A RE-EVALUATION OF BIOMEDICAL NAMED ENTITY–TERM RELATIONS

TOMOKO OHTA*, SAMPO PYYSALO† and JIN-DONG KIM‡,§

Department of Computer Science
University of Tokyo, Tokyo, Japan
*okap@is.s.u-tokyo.ac.jp
†smp@is.s.u-tokyo.ac.jp
‡jdkim@is.s.u-tokyo.ac.jp

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
tsujii@is.s.u-tokyo.ac.jp

Received 31 May 2010
Revised 3 July 2010
Accepted 3 July 2010

Text mining can support the interpretation of the enormous quantity of textual data produced in biomedical field. Recent developments in biomedical text mining include advances in the reliability of the recognition of named entities (NEs) such as specific genes and proteins, as well as movement toward richer representations of the associations of NEs. We argue that this shift in representation should be accompanied by the adoption of a more detailed model of the relations holding between NEs and other relevant domain terms. As a step toward this goal, we study NE–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.

Keywords: Text mining; information extraction; relation representation; event representation.

1. Introduction

In the recent decades, with the development of high-throughput screening methods, researchers in molecular biology and biomedicine need to interpret a tremendous
amount of data. To reduce the load on researchers, demands on methods for automatic data analysis are increasing. To improve access to information in domain research papers, there has been considerable focus in the preceding decade on methods for information extraction (IE), the automatic analysis and structured representation of information in natural language text. This focus has brought significant advances in the state of the art, both in basic methods for detecting mentions of entity names in text as well as in representations of their associations.

Genes and their products have a key role in the interpretation of major biological phenomena, and the recognition of mentions of their names in text is consequently a key task in domain IE. Recently, named entity (NE) recognition systems capable of detecting gene, protein and RNA name mentions at practically applicable performance levels have been introduced, and large-scale automatic annotation of key NE types including gene/protein entities is currently being pursued. A further notable development is the increase of interest in rich representations of extracted information, such as the event representation considered in the BioNLP 2009 Shared Task on Event Extraction. The shared task represents the first broad move in domain IE efforts toward a representation capable of capturing complex, structured associations involving multiple entities in different roles. While the event representation features an expressive model of the ways entities are associated, the core entities considered in the task are themselves 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. However, in natural language text, participants of events include not only GGP but also their variants (e.g. isoforms or mutants), specific regions (e.g. motifs or regulatory elements), complexes, families and groups, among others. These are referred to by terms with some structure and variability that are typically not specific enough to be considered names, but that can together with a name constitute specific references; consider, for instance, “p53 promoter region” or “promoter region of p53”.

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. On the one hand, this choice takes into account the focus in the domain on GGP NEs as precise references to relevant “real world” entities and allows us to build on the success of NE recognition systems. On 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 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 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. \( \text{NE}_1 \text{ affects } \text{NE}_2 \)), but often this approach requires approximation. Even in cases where the approximation is reasonable for many applications (e.g. \( \text{NE}_1 \text{ affects } \text{NE}_2 \) \text{ domain}, \( \text{NE}_1 \text{ affects } \text{NE}_2 \) \text{ 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 \( \text{NE}_1 \) from those involving \( \text{NE}_1 \text{ mutant} \), 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. 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 corpus 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)\(^a\) 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 19 million citations are manually labeled 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 full e.g. as \(\text{GGP-Protein isoform}(\text{GGP:NE, Protein isoform:term})\). As we consider relations of the form \(R(\text{r}_1: \text{NE}, \text{r}_2: \text{term})\) where the roles of participants \((\text{r}_1, \text{r}_2)\) are fixed by the relation type \(R\), below we will 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 \(\text{NE}_1 \text{ binding domain of NE}_2\), two distinct relation types hold between the term binding domain 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,\(^8\) giving a total of 12,520 terms. Thus, unlike in our previous work\(^9\) 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\(^b\) and that the specific name involved would not affect the relation. We could

\(^a\)http://www.nlm.nih.gov/mesh.
\(^b\)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.
thus replace NEs with placeholders and judge unique cases of terms simplified in this way, reducing the number to 2,554 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 10,368 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 characterizes how the contained NE is associated with the term. With the exception of some classes of relations excluded from more fine-grained characterization (see Sec. 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, the number of annotated cases labeled with each type, and the number of instances that these cases represent are illustrated in Fig. 1, which also shows our current organization of the types into an 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 our prior study, including the subdivision of part–whole relations following the taxonomy of Winston et al. 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 annotated cases could potentially benefit many applications as they suggest terms that could be simplified e.g. by replacing their text with the...
Fig. 1. Relation types. Number of annotated cases/number of annotated instances (see Sec. 4) of each type shown in parentheses. 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.
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-Precurisor relation to the p50 protein refers to the p105 protein. 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 NE1, an extracted Binding event involving a term with a GGP-Modified Protein relation (implying chemical modification) to NE1 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 at the chosen granularity; all were instances of the relation holding between a protein complex (term) and its subunit (NE). 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 NE1 binds T and that T is a component of NE2, we can infer that NE1 binds NE2. The detailed relation types allow more specific inferences: for example, from binding of NE1 to a T that has a GGP-Regulatory Element relation to NE2 we can infer that NE1 regulates NE2.

\[\text{see http://www.uniprot.org/uniprot/P19838.}\]

\[\text{Here, we use the term part-of broadly, without regard to the ordering of the arguments, thus including both Object-Component and Component-Object relations.}\]
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 NE1) 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 NE1–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 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. 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 in their study of relations involving 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, 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. However, the present study, which continues and extends our previous work on non-causal relations and their role in biomedical IE, 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 annotation offers a number of opportunities for the development of event extraction systems,\(^{15,16}\) potentially facilitating the introduction of more detailed representations of extracted information as well as more accurate extraction of presently targeted representations.\(^9\) Careful exploration of these opportunities remains future work. The annotated data will be made available through the GENIA website.\(^{1}\)

**Acknowledgments**

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

**References**


\(^{1}\)http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/.

Tomoko Ohta received her Ph.D. degree in Biotechnology from Tokyo Institute of Technology in 1996. From 1996 to 1999, she was at National Cancer Center Research Institute and the Institute of Medical Science, University of Tokyo as a Ph.D. candidate and postdoctoral fellow, respectively. In 1999, she joined Tsujii Laboratory at the University of Tokyo, where she has been working for the GENIA project. She is working on the GENIA corpus annotation, the GENIA ontology and semantic representation, and is one of the organizers of the BioNLP 2009 shared task.

Sampo Pyysalo received his M.Sc. degree in Computer Science from the University of Oulu, Finland in 2003, defended his Ph.D. thesis in Computer Science at the University of Turku, Finland, in 2008, and has since been working as a researcher in Tsujii Laboratory at the University of Tokyo. He is working primarily for the biotext mining project GENIA and is one of the organizers of the BioNLP 2009 shared task.

Jin-Dong Kim is Project Associate Professor of Text Mining and Bioinformatics in the Database Center for Life Science (DBCLS). He received both his M.Sc. and Ph.D. degrees in Computer Science from Korea University in 1996 and 2000, respectively. Since he joined Tsujii Laboratory at the University of Tokyo in 2001, he has been working primarily biotext mining. He is one of the main authors of the GENIA resources, and one of the organizers of the BioNLP shared tasks in 2004 and 2009.
Jun’ichi Tsujii received his B.Eng., M.Eng. and Ph.D. degrees in Electrical Engineering from Kyoto University, Japan, in 1971, 1973, and 1978, respectively. He was Assistant Professor and Associate Professor at Kyoto University, before accepting his position as Professor of Computational Linguistics at the University of Manchester Institute for Science and Technology (UMIST) in 1988. Since 1995, he has been professor at the Department of Computer Science at the University of Tokyo. He is also Professor of Text Mining at the University of Manchester (half-time), and Research Director of UK National Centre for Text mining (NaCTeM) since 2004. He was President of ACL (Association for Computational Linguistics) in 2006 and has been a permanent member of ICCL (International Committee on Computational Linguistics) since 1992.