Open Information Extraction from Biomedical Literature Using Predicate-Argument Structure Patterns

Nhung T. H. Nguyen[†*], Makoto Miwa[‡], Yoshimasa Tsuruoka[†] and Satoshi Tojo[*]
[†]The University of Tokyo, 3-7-1 Hongo, Bunkyo-ku, Tokyo, Japan
{nhung, tsuruoka}@logos.t.u-tokyo.ac.jp
[‡]The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK
makoto.miwa@manchester.ac.uk
[*]Japan Advanced Institue of Science and Technology, Ishikawa, Japan
{nthnhung, tojo@jaist.ac.jp}

Abstract

In this paper, we propose an open information extraction (Open IE) system, which attempts to extract relations (or facts) of any type from biomedical literature. What distinguishes our system from existing Open IE systems is that it uses predicate-argument structure patterns to detect the candidates of possible biomedical facts. We have manually evaluated the output of our system and found that it is reasonably accurate (50% precision). We have also applied our system to the whole MEDLINE and revealed that the relations between 'Amino Acid, Peptide, or Protein' entities are the most frequently described type of relations.

1 Introduction

Relation extraction is one of the most important tasks in biomedical text mining. Most of the studies on this topic have focused on specific or predefined types of relations, such as protein-protein interaction (Yakushiji et al., 2006; Airola et al., 2008; Miwa et al., 2009), drug-drug interaction (Segura-Bedmar et al., 2013), and biomolecular events (Nédellec et al., 2013). The scope of the types of relations that can be extracted by existing approaches is, therefore, inherently limited.

Recently, an information extraction paradigm called Open Information Extraction (Open IE) has been introduced to overcome the above-mentioned limitation (Banko et al., 2007; Fader et al., 2011; Mausam et al., 2012). Open IE systems aim to extract all possible relations from text. Although the concept of Open IE is certainly appealing, we have found that state-of-the-art Open IE systems, namely Reverb (Fader et al., 2011) and OLLIE (Mausam et al., 2012), do not perform well on biomedical text – they can capture relational phrases with reasonable accuracy but often fails to correctly identify their arguments.

This observation has motivated us to develop an Open IE system specifically designed for biomedical texts. Our system uses Predicate-Argument Structures (PAS) patterns to detect the candidates of possible biomedical facts. We decided to use PAS patterns because they are well normalized forms that represent deep syntactic relations. In other words, multiple syntactic variations are reduced to a single PAS, thereby allowing us to cover many kinds of expressions with a small number of PAS patterns. We first apply an HPSG-based syntactic parser to input sentences, and then match its output to predefined PAS patterns to detect pairs of relevant noun phrases (NPs). Named entities in these pairs are then detected; and finally, relations between these entities are extracted. The output of our system is, hopefully, a set of all semantic relations contained in the input.

Our contribution in this paper is twofold: (1) a simple but effective set of syntactic patterns for general relation extraction, and (2) an Open IE system that extracts biomedical facts from biomedical text; to the best of our knowledge, our system is the first Open IE system that attempts to detect relations from the whole MEDLINE in a general schema.

2 Related Work

Banko et al. (2007) introduced Open IE as a novel paradigm that facilitates domain independent discovery of relations extracted from text and readily scales to the diversity and size of the Web corpus. The system detects the tuples in the format of (argument 1; relational phrase; argument 2) without a pre-specified set of relations or domain-specific knowledge engineering. Several Open IE systems have been proposed up to now, including TextRunner (Banko et al., 2007), ReVerb (Fader et al., 2011), OLLIE (Mausam et al., 2012).

In the biomedical domain, large-scale event extraction has attracted many researchers (Rindflesch and Fiszman, 2003; Miyao et al., 2006; Björne et al., 2010; Taura et al., 2010; Rindflesch et al., 2011; Kilicoglu et al., 2012; Van Landeghem et al., 2013). Miyao et al. (2006) propose a system that extracts verb-mediated relations between genes, gene products, and diseases from MEDLINE. The output of their system is served as a database for MEDIE (Ohta et al., 2010), a semantic search engine on MEDLINE. Björne et al. (2010) apply their system to the titles and abstracts of all PubMed citations. Kilicoglu et al. (2012) also run their system on the entire set of PubMed citations to create SemMedDB, a repository of semantic predications.

SemRep (Rindflesch and Fiszman, 2003; Rindflesch et al., 2011) extracts semantic relationships from the titles and abstracts of all PubMed citations. Their relationships are represented by 30 specific *predicates* restricted to a limited number of verbs. Nebot and Berlanga (2012) extracted explicit binary relations of the form <*subject, predicate, object*> from CALBC initiative. To detect candidate relations, they proposed seven simple lexico-syntactic patterns. These two systems perform general relations extraction similar to ours, but unlike our system, neither of them use PAS patterns.

3 Our Open IE Framework

Since we focus on a general schema of relations, there is no labeled corpus suitable for learning the extraction model. Our system, therefore, relies solely on the input text and its linguistic characteristics such as the form, meaning, and context of the words. More specifically, we create patterns to capture these characteristics of text and then extract relations.

In order to find appropriate PAS patterns, we have first observed textual expressions that represent biomedical relations in GENIA corpus and found that those relations are usually expressed with verbs and prepositions; for example, $Entity_A$ {*affect, cause, express, inhibit ...*} $Entity_B$ or $Entity_A$ {*arise, happen, ...*} {*in, at, on ...*} *Location*. Our patterns in predicate-argument form and their corresponding examples are presented in Table 1. Patterns 1, 2, 3 and 4 are presented for transitive verbs. Intransitive verbs are captured by Pattern 5. The final pattern (Pattern 6) is used for prepositions, which would capture localization and whole-part relations. The elements NP_1 and NP_2 in each pattern are considered as candidate relations. In our system, Enju, an HPSG parser (Matsuzaki et al., 2007; Miyao et al., 2008), is employed to extract these candidates.

We then apply MetaMap[1] (Aronson and Lang, 2010) to identify named entities in the extracted NP pairs. At this stage, we apply two post-processes to remove false positive output from MetaMap. In the first process, we remove all entities that are verbs, adjectives, prepositions or numbers because we are only interested in noun or noun phrase ones. The second post-process is used to avoid common noun entities, such as 'binding', 'behaviors' and 'kinds'. In this process, we apply MetaMap to the whole MEDLINE and construct a dictionary of named entities and their occurrences. We then remove highly frequent entities from the dictionary. This dictionary is used to check the validity of named entities. Our statistical results on the whole MEDLINE revealed that the post-processes filtered out **70.83%** of the entities extracted by MetaMap. This filtering will help our system avoid extracting irrelevant relations.

After the above two post-processes, we obtain named entities in relevant NP pairs. Let us denote by $<NP_1, NP_2>$ a relevant NP pair, by e_{1i} ($i = 1, 2, ...$) entities in NP_1, and by e_{2j} ($j = 1, 2, ...$) entities in NP_2. Since NP_1 and NP_2 are relevant, we assume that every pair of entities $<e_{1i}, e_{2j}>$ is relevant, which means that they constitute a semantic relation. However, this assumption is so strong that it may create incorrect relations. In order to improve the precision of our system, we use the UMLS semantic network[2] as a constraint in extracting semantic relations. Let us denote by $<s_1, s_2>$ the pair of semantic types of $<e_{1i}, e_{2j}>$. If and only if $<s_1, s_2>$ exists in this semantic network, $<e_{1i}, e_{2j}>$ can constitute a relation.

4 Experimental Results

4.1 Performance on General Relations

Since there is no available labeled corpus for a general schema of relations, we manually evaluated our system on our own test set. This test set was created by randomly selecting 500 sentences from MEDLINE. Our system was given this set

[1] We employed MetaMap 2012 version 2 from `http://metamap.nlm.nih.gov/#Downloads`

[2] `http://semanticnetwork.nlm.nih.gov/`

No.	Type	PAS Patterns	Examples
1	Verb	$NP_1 \leftarrow \mathbf{Verb} \rightarrow NP_2$	protein RepA(cop) ← affects →a single amino acid
2		$NP_1 \leftarrow \mathbf{Verb} \rightarrow by + NP_2$	Diabetes ← induced → by streptozotocin injection
3		$NP_1 \leftarrow \mathbf{Verb} \rightarrow NP'$ $\uparrow$ $Prep. \rightarrow NP_2$	Endothelin-1 (ET-1) and ET-3 ← had → a strong effect $\uparrow$ $in \rightarrow$ all trabeculae
4		$NP_1 \leftarrow \mathbf{be} \rightarrow ADJP \leftarrow Prep. \rightarrow NP_2$	EPO receptor ← be → present ← in → tubular epithelial cells
5		$NP_1 \leftarrow \mathbf{Verb} \leftarrow Prep. \rightarrow NP_2$	subacute hepatitis ← results ← from → intravenous drug use
6	Prep.	$NP_1 \leftarrow \mathbf{Prep.} \rightarrow NP_2$	vitronectin ← in → the connective tissue

Table 1: Our PAS patterns focus on verb and preposition predicates. An arrow goes from a to b means a modifies b and a is called a predicate, b is called an argument. $<NP_1, NP_2>$ is a relevant NP pair in each pattern.

	Conf.	# of Rel.	Precision
ReVerb	≥0.3	75	46.67
	≥0.5	72	47.22
	≥0.7	58	46.55
OLLIE	≥0.3	124	38.71
	≥0.5	114	41.22
	≥0.7	89	42.69
Our patterns	-	**438**	**50.00**

Table 2: The precisions of relation extraction on our test set when using ReVerb and OLLIE with three confidence scores of 0.3, 0.5 and 0.7, and our PAS patterns to extract NP pairs.

as input, and returned a set of binary relations as output.

For comparison, we conducted experiments using two state-of-the-art Open IE systems, namely, ReVerb (Fader et al., 2011) and OLLIE (Mausam et al., 2012). We employed these two systems to extract relevant NP pairs in place of our PAS patterns. We chose confidence scores of 0.3, 0.5 and 0.7 as the thresholds for accepting generated tuples as candidate relations in our experiments. Next, the other processes were applied in the same way as our system. We report our evaluation results in Table 2. Compared with ReVerb and OLLIE, our PAS patterns generated the highest number of relations with the highest precision. This indicates that our PAS patterns perform better than the other approaches.

The causes of false positive relations include MetaMap errors, parser errors, and our greedy extraction. Since our system is based on the Enju parser, if the parser captures wrong noun phrases, our system will generate incorrect relevant pairs. For example, with this input "{[Laminin]}$_{NP_1}$ was *located in* {the zone of the basal [membrane], whereas [tenascin] was mainly found in the mucosal [vessels]}$_{NP_2}$", based on the NP pair $<NP_1, NP_2>$, the system returned two relations r_1 (Laminin, membrane) and r_2 (Laminin, vessels). In this example, the parser failed to detect the second NP of the pair; the correct one should be 'the zone of the basal membrane', not including 'whereas' clause. This error caused a false positive relation of *(Laminin, vessels)*. Extracted relation r_1 (Laminin, membrane) is also not correct because of the MetaMap error, i.e., the entity 'membrane' should be 'basal membrane'.

Although we use the Semantic Network to limit the generated relations, there are several false positive ones. For instance, given an input sentence: "{Efficiency of presentation of a peptide epitope by a [MHC class I molecule]}$_{NP_1}$ *depends on* {two parameters: its binding to the [MHC] molecule and its generation by intracellular Ag processing}$_{NP_2}$", the pair $<NP_1, NP_2>$ created a relation of (MHC class I molecule, MHC). This relation resulted from our greedy extraction. However, it is incorrect because 'MHC class I molecule' or 'MHC' is not the main subject of this sentence.

Table 2 shows that when using ReVerb and OLLIE to generate NP pairs, the numbers of extracted relations are significantly lower than those when using our patterns. The main reason is that these systems have failed to capture NP pairs in many sentences. In our test set, ReVerb and OLLIE could not extract NP pairs from 150 sentences and 95 sentences respectively; while our system could not extract pairs from 14 sentences only. Given the input sentence "{[Total protein], [lactate dehydrogenase] (LDH), [xanthine oxidase] (XO), [tumor necrosis factor] (TNF), and [interleukin 1] (IL-1)}$_{NP_1}$ were *measured in* {[bronchoalveolar lavage fluid] (BALF)}$_{NP_2}$.", ReVerb and OLLIE cannot extract any tuples, while our system generated a NP pair of $<NP_1, NP_2>$ and returned five

	AIMed		BioInfer		LLL	
	Pre.	Re.	Pre.	Re.	Pre.	Re.
(1)	71.8	*48.4*	-	-	-	-
(2)	52.9	61.8	47.7	59.9	72.5	87.2
(3)	55.0	68.8	65.7	71.1	77.6	86.0
Our system	30.3	**52.5**	51.2	44.9	87.5	81.5

Table 3: Performance of our system on AIMed, BioInfer and LLL corpora, compared with some notable systems for PPI: (1) Yakushiji et al. (2006), (2) Airola et al. (2008), and (3) Miwa et al. (2009).

	MedLine		DrugBank	
	Pre.	Re.	Pre.	Re.
Best system	55.8	*50.5*	81.6	83.8
Worst system	*62.5*	42.1	38.7	73.9
Our PAS patterns	27.0	**62.5**	41.0	61.6

Table 4: Performance of our system on MedLine and DrugBank corpora of SemEval-2013 Task 9 (Segura-Bedmar et al., 2013), compared with the best and worst system in that shared task.

correct relations between 'bronchoalveolar lavage fluid' and five entities in NP_1. This is a representative example showing the advantage of our PAS patterns in extracting candidate relations.

4.2 Performance on Predefined Relations

We also conducted experiments to check if our PAS patterns could cover other predefined relations, including Protein-Protein Interaction (PPI) and Drug-Drug Interaction (DDI). Regarding PPI, we applied our patterns to AIMed, BioInfer and LLL (Airola et al., 2008; Pyysalo et al., 2008). The available gold standard entities in these corpora were used instead of MetaMap output. Our experimental results and the results of some machine learning-based systems on PPI are shown in Table 3. It should be noted that these systems were evaluated by using 10-fold cross validation or using the test set; while our method is rule-based and thus we simply applied our patterns to the whole labeled corpora.

We conducted the same experiment for DDI on the SemEval-2013 task 9 corpus (Segura-Bedmar et al., 2013) and report the results in Table 4.

Results in Table 3 and Table 4 show that although our PAS patterns are very simple, their performance is competitive with other machine learning methods on both PPI and DDI. In some cases, our method even outperforms the other ones such as PPI on AIMed corpus and DDI on MedLine corpus in recall.

Rank	Semantic Relation		Count
	Entity 1	Entity 2	
1	aapp	aapp	**2,006,301**
2	cell	aapp	1,770,561
3	bpoc	aapp	1,046,523
4	gngm	aapp	1,008,017
5	dsyn	dsyn	909,195
6	aapp	dsyn	869,143
7	aapp	bacs	680,349
8	bpoc	mamm	676,325
9	lbpr	aapp	650,571
10	bpoc	dsyn	626,644

Table 5: The ten most frequent types of semantic relations found in the whole MEDLINE.

4.3 Extracting Semantic Relations in MEDLINE

We have applied our system to the whole MEDLINE[3] to extract semantic relations and calculated their frequencies to see which relations are common in this corpus. The statistical results in Table 5 show that the most common semantic relation in MEDLINE is the relation between 'Amino Acid, Peptide or Protein' (aapp) entities[4]. This explains why researchers in BioNLP have been focusing on protein-protein interaction. We can also see that 'Amino Acid, Peptide or Protein' entities contribute in 7 over 10 most popular relations, which shows their important role in the biomedical domain.

5 Conclusion

In this work, we have developed an Open IE system for biomedical literature by employing six PAS patterns to extract the candidates of possible biomedical facts. The system extracted 438 relations from our test set and 50% of those were correct. Compared with ReVerb and OLLIE, our patterns have presented better performance in extracting relevant NP pairs. The experimental results show that our patterns are effective on both general and specific relations. The statistical analysis on the result of the whole MEDLINE provides support for the intuition that the most common semantic relations are the ones between 'Amino Acid, Peptide and Protein' entities.

[3]The version used in this paper is the 2012 MEDLINE/PubMed baseline database.

[4]The semantic types of entities in Table 5 are in short form for our convenience, for their full form, please refer to `http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF`

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 1–9.

Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *JAMIA*, 17(3):229–236.

Michele Banko, Michael Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*, pages 2670–2676.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up Biomedical Event Extraction to the Entire Pubmed. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)*, pages 28–36. ACL.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP*, pages 1535–1545. ACL.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. 2012. SemMedDB: a pubmed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160.

Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2007. Efficient HPSG parsing with supertagging and cfg-filtering. In *Proceedings of IJCAI*, pages 1671–1676.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of EMNLP-CoNLL*, pages 523–534. ACL.

Makoto Miwa, Rune Sætre, Yusuke Miyao, and Jun'ichi Tsujii. 2009. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *I. J. Medical Informatics*, 78(12):39–46.

Yusuke Miyao, Tomoko Ohta, Katsuya Masuda, Yoshimasa Tsuruoka, Kazuhiro Yoshida, Takashi Ninomiya, and Jun'ichi Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. In *Proceedings of ACL*.

Yusuke Miyao, Rune Sætre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of ACL*, pages 46–54.

Victoria Nebot and Rafael Berlanga. 2012. Exploiting semantic annotations for open information extraction: an experience in the biomedical domain. *Knowledge and Information Systems*.

C. Nédellec, R. Bossy, J.-D. Kim, J-J. Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 1–7, August.

Tomoko Ohta, Takuya Matsuzaki, Naoaki Okazaki, Makoto Miwa, Rune Stre, Sampo Pyysalo, and Jun'ichi Tsujii. 2010. Medie and info-pubmed: 2010 update. *BMC Bioinformatics*, 11(S-5):P7.

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

Thomas C. Rindflesch and Marcelo Fiszman. 2003. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, 36(6):462–477.

Thomas C. Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, and Dongwook Shin. 2011. Semantic MEDLINE: An advanced information management application for biomedicine. *Information Services & Use*, (31):15–21.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. SemEval-2013 task 9 : Extraction of Drug-Drug interactions from Biomedical Texts. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, June.

Kenjiro Taura, Takuya Matsuzaki, Makoto Miwa, Yoshikazu Kamoshida, Daisaku Yokoyama, Nan Dun, Takeshi Shibata, Choi Sung Jun, and Jun'ichi Tsujii. 2010. Design and implementation of GXP make - a workflow system based on make. In *eScience*, pages 214–221. IEEE Computer Society.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One*, 8(4).

Akane Yakushiji, Yusuke Miyao, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2006. Automatic construction of predicate-argument structure patterns for biomedical information extraction. In *Proceedings of EMNLP*, pages 284–292. ACL.