 830 unique cases representing a total of 1601 entities containing NEs were annotated.

The nesting structure of the term annotation does not, however, capture all relevant static relations: the term annotation scheme disallows discontinuous terms and annotation of terms with structure more complex than base NPs. Thus, the possible relations of NEs to entities to which they were connected e.g. by a prepositional phrase cannot be directly derived from the existing annotation. As an example, the nesting in *[*NE *region]* directly suggest the existence of a relation, while no such connection appears in *[region] of* NE. To annotate relations for entities for which the term annotation does not identify a candidate related NE, it is necessary to form (NE, entity) pairs with co-occurring NEs. Even when the candidate NEs were restricted to those occurring in the same sentence, the number of such pairs in the corpus was over 17,000, beyond the scope of what could be annotated as part of this effort. Further, as the great majority of co-occurring (NE, entity) pairs will have no relevant static relation, we used heuristics to increase the proportion of relevant and near-miss cases in the annotated data.

We first converted the gold standard annotation of the GENIA treebank (Tateisi et al., 2005) into a dependency representation using the Stanford parser tools (de Marneffe et al., 2006) and then determined the shortest paths in the dependency analyses connecting each relevant entity with each NE. The (NE, entity) pairs were then ordered according to the length of these paths, on the assumption that syntactically more closely related entities are more likely to have a relevant static relation. Annotation then proceeded on the ordered list of pairs. During the annotation, we further developed more ordering heuristics, such as giving higher ranking to candidate pairs connected by a path that contains a subpath known to connect pairs with relevant relations. Such known paths were first derived from the BioInfer static relation annotation (Pyysalo et al., 2007) and later extracted from previously annotated cases. In this annotation process, judgments were performed with reference to the full sentence context. In total, 1100 cases were annotated in this way. All stages of the annotation process involved only lists formatted as simple text files for markup and custom-written software for processing.

	Annotated instances		
Relation	cont.	nonc.	total
PW.Object-Component	394	133	527
PW.Component-Object	299	44	343
Variant	253	20	273
PW.Member-Collection	25	124	149
PW.Place-Area	4	1	5
Other/Out	626	778	1404
total	1601	1100	2701

Table 2: Statistics for annotated data. Number of instances given separately for relations annotated between entities with contained (cont.) and non-contained (nonc.) NEs.

Table 2 contains statistics for the annotated data, showing separately the number of annotated relations of entities to contained and non-contained NEs. There are interesting differences in the relation type distribution between these two categories, reflecting the different ways in which relations are typically stated. This difference in distribution suggests that it may be beneficial to give the two cases different treatment in extraction.

4.2 Representation

For simplicitly of use, we provide the annotated data in two equivalent representations: a simple inline XML format and a standoff format. The XML format closely resembles the representation used for the SemEval-2007 Semantic Relations between Nominals task (Girju et al., 2007). Here, each NE-Entity pair is given its own entry with its sentence context in which only the pair is marked. In the alternate standoff representation, all entities appearing in each sentence are tagged, and the annotated relations given separately. These representations are easily processed and should be usable with little modification with many existing relation extraction methods.

We further split the data into training, development-test and test sets according to the same division applied in the BioNLP'09 shared task on event extraction. This division allows the dataset to be easily integrated into settings using the shared task data, combining static relation and event extraction approaches.

5 Experiments

The selected task setting and representation form a natural basis for two alternative classification problems: a binary classification problem for detecting the presence of any relevant relation, and a multiclass classification problem where the correct relation type must also be determined. In the following, we describe experiments using the dataset in these two settings. While we apply a state-of-the-art machine learning method and a fairly expressive representation, the aim of the experiments is only to determine the relative difficulty of the relation extraction task and to establish a moderately competitive baseline result for the newly created dataset.

We use a linear Support Vector Machine (SVM) classifier (Chang and Lin, 2001) with N-gram features defined over token sequences delimited by the beginning and end of the entity and the position of the NE. The NE is treated as a single token and its text content blinded from the classifier to avoid overfitting on specific names. Features are generated from two sequences of tokens: those inside the entity and, when the NE is not contained in the entity, those between the entity and the NE (inclusive of the entity and NE at the sequence boundaries). In preliminary experiments on the development test set we found no clear benefit from including N-gram features extracted from a broader context, supporting an assumption that the problem can be mostly addressed on the basis of local features. By contrast, preliminary experiments supported the use of the simple Porter algorithm (Porter, 1980) for stemming, the inclusion of uni-, bi- and trigram features, and normalization of the feature vectors to unit length; these were adopted for the final experiment. The SVM regularization parameter was optimized using a sparse search with evaluation on the development test set.

We first reduced the annotated data into a binary classification problem with the Other/Out class representing negative (irrelevant) and the other relations positive (relevant) cases. The results for this experiment were very encouraging, giving both a high classification accuracy of 86.8% and an F-score of 84.1%. The test set contains 179 positive and 269 negative cases, giving a majority baseline accuracy of 60.0% and an all-true baseline F-score of 57.1%. The classifier notably and statistically significantly (McNemar's test, $p < 0.01$) outperforms these simple baselines. We then performed a separate multiclass classification experiment, predicting the specific type of the relation, also including the Other/Out type. In this experiment, accuracy remained relatively high at 81.9%, while per-class precision and recall results (considering each class in turn positive and all others negative, see Table 3) indicate some remaining challenges. The results vary somewhat predictably with the number of examples per relation type (Table 2): while PW.Object-Component relations can be predicted at high precision and fair recall, performance for PW.Member-Collection relations falls behind expectations for a local relation extraction problem.

	P	R	F
Relevant	81.2	87.2	84.1
PW.Object-Component	94.2	75.4	83.8
PW.Component-Object	60.0	71.2	65.1
Variant	88.0	57.9	69.8
PW.Member-Collection	54.5	37.5	44.4

Table 3: Classification results with (P)recision, (R)ecall and (F)-score for the binary Relevant/Irrelevant experiment and classwise results for the relevant classes (PW.Place-Area excluded for lack of data).

To briefly relate these results to domain causal RE results, we note that the recently proposed state-of-the-art method of (Airola et al., 2008) was reported to achieve F-scores ranging between 56.4–76.8% on five different causal RE corpora in a binary classification setting. As our relatively simple method achieves a notably higher 84.1% F-score at the binary static RE task, we can conclude that this static RE task is not as difficult as the causal RE tasks. This is encouraging for the prospects of static RE in support of domain causal RE and event extraction.

6 Related work

Relations of types that we have here termed static have figured prominently in the MUC and ACE series of events that have largely defined the "general domain" IE research program (Sundheim, 1995; Doddington et al., 2004). In this line of research, event-type annotation is used (as the name implies) to capture events, defined as "*[...] something that happens [...] [that] can frequently be described as a*

change of state" (LDC, 2005) and relation-type annotation is applied for relevant non-causal relationships. General static relations have been studied extensively also in broader, non-IE contexts (see e.g. (Girju et al., 2007)).

In the biomedical domain, static relations have received relatively little attention. Domain noun compound semantics, including static relations, have been considered in studies by (Rosario and Hearst, 2001) and (Nakov et al., 2005), but in IE settings static relations tend to appear only implicitly, as in the RelEx causal RE system of (Fundel et al., 2007), or through the causal relations they imply: for example, in the AIMed corpus (Bunescu et al., 2005) statements such as NE_1/NE_2 *complex* are annotated as a binding relation between the two NEs, not Part-Whole relations with the broader entity. By contrast, there has been considerable focus on the extraction of "things that happen," dominantly making use of relation-type corpus annotation and extraction approaches: a study of five corpora containing primarily causal relation annotation is found in (Pyysalo et al., 2008); more complete lists of domain corpora are maintained by Kevin Cohen[2] and Jörg Hakenberg[3]. For a thorough review of recent work in domain RE, we refer to (Zweigenbaum et al., 2007).

BioInfer (Pyysalo et al., 2007), to the best of our knowledge the first domain corpus to include event-type annotation, also includes annotation for a set of static relation types. The design of the BioInfer corpus and relationship type ontology as well as work applying the corpus in jointly targeting event extraction and static relation extraction (Heimonen et al., 2008; Björne et al., 2008) have considerably influenced the present study. A key difference in focus is that BioInfer primarily targets NE-NE relations, while our concern here has been the relations of NEs with other, non-NE entities, specifically focusing on the requirements of the BioNLP'09 shared task. A class of static relations, connecting Mutants and Fragments with their parent proteins, is annotated in the recently introduced ITI TXM corpora (Alex et al., 2008). While somewhat limited in the scope of static relations, this annotation covers an extensive number of instances, over 20,000, and could likely support the development of high-reliability methods for the extraction extraction of these specific static relations. As discussed in detail in Section 4.1, previously published versions of the GENIA corpus (Kim et al., 2008) contain NE, term and event annotation, but no static relations have been annotated in GENIA prior to this effort.

While previously introduced corpora thus cover aspects of the annotation required to address the static relation extraction task considered in this paper, we are not aware of previously published resources that would address this task specifically or contain annotation supporting the entire task as envisioned here.

7 Conclusions and future work

In this paper, we have argued for a position for static relations in biomedical domain IE, specifically advancing the subtask of extracting static relations between named entities and other entities appearing in their context. We explored this subtask in the specific IE context of the BioNLP'09 shared task on event extraction, identifying possible instances of static relations relevant to the task setting. We then studied these instances of detail, defining a minimal set of basic static relations argued to be sufficient to support the type of IE envisioned in the shared task. We annotated 2701 instances of candidate static relations, creating the first domain corpus of static relations explicitly designed to support IE, and performed experiments demonstrating that the static relation extraction task can be performed accurately, yet retains challenges for future work. The newly annotated corpus is publicly available at `www-tsujii.is.s.u-tokyo.ac.jp/GENIA` to encourage further research on this task.

Acknowledgments

Discussions with members of the BioInfer group were central for developing many of the ideas presented here. We are grateful for the efforts of Maki Niihori in producing supporting annotation applied in this work. This work was partially supported by Grant-in-Aid for Specially Promoted Research (Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan), and Genome Network Project (MEXT, Japan).

[2] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml

[3] http://www2.informatik.hu-berlin.de/∼hakenber/links/benchmarks.html

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.

Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of LREC'08*.

Jari Björne, Sampo Pyysalo, Filip Ginter, and Tapio Salakoski. 2008. How complex are complex protein-protein interactions? In *Proceedings SMBM'08*.

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. *Artificial Intelligence in Medicine*, 33(2):139–155.

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

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program: Tasks, data, and evaluation. In *Proceedings of LREC'04*, pages 837–840.

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

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of SemEval'07*, pages 13–18.

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'09*.

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.

LDC. 2005. ACE (automatic content extraction) english annotation guidelines for events. Technical report, Linguistic Data Consortium.

R. Leaman and G. 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, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC'06*, pages 449–454.

Preslav Nakov, Ariel Schwartz, Brian Wolf, and Marti Hearst. 2005. Scaling up bionlp: Application of a text annotation architecture to noun compound bracketing. In *Proceedings of BioLINK'05*.

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

Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT'02)*, pages 73–77.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(2):130–137.

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.

Barbara Rosario and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of EMLNP'01*, pages 82–90.

Beth M. Sundheim. 1995. Overview of results of the MUC-6 evaluation. In *Proceedings of MUC-6*, pages 13–31.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the GENIA corpus. In *Proceedings of IJCNLP'05*, pages 222–227.

John Wilbur, Larry Smith, and Lorrie Tanabe. 2007. Biocreative 2 gene mention task. In *Proceedings of BioCreative 2*, pages 7–16.

Morton E. Winston, Roger Chaffin, and Douglas Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive Science*, 11.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. 2007. Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*.