A Re-evaluation of Biomedical Named Entity - Term Relations

Tomoko Ohta* **Sampo Pyysalo*** **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
{okap,smp,jdkim,tsujii}@is.s.u-tokyo.ac.jp

Abstract

Recent developments in biomedical text mining include advances at the reliability of named entity recognition as well as movement toward richer representations of the associations of named entities. We argue that this shift in representation should be accompanied by the adoption of a a more detailed model of the relations holding between named entities and other relevant domain terms. As a step toward this goal, we study named entity - term relations with the aim of defining a detailed, broadly applicable set of relation types based on accepted domain standard concepts for use in corpus annotation and domain information extraction approaches.

1 Introduction

The preceding decade of research focused on biomedical information extraction (IE), the automatic detection and structured representation of relevant information from biomedical research papers, has brought significant advances in the state of the art in the domain. One recent development is the introduction of named entity (NE) recognition systems capable of detecting gene, protein and RNA entity mentions at practically applicable performance levels, as demonstrated in the BioCreative II challenge (Krallinger et al., 2008); another is the increase of interest in rich representations of extracted information, driven in particular by the BioNLP Shared Task on Event Extraction (Kim et al., 2009). The latter represents the first wider move in domain IE from the *relation representation* – where extracted information is captured only in the form of (possibly typed) pairs of entities – toward a representation capable of capturing complex associations involving multiple entities in different roles.

The *event representation* of the BioNLP Shared Task features an expressive model of the ways entities are associated in events. However, the entities considered in the task are limited to the basic gene, RNA and protein types (genes and gene products, below *GGP* for short), and their associations are only through events involving change or causal relations of the entities. We argue that the move toward rich representations for biomedical IE should be accompanied by broader consideration of entities and their relations, including entities referred to by non-NE terms and non-causal relations such as *part-of*. Representation of such relations would allow statements of entity associations to be modeled in greater detail, facilitating more accurate information extraction and extending the applicability of extracted representations.

In this study, we aim to advance toward broadly applicable resources for capturing such relations. Our suggested focus in the vast space of possible relations between biomedical domain entities is on relations between a GGP NE and non-NE terms. One one hand, this choice takes into account the focus in the domain on GGP NEs entities as precise references to relevant "real world" entities and allows us to build on the success of NE recognition systems. One the other hand, including non-NE terms allows us to considerably extend the coverage of represented information past that captured by purely NE-driven models and, as we will argue, fill a gap in

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine,
Short Papers, pages 97-102, Jeju Island, South Korea, 8-10 November 2009.

the commonly applied representation of the connection between NEs and events they participate in. We present a study of the relations between GGPs and terms that contain them as annotated in the GENIA corpus (Ohta et al., 2002; Kim et al., 2008) with the aim of discovering the key relations, establishing a classification system for annotating their types, and organizing these relation types in a type hierarchy.

2 Named Entity - Term relations

With few exceptions, biomedical IE efforts target relations or events directly involving NEs as participants. In some cases the source texts offer no more information (e.g. NE_1 *affects* NE_2), but often this approach requires approximation. Even in cases where the approximation is reasonable for many applications (e.g. NE_1 *affects* NE_2 *domain*, NE_1 *affects* NE_2 *mutant*), it necessarily limits the applicability of both the extraction method and the extracted information to those applications: if it is necessary to distinguish between, for example, statements involving NE_1 from those involving NE_1 *mutant*, a model that abstracts away the difference is not usable.

As a step toward a representation not limited in this way, we recently presented a task setting and representation making a number of such relations explicit (Pyysalo et al., 2009). The specific focus of the study was on relations such as part-of that "[...] hold between two entities without implication of change or causality". We presented relation types and annotation motivated by the data processing needs of the BioNLP event extraction task, capturing four different part-whole relation types, a task-specific *Variant* relation and a catch-all category *Other/Out* used to annotate cases not involving relevant relations. While sufficient for the specific need, these categories are arguably quite idiosyncratic and the annotation somewhat limited in applicability.

In the present study, we adopt the general task setting defined in our previous work and the representation of relations as ordered pairs of entities where both participating entities must be specified and their roles are fixed by the relation. By contrast, we seek to define a classification system with finer-grained distinctions and broader applicability.

3 Reference standard

The choice of relation types has far-ranging effects spanning from the effort to create annotation to its applicability and the feasibility of automatic extraction. One key issue is the granularity of the types: for example, whether to distinguish the relation of a gene to its *3' flanking region* from that to the *5' flanking region*, to annotate both as *gene-flanking region*, or, possibly, to simply capture these as instances of a general *object-component* relation. In explicitly seeking to identify types that are applicable more broadly than in the context of a specific task, we lose the ability to evaluate questions relating to issues such as granularity according to whether the task requires a specific distinction or not. To avoid having to rely entirely on subjective judgments, we chose to base our classification on an existing resource with broad community support.

As the relations need to be annotated by biomedical domain experts, we chose to base them on a domain reference standard instead of e.g. a general top-level ontology. We preliminarily considered a number of domain resources and chose as the most promising alternative to use the Medical Subject Headings (MeSH)[1] hierarchical controlled vocabulary as a reference. MeSH contains over 25,000 terms ("descriptors") covering a broad range of concepts in medicine and biology and is widely studied and applied in domain research. Further, entries in the PubMed literature database of currently approximately 17 million citations are manually labelled with MeSH descriptors, providing a rich potential source of related texts for each concept.

In considering the use of MeSH as a reference for relation types, it is important to note that MeSH terms primarily characterize individual entities, not their associations. However, given that the relations are specified to hold between an NE and a non-NE term and the participating NEs are limited to GGP types, the simple change of perspective of using the MeSH term to fix the role of the term in a GGP-term relation was found sufficient to suggest corresponding relation types. For example, the MeSH term *Protein Isoforms* suggests the GGP-term relation holding between a protein and its isoform, which we can specify in

[1] http://www.nlm.nih.gov/mesh

full e.g. as `GGP-Protein_isoform(GGP:NE, Protein_isoform:term)`. As we consider relations of the form $R(r_1:NE, r_2:term)$ where the roles of participants (r_1,r_2) are fixed by the relation type R, we will below simply use the relation type to refer to the relations.

We note that this formulation suggests that in the special case we consider here for purposes of relation type discovery the task could alternatively be cast as high-granularity term typing. However, relation-type annotation is necessary for the general case. For example, in the noun phrase NE_1 *binding domain of* NE_2 two distinct relation types hold between the term and the two NEs.

4 Data and annotation process

As the starting point for our work, we selected all terms annotated in the GENIA corpus that directly involve (contain) GGP NEs (Ohta et al., 2009), giving a total of 12520 terms. Thus, unlike in our previous work (Pyysalo et al., 2009) where only terms involved in specific events were considered, we here consider the entire set of terms annotated in GENIA. In focusing on terms that contain GGPs, the selection excludes many forms of statements of relations. However, based on our previous experience with the corpus we expect it to provide sufficient data to identify general classes of relations. To reduce annotation effort, we relied on two simplifying assumptions: that the relation between the term and the contained NE can be determined without reference to context[2] and that the specific name involved would not affect the relation. We could thus replace NEs with placeholders and judge unique cases of terms simplified in this way, reducing the number to 2554 cases. Finally, as our aim is to identify types that generalize to characterize a reasonable number of relations, we assumed that we could ignore terms whose (simplified) content appeared only once in the entire corpus. After this filtering, the final annotated dataset contained 518 unique cases representing 10368 term-NE instances, i.e. approximately 83% of the original unfiltered instances.

In the annotation process, each case was considered independently to determine the relation (or relations) that characterize how the contained NE is associated with the term. With the exception of some classes of relations excluded from more fine-grained characterization (see Section 5.5), the MeSH hierarchy was then consulted to determine the most specific MeSH concept applicable to define the relation. In cases where no applicable entry was found in MeSH, new types were considered. Finally, to avoid overlap with existing annotation and issues relating to gene/protein disambiguation, we as a general principle did not distinguish between a gene and its products, e.g. generalizing specific MeSH terms such as *RNA Precursor* and *Protein Precursor* to non-type specific terms such as *Precursor*.

5 Results

The identified relation types and the number of instances labeled with each are illustrated in Figure 1, which also shows our current organization of the types into a *is-a* hierarchy of relations. Our primary focus in this work is the definition of the relation types, not the specifics of their organization into a general taxonomy, and for organizing the types we have largely adopted the top-level structure of (Pyysalo et al., 2009), including the subdivision of part-whole relations following the taxonomy of (Winston et al., 1987).[3] In the following, we discuss the key relations and highlight some features of the proposed categorization and possible uses.

5.1 Equivalent entities

The most frequent type, *Equivalent*, is an important general relation we define as holding between an NE and a term that, in a neutral context, refers to the same entity as the NE or one that is equivalent under the equivalence relation holding between a gene and its products. In addition to cases such as *NE gene* or *NE protein*, the relation is used to mark e.g. *wild-type NE* as well as cases such as *transcription factor NE* involving (somewhat redundantly) an inherent characteristic of the NE. We expect that these

[2] This assumption, common in work on noun phrase semantics, was found not to hold in a small number of cases, in which the original context was studied.

[3] We note that while MeSH is primarily organized as a hierarchy, the relations implied by one term being the parent of another are not entirely consistent and we thus cannot rely on the structure of MeSH to suggest a consistent hierarchy of relation types.

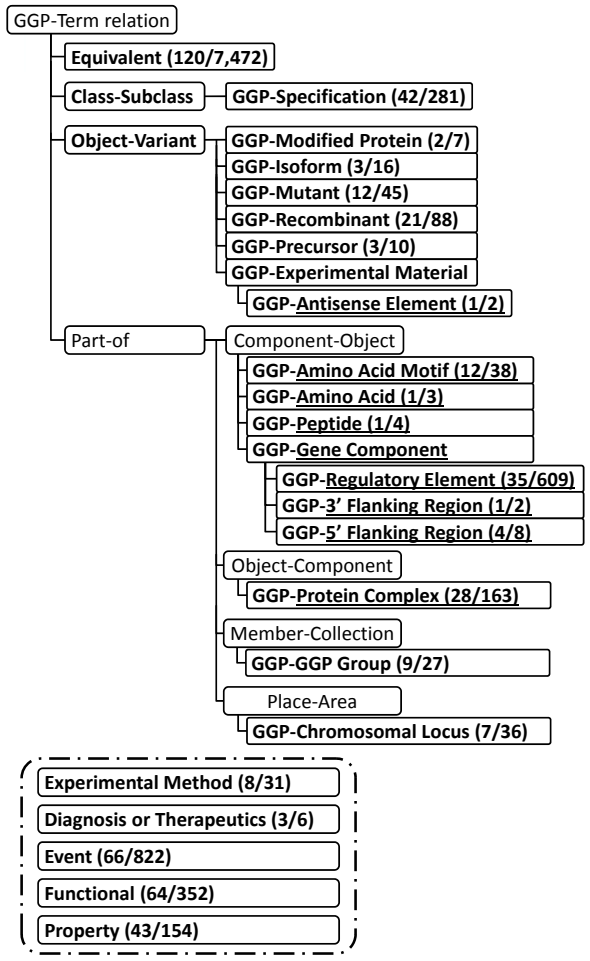

Figure 1: Relation types with number annotated (different cases/instances). Types in bold newly introduced in this study, underlined types drawn from MeSH. The separately shown unstructured terms identify categories of cases excluded from detailed classification. The lines connecting relation types represent *is-a* relations; e.g. the *Object-Component* relation *is-a Part-of* relation.

annotated cases could potentially benefit many applications as they suggest terms that could be simplified e.g. by replacing their text with the NE without altering meaning, a possibility that the inclusion of these cases under the general *Variant* type applied in our previous work did not allow.

5.2 Variants

The domain-specific relation class *Variant* suggested in our prior study is preserved in the hierarchy. However, the original single type covering highly heterogeneous cases was refined into types that can be used to identify the relation of the NE to the term in detail. In addition to the separation of Equivalent cases, the new categorization distinguishes between e.g. *GGP-Mutant* and *GGP-Isoform* relations. The use of terms involving different *Variant* relations is likely to vary considerably by application. In some cases, the relation can allow the identification of a specific entity that is referred to but not directly named: for example, any term with a *GGP-Precursor* relation to the *p50* protein refers to the *p105* protein.[4] A detailed (sub)domain ontology or database could support such remapping automatically. Distinctions between different *Variant* types also offer the general capacity to differentiate between terms by expected "functional distance" to their related NE. For example, assuming an information need for the binding partners of NE_1, an extracted *Binding* event involving a term with a *GGP-Modified Protein* relation (implying chemical modification) to NE_1 is more likely to be informative than one with a *GGP-Mutant* relation, which is in turn more likely to be informative than one with a *GGP-Recombinant* relation.

5.3 Part-of relations

Part-of relations, the most common category in our previous study, were also considerably refined. Interestingly, we found that while the data contained 163 instances of relations where the NE is a component of an object referred to by the term, these were fully homogeneous; all instances were cases of the relation holding between a subunit (NE) and a protein complex (term). The somewhat less frequent *Member-Collection* and *Place-Area* classes were similarly homogeneous. By contrast, *Component-Object* relation types, where the term refers to a component of the NE, were frequent in the data, of highly varying types, and represented to considerable detail in MeSH. The *Part-of* classes of relations can support some cases of simple, sound inference: for example, given the information that NE_1 binds T and that T is a component of NE_2, we can infer that NE_1 binds NE_2. The detailed relation types allow more specific inferences: for example, from binding of NE_1 to a T that has a *GGP-Regulatory Element* relation to NE_2 we can infer that NE_1 regulates NE_2.

[4] see http://www.uniprot.org/uniprot/P19838

5.4 Class-Subclass relations

In the current categorization, we added *Class-Subclass* as a separate top-level relation category. Cases where a GGP refers to a class of entities of which the term refers to a subclass (e.g. *Human* NE_1) were previously somewhat arbitrarily grouped together with *Member-Collection* relations. The new categorization allows clear differentiation between relations based on inherent characteristics and those involving more arbitrary groupings. The assumption that properties generalize across the class-subclass boundaries separating e.g. homologous human and mouse proteins is important in biology and frequently provisionally accepted by researchers. By contrast, an assumption that members of a same collection generally share their properties would be much less likely to hold: for example, members of a NE_1-*binding protein family* don't necessarily resemble each other in any other way than sharing the function defining the family.

While in the annotation process we generally allowed more than one relation type to be used to characterize the relation between a given NE and term, *Class-Subclass* relations were the only type found to occur together with other relations in the data; a typical case is *human NE promoter*. This case shows a specific benefit of recognizing multiple relations at once instead of annotating multiple levels of nested terms each involving a single relation: the annotation scheme does not force the arbitrary choice whether the term refers to the human variant of *NE promoter* or the promoter of the *human NE*.

5.5 Other relations

As in our previous work, we aimed to define relations that would complement existing annotations without overlap, further fitting the general focus of the GENIA corpus annotation. We thus identified but excluded from more detailed classification the following classes of relations: terms referring to a processes or events and NEs participating as *cause* or *theme* (e.g. *NE expression*; such cases are annotated in the current GENIA Event corpus), terms referring to separate entities identified through a functional or causal relation to the NE (e.g. *NE inhibitor*) terms containing an NE to characterize a property of the referred entity, not stating a simple direct relation (*NE-deficient mice*), and terms referring to entities considered out of scope of the annotation, such as experimental methods or diagnoses. For many applications, some of these annotations for "excluded" cases can also be applied e.g. to filter out irrelevant NE mentions from consideration. For example, proteins whose names occur only to define a property of another entity can be removed as candidate event participants in event extraction, thus potentially improving the precision of extraction.

6 Related work

Relation extraction has been extensively studied in both "general domain" and biomedical domain IE (see e.g. (Doddington et al., 2004; Zweigenbaum et al., 2007)). However, while relations targeted e.g. in the Automatic Content Extraction task focus on "static" types such as *Citizen-Of*, *Part-Of* and *Located*, the relations targeted by biomedical domain IE methods and corpora are almost exclusively of types that involve change or causal relations of the related entities. There are thus few domain studies or resources focusing on the types of relations we have considered here. Relation types similar to some of those we have identified here were considered also by (Rosario and Hearst, 2001) in their study of relations involved in biomedical compounds of two nouns, though their study largely excluded NEs and considered a broader domain, defining more generic relation types. A number of relations of types considered here are annotated in the BioInfer corpus (Pyysalo et al., 2007), likewise using somewhat more generic relations types (e.g. a single *Substructure* type covering what we have here subdivided as different *Component-Object* relation types), and the ITI TXM corpora contain extensive annotation for the specific relations connecting Mutants and Fragments with their parent proteins (Alex et al., 2008). However, the present study, which continues and extends our previous work on non-causal relations and their role in biomedical IE (Pyysalo et al., 2009), is to the best of our knowledge the first domain effort to characterize and annotate these relations at large scale (in terms of both corpus size and the number of relation types) or to the present level of detail.

7 Conclusions, discussion and future work

We argued that the move toward richer representation of the associations of named entities in biomedical information extraction should be accompanied by a more detailed model of the relations of named entities with other domain terms, including non-causal relations. To advance toward generally applicable resources for capturing such relations, we presented a study of relations holding between named entities and terms annotated in the GENIA corpus, aiming to create a relation classification system that could be applied together with rich representations for domain information extraction.

We studied 518 cases representing over 10,000 instances of NE-term relations, identifying for each the most specific MeSH terms that can be used to characterize the relation. Based on the study, we created a candidate hierarchy of relation types proposed for use in NE-term relation annotation and domain IE systems. The hierarchy is considerably more refined than that used in previous GENIA relation annotation and should not only allow better generalization through the removal of task-specific aspects but, we argued, can also support more types of inference. Nevertheless, the relation type hierarchy preserves some specific characteristics of both the GENIA data as well as the applied reference standard MeSH, suggesting that further development may be necessary to increase its applicability.

As future work, we intend to apply the identified relation types in creating NE-term relation annotation covering all NE-term pairs co-occurring within sentence scope in the GENIA corpus. We will also aim to refine the defined set of types to include a computationally implementable specification of types of inference they can support in the context of an event-type representation. The annotated data will be made available through the GENIA website.[5]

Acknowledgments

This work was partially supported by Grant-in-Aid for Specially Promoted Research (Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan).

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

References

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*.

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.

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 Shared Task*, pages 1–9, June.

Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. 2008. Evaluation of text-mining systems for biology: overview of the second biocreative community challenge. *Genome Biology*, 9(Suppl 2):S1.

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.

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

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, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the BioNLP 2009 Workshop*.

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.

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*.