Arguments of Nominals in Semantic Interpretation of Biomedical Text

Halil Kilicoglu,[1,2] Marcelo Fiszman,[2] Graciela Rosemblat,[2]
Sean Marimpietri,[3] Thomas C. Rindflesch[2]
[1]Concordia University, Montreal, QC, Canada
[2]National Library of Medicine, Bethesda, MD, USA
[3]University of California, Berkeley, CA, USA
h_kilico@cse.concordia.ca, sean.marimpietri@gmail.com
{fiszmanm,grosemblat,trindflesch}@mail.nih.gov

Abstract

Based on linguistic generalizations, we enhanced an existing semantic processor, SemRep, for effective interpretation of a wide range of patterns used to express arguments of nominalization in clinically oriented biomedical text. Nominalizations are pervasive in the scientific literature, yet few text mining systems adequately address them, thus missing a wealth of information. We evaluated the system by assessing the algorithm independently and by determining its contribution to SemRep generally. The first evaluation demonstrated the strength of the method through an F-score of 0.646 (P=0.743, R=0.569), which is more than 20 points higher than the baseline. The second evaluation showed that overall SemRep results were increased to F-score 0.689 (P=0.745, R=0.640), approximately 25 points better than processing without nominalizations.

1 Introduction

Extracting semantic relations from text and representing them as predicate-argument structures is increasingly seen as foundational for mining the biomedical literature (Kim et al., 2008). Most research has focused on relations indicated by verbs (Wattarujeekrit et al., 2004; Kogan et al., 2005). However nominalizations, gerunds, and relational nouns also take arguments. For example, the following sentence has three nominalizations, *treatment*, *suppression*, and *lactation* (nominalized forms of the verbs *treat*, *suppress*, and *lactate*, respectively). *Agonist* is derived from *agonize*, but indicates an agent rather than an event.

> *Bromocriptine, an ergot alkaloid dopamine **agonist**, is a recent common **treatment** for **suppression** of **lactation** in postpartum women.*

In promoting economy of expression, nominalizations are pervasive in scientific discourse, particularly the molecular biology sublanguage, due to the highly nested and complex biomolecular interactions described (Friedman et al., 2002). However, Cohen et al. (2008) point out that nominalizations are more difficult to process than verbs. Although a few systems deal with them, the focus is often limited in both the nominalizations recognized and the patterns used to express their arguments. Inability to interpret nominal constructions in a general way limits the effectiveness of such systems, since a wealth of knowledge is missed.

In this paper, we discuss our recent work on interpreting nominal forms and their arguments. We concentrate on nominalizations; however, the analysis also applies to other argument-taking nouns. Based on training data, we developed a set of linguistic generalizations and enhanced an existing semantic processor, SemRep, for effective interpretation of a wide range of patterns used to express arguments of nominalization in clinically oriented biomedical text. We evaluated the enhancements in two ways: by examining the ability to identify arguments of nominals independently and the effect these enhancements had on the overall quality of SemRep output.

2 Background

The theoretical linguistics literature has addressed the syntax of nominalizations (e.g. Chomsky, 1970; Grimshaw, 1990; Grimshaw and Williams, 1993), however, largely as support for theoretical argumentation, rather than detailed description of the facts. Quirk et al. (1985) concentrate on the morphological derivation of

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 46–54,
Uppsala, Sweden, 15 July 2010.

nominalizations from verbs. Within the context of NomBank, a project dedicated to annotation of argument structure, Meyers et al. (2004a) describe the linguistics of nominalizations, emphasizing semantic roles. However, major syntactic patterns of argument realization are also noted. Cohen et al. (2008) provide a comprehensive overview of nominalizations in biomedical text. They include a review of the relevant literature, and discuss a range of linguistic considerations, including morphological derivation, passivization, transitivity, and semantic topics (e.g. agent/instrument (*activator*) vs. action/process/state (*activation*)). Based on an analysis of the PennBioIE corpus (Kulick et al., 2004), detailed distributional results are provided on alternation patterns for several nominalizations with high frequency of occurrence in biomedical text, such as *activation* and *treatment*.

In computational linguistics, PUNDIT (Dahl et al., 1987) exploited similarities between nominalizations and related verbs. Hull and Gomez (1996) describe semantic interpretation for a limited set of nominalizations, relying on WordNet (Fellbaum, 1998) senses for restricting fillers of semantic roles. Meyers et al. (1998) present a procedure which maps syntactic and semantic information for verbs into a set of patterns for nominalizations. They use NOMLEX (MacLeod et al., 1998), a nominalization lexicon, as the basis for this transformation. More recently, the availability of the NomBank corpus (Meyers et al., 2004b) has supported supervised machine learning for nominal semantic role labeling (e.g. Pradhan et al., 2004; Jiang and Ng, 2006; Liu and Ng, 2007). In contrast, Padó et al. (2008) use unsupervised machine learning for semantic role labeling of eventive nominalizations by exploiting similarities between the argument structure of event nominalizations and corresponding verbs. Gurevich and Waterman (2009) use a large parsed corpus of Wikipedia to derive lexical models for determining the underlying argument structure of nominalizations.

Nominalizations have only recently garnered attention in biomedical language processing. GeneScene (Leroy and Chen, 2005) considers only arguments of nominalizations marked by prepositional cues. Similarly, Schuman and Bergler (2006) focus on the problem of prepositional phrase attachment. In the BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009), the most frequent predicates were nominals. Several participating systems discuss techniques that accommodate nominalizations (e.g. K. B. Cohen et al., 2009; Kilicoglu and Bergler, 2009). Nominalizations have not previously been addressed in clinically oriented text.

2.1 SemRep

SemRep (Rindflesch and Fiszman, 2003) automatically extracts semantic predications (logical subject-predicate-logical object triples) from unstructured text (titles and abstracts) of MEDLINE citations. It uses domain knowledge from the Unified Medical Language System® (UMLS ®) (Bodenreider, 2004), and the interaction of this knowledge and (underspecified) syntactic structure supports a robust system. SemRep extracts a range of semantic predications relating to clinical medicine (e.g. TREATS, DIAGNOSES, ADMINISTERED_TO, PROCESS_OF, LOCATION_OF), substance interactions (INTERACTS_WITH, INHIBITS, STIMULATES), and genetic etiology of disease (ASSOCIATED_WITH, PREDISPOSES, CAUSES). For example, the program identifies the following predications from input text *MRI revealed a lacunar infarction in the left internal capsule*. Arguments are concepts from the UMLS Metathesaurus and predicates are relations from the Semantic Network.

Magnetic Resonance Imaging DIAGNOSES Infarction, Lacunar

Internal Capsule LOCATION_OF Infarction, Lacunar

Processing relies on an underspecified syntactic analysis based on the UMLS SPECIALIST Lexicon (McCray et al., 1994) and the MedPost part-of-speech tagger (Smith et al., 2004). Output includes phrase identification, and for simple noun phrases, labeling of heads and modifiers.

[$_{\text{HEAD}}$(*MRI*)] [*revealed*] [*a* $_{\text{MOD}}$(*lacunar*), $_{\text{HEAD}}$(*infarction*)] [*in the* $_{\text{MOD}}$(*left*) $_{\text{MOD}}$(*internal*), $_{\text{HEAD}}$(*capsule*).]

MetaMap (Aronson and Lang, 2010) maps simple noun phrases to UMLS Metathesaurus concepts, as shown below. Associated semantic types are particularly important for subsequent processing.

[$_{\text{HEAD}}$(*MRI*){Magnetic Resonance Imaging (Diagnostic Procedure)}] [*revealed*] [*a* $_{\text{MOD}}$(*lacunar*), $_{\text{HEAD}}$(*infarction*) {Infarction, Lacunar(Disease or Syndrome)}] [*in the* $_{\text{MOD}}$(*left*) $_{\text{MOD}}$(*internal*), $_{\text{HEAD}}$(*capsule*) {Internal Capsule(Body Part, Organ, or Organ Component)}.]

This structure is the basis for extracting semantic predications, which relies on several mechanisms. Indicator rules map syntactic phenomena, such as verbs, nominalizations, prepositions, and modifier-head structure in the simple noun phrase to ontological predications. Examples include:

reveal (verb) → DIAGNOSES
in (prep) → LOCATION_OF

SemRep currently has 630 indicator rules. Ontological predications are based on a modified version of the UMLS Semantic Network and have semantic types as arguments. For example:

Diagnostic Procedure DIAGNOSES Disease or Syndrome
Body Part, Organ, or Organ Component LOCATION_OF Disease or Syndrome

Construction of a semantic predication begins with the application of an indicator rule, and is then constrained by two things. Arguments must satisfy syntactic restrictions for the indicator and must have been mapped to Metathesaurus concepts that match the arguments of the ontological predication indicated. As part of this processing, several syntactic phenomena are addressed, including passivization, argument coordination, and some types of relativization. For both verb and preposition indicators, underspecified syntactic rules simply ensure that subjects are on the left and objects on the right. Enhancing SemRep for nominalizations involved extending the syntactic constraints for arguments of nominalization indicators.

3 Methods

In order to gain insight into the principles underlying expression of nominal arguments, we first determined the 50 most common nominalizations in MEDLINE citations that also occur in the UMLS SPECIALIST Lexicon, and then analyzed a corpus of 1012 sentences extracted from 476 citations containing those nominalizations. We further limited these sentences to those with nominalizations containing two overt arguments (since SemRep only extracts predications with two arguments), resulting in a final set of 383 sentences. We determined 14 alternation patterns for nominalizations based on this analysis and devised an algorithm to accommodate them. We then conducted two evaluations, one to assess the effectiveness of the algorithm independently of other considerations and another to assess the contribution of enhanced nominalization processing to SemRep generally.

3.1 Nominal Alternations

Much work in identifying arguments of nominalizations assigns semantic role, such as agent, patient, etc., but SemRep does not. In this analysis, arguments are logical subject and object. Relational nouns often allow only one argument (e.g. *the weight of the evidence*), and either one or both of the arguments of a nominalization or gerund may be left unexpressed. SemRep doesn't interpret nominalizations with unexpressed arguments. If both arguments appear, they fall into one of several patterns, and the challenge in nominalization processing is to accommodate these patterns. Cohen et al. (2008) note several such patterns, including those in which both arguments are to the right of the nominalization, cued by prepositions (*treatment of fracture with surgery*), the nominalization separates the arguments (*fracture treatment with surgery*, *surgical treatment for fracture*), and both arguments precede the nominalizations, as modifiers of it (*surgical fracture treatment* and *fracture surgical treatment*).

Cohen et al. (2008) do not list several patterns we observed in the clinical domain, including those in which the subject appears to the right marked by a verb (*the treatment of fracture is surgery*) or as an appositive (*the treatment of fracture, surgery*), and those in which the subject appears to the left and the nominalization is either in a prepositional phrase (*surgery in the treatment of fracture, surgery in fracture treatment*) or is preceded by a verb or is parenthetical (*surgery is (the best) treatment for fracture*; *surgery is (the best) fracture treatment; surgery, the best fracture treatment*). One pattern, in which both arguments are on the right and the subject precedes the object, is seen most commonly in the clinical domain when the nominalization has a lexically specified cue (e.g. *the contribution of stem cells to kidney repair*). The nominal alternation patterns are listed in Table 1.

Generalizations about arguments of nominalizations are based on the position of the arguments, both with respect to each other and to the nominalization, and whether they modify the nominalization or not. A modifying argument is internal to the simple noun phrase of which the nominalization is the head; other arguments (both to the left and to the right) are external. (Relativization is considered external to the simple noun phrase.)

[NOM] [PREP OBJ] [PREP SUBJ] *Treatment of fracture with surgery*
[NOM] [PREP OBJ], [SUBJ] *The treatment of fracture, surgery*
[NOM] [PREP OBJ] ([SUBJ]) *The treatment of fracture (surgery)*
[NOM] [PREP OBJ] [BE] [SUBJ] *The treatment of fracture is surgery*
[NOM] [PREP SUBJ] [PREP OBJ] *Treatment with surgery of fracture*
[SUBJ NOM] [PREP OBJ] *Surgical treatment of fracture*
[SUBJ] [PREP NOM] [PREP OBJ] *Surgery in the treatment of fracture*
[SUBJ] [BE] [NOM] [PREP OBJ] *Surgery is the treatment of fracture*
[OBJ NOM] [BE] [SUBJ] *Fracture treatment is surgery*
[OBJ NOM] [PREP SUBJ] *Fracture treatment with surgery*
[SUBJ] [PREP OBJ NOM] *Surgery for fracture treatment*
[SUBJ] [BE] [OBJ NOM] *Surgery is the fracture treatment*
[SUBJ OBJ NOM] *Surgical fracture treatment*
[OBJ SUBJ NOM] *Fracture surgical treatment*

Table 1. Patterns

Argument cuing plays a prominent role in defining these patterns. A cue is an overt syntactic element associated with an argument, and can be a preposition, a verb (most commonly a form of *be*), a comma, or parenthesis. A cued argument is in a dependency with the cue, which is itself in a dependency with the nominalization. The cue must occur between the nominalization and the argument, whether the argument is to the right (e.g. *treatment* ***of*** *fracture*) or to the left (e.g. *surgery* ***in*** *the treatment*). Prepositional cues for the objects of some nominalizations are stipulated in the lexicon; some of these are obligatory (e.g. *contribution – to*), while others are optional (*treatment – for*).

External arguments of nominalizations must be cued, and cues unambiguously signal the role of the argument, according to the following cuing rules (Cohen et al., 2008). Verbs, comma, parenthesis, and the prepositions *by, with,* and *via* cue subjects only. (*By* is used for semantic role agent and *with* for instrument, but SemRep does not exploit this distinction.) *Of* cues subjects only if the nominalization has an obligatory (object) cue; it must cue objects otherwise. There is a class of nominalizations (e.g. *cause*) that do not allow a prepositionally cued subject. Considerable variation is seen in the order of subject and object; however, if the subject intervenes between the nominalization and the object, both must have equal cuing status (the only possibilities are that both be either uncued or cued with a preposition).

3.2 Algorithm

In extending SemRep for identifying arguments of nominalizations, existing machinery was exploited, namely shallow parsing, mapping simple noun phrases to Metathesaurus concepts, and the application of indicator rules to map nominalizations to enhanced Semantic Network ontological predications (which imposes restrictions on the semantic type of arguments). Finally, syntactic argument identification was enhanced specifically for nominalizations and exploits the linguistic generalizations noted. For example in the sentence below, phrases have been identified and *cervical cancer* has been mapped to the Metathesaurus concept “Cervix carcinoma” with semantic type ‘Neoplastic Process’, and *vaccination* to “Vaccination” (‘Therapeutic or Preventive Procedure’). An indicator rule for *prevention* maps to the ontological predication “Therapeutic or Preventive Procedure PREVENTS Neoplastic Process” (among others) in generating the predication: “Vaccination PREVENTS Cervix carcinoma.”

> *Therefore,* ***prevention of*** *cervical cancer* ***with*** *HPV vaccination may have a significant financial impact.*

Processing to identify arguments for *prevention* begins by determining whether the nominalization has a lexically specified object cue. This information is needed to determine the cuing function of *of*. Since it is common for there to be at least one argument on the right, identification of arguments begins there. Arguments on the right are external and must be cued. If a cued argument is found, its role is determined by the argument cuing rules. Since *prevention* does not have a lexically specified cue, *of* marks its object. Further, the semantic type of the concept for the object of *of* matches the object of the ontological predication (‘Neoplastic Process’).

The algorithm next looks to the right of the first argument for the second argument. Since processing addresses only two arguments for nominalizations, subject and object, once the role of the first has been determined, the second can be inferred. For cued arguments, the process checks that the cue is compatible with the cuing rules. In all cases, the relevant semantic type must match the subject of the ontological predication. In this instance, *with* cues subjects and ‘Therapeutic or Preventive Process’ matches the subject of the ontological predication indicated.

If only one noun phrase to the right satisfies the argument cuing rules, the second argument must be on the left. A modifier immediately to the left of the nominalization (and thus an internal argument) is sought first, and its role inferred from the first argument. Since internal arguments are not cued, there is no need to ensure cuing compatibility. The predication “Operative Surgical Procedures TREATS Pregnancy, Ectopic” is found for *resolution* in

> *Surgical* ***resolution of*** *an ectopic pregnancy in a captive gerenuk (Litocranius walleri walleri).*

Resolution is an indicator for the ontological predication “Therapeutic or Preventive Procedure TREATS Disease or Syndrome.” S*urgical* maps to “Operative Surgical Procedures” (‘Therapeutic or Preventive Procedure’), which matches the subject of this predication, and *ectopic pregnancy* maps to “Pregnancy, Ectopic” (‘Disease or Syndrome’), which matches its object. O*f* marks the object of *resolution*.

An argument to the left of a nominalization may be external, in which case a cue is necessary. *For* preceding *treatment* satisfies this requirement in the following sentence.

> *Preclinical data have supported the use of fludarabine and cyclophosphamide (FC) in combination* ***for*** *the* ***treatment of*** *indolent lymphoid malignancies.*

The two drugs in this sentence map to concepts with semantic type ‘Pharmacologic Substance’ and the malignancy has ‘Neoplastic Process’, as above. There is an ontological predication for TREATS with subject ‘Pharmacologic Substance’. After coordination processing in SemRep, two predications are generated for *treatment*:

Cyclophosphamide TREATS Malignant lymphoid neoplasm
Fludarabine TREATS Malignant lymphoid neoplasm

If there is no argument to the right, both arguments must be on the left. A modifier immediately to the left of the nominalization is sought first. Given the properties of cuing (the cue intervenes between the argument and the nominalization), if both arguments occur to the left, at least one of them must be internal, since it is not possible to have more than one external argument on the left (e.g. **Surgery is fracture for treatment*). The role of the first argument is found based on semantic type. The first modifier to the left of *treatment* in the following sentence is *epilepsy*, which has semantic type ‘Disease or Syndrome’, matching the object of the ontological predication for TREATS.

> *Patients with most chances of benefiting from surgical epilepsy* ***treatment***

The second modifier to the left, *surgical* maps to the concept “Operative Surgical Procedures,” whose semantic type matches the subject of the ontological predication. These conditions allow construction of the predication “Operative Surgical Procedures TREATS Epilepsy.”

In the next sentence, the indicator rule for *prediction* maps to the ontological predication “Amino Acid, Peptide, or Protein PREDISPOSES Disease or Syndrome.”

> *The potential clinical role of measuring these apolipoproteins* ***for*** *ischemic stroke* ***prediction*** *warrants further study.*

Ischemic stroke satisfies the object of this predication and *apolipoproteins* the subject. Since the external subject is cued by *for*, all constraints are satisfied and the predication “Apolipoproteins PREDISPOSES Ischemic stroke” is generated.

3.3 Evaluation

Three-hundred sentences from 239 MEDLINE citations (titles and abstracts) were selected for annotating a test set. Some had previously been selected for various aspects of SemRep evaluation; others were chosen randomly. A small number (30) were sentences in the GENIA event corpus (Kim et al., 2008) with bio-event-triggering nominalizations. Annotation was conducted by three of the authors. One, a linguist (A), judged all sentences, while the other two, a computer scientist (B) and a medical informatics researcher (C), annotated a subset. Annotation was not limited to nominalizations. The statistics regarding the individual annotations are given below. The numbers in parentheses show the number of annotated predications indicated by nominalizations.

Annotator	# of Sentences	# of Predications
A	300	533 (286)
B	200	387 (190)
C	132	244 (134)

Table 2. Annotation statistics

As guidance, annotators were provided UMLS Metathesaurus concepts for the sentences. However, they consulted the Metathesaurus directly to check questionable mappings. Annotation focused on the 25 predicate types SemRep addresses.

We measured inter-annotator agreement, defined as the F-score of one set of annotations, when the second is taken as the gold standard. After individual annotations were complete, two annotators (A and C) assessed all three sets of annotations and created the final reference standard. The reference standard has 569 predications, 300 of which (52.7%) are indicated by nominalizations. We further measured the agreement between individual sets of annotations and the reference standard. Results are given below:

Annotator pair	# of Sentences	IAA
A-B	200	0.794
A-C	132	0.974
B-C	103	0.722
A-Gold	300	0.925
B-Gold	200	0.889
C-Gold	132	0.906

Table 3. Inter-annotator agreement

We performed two evaluations. The first (*eval1*) evaluated nominalizations in isolation, while the second (*eval2*) assessed the effect of the enhancements on overall semantic interpretation in SemRep. For *eval1*, we restricted SemRep to extract predications indicated by nominalizations only. The baseline was a nominalization argument identification rule which simply stipulates that the subject of a predicate is a concept to the left (starting from the modifier of the nominalization, if any), and the object is a concept to the right. This baseline implements the underspecification principle of SemRep, without any additional logic. We compared the results from this baseline to those from the algorithm described above to identify arguments of nominalizations. The gold standard for *eval1* was limited to predications indicated by nominalizations.

We investigated the effect of nominalization processing on SemRep generally in *eval2*, for which the baseline implementation was SemRep with no nominalization processing. The results for this baseline were evaluated against those obtained using SemRep with no restrictions. Typical evaluation metrics, precision, recall, and F-score, were calculated.

4 Results and Discussion

The results for the two evaluations are presented below.

	Precision	Recall	F-Score
eval1			
Baseline	0.484	0.359	0.412
With NOM	0.743	0.569	0.645
eval2			
Baseline	0.640	0.333	0.438
With NOM	0.745	0.640	0.689

Table 4. Evaluation results

Results illustrate the importance of nominalization processing for effectiveness of semantic interpretation and show that the SemRep methodology naturally extends to this phenomenon. With a single, simple, rule (*eval1 baseline*), SemRep achieves an F-score of 0.412. With additional processing based on linguistic generalizations, F-score improves more than 20 points. Further, the addition of nominalization processing not only enhances the coverage of SemRep (more than 30 points), but also increases precision (more than 10 points). While nominalizations are generally considered more difficult to process than verbs (Cohen et al., 2008), we were able to accommodate them with greater precision than other types of indicators, including verbs (0.743 vs. 0.64 in *eval1 with NOM* vs. *eval2 baseline*) with our patterns.

	Precision	Recall	F-Score
eval1			
Baseline	0.233	0.140	0.175
With NOM	0.690	0.400	0.506
eval2			
Baseline (No NOM)	0.667	0.278	0.392
With NOM	0.698	0.514	0.592

Table 5. Results for molecular biology sentences

Limiting the evaluation to sentences focusing on biomolecular interactions (from GENIA), while not conclusive due to the small number of sentences (30), also shows similar patterns, as shown in Table 5. As expected, while overall quality of predications is lower, since molecular biology text is significantly more complex than that in the clinical domain, improvements with nominalization processing are clearly seen.

Errors were mostly due to aspects of SemRep orthogonal to but interacting with nominalization processing. Complex coordination structure was the main source of recall errors, as in the following example.

> *RESULTS: The best* ***predictors of*** *incident metabolic syndrome* ***were*** *waist circumference (odds ratio [OR] 1.7 [1.3-2.0] per 11 cm), HDL cholesterol (0.6 [0.4-0.7] per 15 mg/dl), and proinsulin (1.7 [1.4-2.0] per 3.3 pmol/l).* [PMID 14988303]

While the system was able to identify the predication "Waist circumference PREDISPOSES Metabolic syndrome," it was unable to find the predications below, due to its inability to identify the coordination of *waist circumference, HDL cholesterol,* and *proinsulin.*

(FN) Proinsulin PREDISPOSES Metabolic syndrome
(FN) High Density Lipoprotein Cholesterol PREDISPOSES Metabolic syndrome

Mapping of noun phrases to the correct UMLS concepts (MetaMap) is a source of both false positives and false negatives, particularly in the context of the molecular biology sentences, where acronyms and abbreviations are common and their disambiguation is nontrivial (Okazaki et al., 2010). For example, in the following sentence

> *PTK* ***inhibition with*** *Gen attenuated both LPS-induced NF-kappaB DNA binding and TNF-alpha production in human monocytes.* [PMID 10210645]

PTK was mapped to "Ephrin receptor EphA8" rather than to "Protein Tyrosine Kinase", causing both a false positive and a false negative.

(FP) Genistein INHIBITS Ephrin receptor EphA8
(FN) Genistein INHIBITS Protein Tyrosine Kinase

Some errors were due to failure to recognize a relative clause by SemRep. Only the head of such a structure is allowed to be an argument outside the structure. In the sentence below, the subject of *treatment* is *hyperthermic intraperitoneal intraoperative chemotherapy*, which is the head of the reduced relative clause, *after cytoreductive surgery*.

> *Hyperthermic intraperitoneal intraoperative chemotherapy after cytoreductive surgery* ***for the treatment of*** *abdominal sarcomatosis: clinical outcome and prognostic factors in 60 consecutive patients.* [PMID 15112276]

SemRep failed to recognize the relative clause, and therefore the nominalization algorithm took the noun phrase inside it as the subject of *treatment*, since it satisfies both semantic type and argument constraints.

(FP) Cytoreductive surgery TREATS Sarcamatosis NOS
(FN) intraperitoneal therapy TREATS Sarcamatosis NOS

A small number of errors were due solely to nominalization processing. In the following sentence, the object of *contribution* is cued with *in*, rather than lexically specified *to*, which causes a recall error.

> *Using SOCS-1 knockout mice, we investigated the* ***contribution of*** *SOCS-1* ***in*** *the development of insulin resistance induced by a high-fat diet (HFD).* [PMID 18929539]

(FN) Cytokine Inducible SH-2 Containing Protein PREDISPOSES Insulin Resistance

Accurate identification of the arguments of nominalizations in the molecular biology subdomain is more challenging than in clinically-oriented text. Some of the syntactic structure responsible for this complexity is discussed by K. B. Cohen et al. (2009). In particular, they note the problem of an argument being separated from the nominalization, and point out the problem of specifying the intervening structure. Although we have not focused on molecular biology, the analysis developed for clinical medicine shows promise in that domain as well. One relevant extension could address the syntactic configuration in which intervening structure involves an argument of a nominalization shared with a verb occurring to the left of the nominalization, as *induced* and *activation* interact in the following sentence:

> *IL-2* ***induced*** *less STAT1 alpha* ***activation*** *and IFN-alpha* ***induced*** *greater STAT5* ***activation*** *in NK3.3 cells compared with preactivated primary NK cells.* [PMID 8683106]

This could be addressed with an extension of our rule that subjects of nominalizations can be cued with verbs. With respect to argument identification, *induce* can function like a form of *be.*

5 Conclusion

We discuss a linguistically principled implementation for identifying arguments of nominalizations in clinically focused biomedical text. The full range of such structures is rarely addressed by existing text mining systems, thus missing valuable information. The algorithm is implemented inside SemRep, a general semantic interpreter for biomedical text. We evaluated the system both by assessing the algorithm independently and by determining the contribution it makes to SemRep generally. The first evaluation resulted in an F-score of 0.646 (P=0.743, R=0.569), which is 20 points higher than the baseline, while the second showed that overall SemRep results were increased to F-score 0.689 (P=0.745, R=0.640), approximately 25 points better than processing without nominalizations.

Since our nominalization processing is by extending SemRep, rather than by creating a dedicated system, we provide the interpretation of these structures in a broader context. An array of semantic predications generated by mapping to an ontology (UMLS) normalizes the interpretation of verbs and nominalizations. Processing is linguistically based, and several syntactic phenomena are addressed, including passivization, argument coordination, and relativization. The benefits of such processing include effective applications for extracting information on genetic diseases from text (Masseroli et al., 2006), as well as research in medical knowledge summarization (Fiszman et al., 2004; Fiszman et al., 2009), literature-based discovery (Ahlers et al., 2007; Hristovski et al., 2010), and enhanced information retrieval (Kilicoglu et al., 2008; T. Cohen et al., 2009).

Acknowledgments

This study was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

C. B. Ahlers, D. Hristovski, H. Kilicoglu, T. C. Rindflesch. 2007. Using the literature-based discovery paradigm to investigate drug mechanisms. In *Proceedings of AMIA Annual Symposium*, pages 6-10.

A. R. Aronson and F.-M. Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17:229-236.

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267-70.

N. Chomsky. 1970. Remarks on nominalization. In Jacobs, Roderick, and Peter S. Rosenbaum (eds.) *Readings in English transformational grammar.* Boston: Ginn and Company, pages 184-221.

K. B. Cohen, M. Palmer, L. Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE,* 3(9): e3158.

K. B. Cohen, K. H. Verspoor, H. L. Johnson, C. Roeder, P. V. Ogren, W. A. Baumgartner, E. White, H. Tipney, L. Hunter. 2009. High-precision biological event extraction with a concept recognizer. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 50-58.

T. Cohen, R. Schvaneveldt, T. C. Rindflesch. 2009. Predication-based semantic indexing: Permutations as a means to encode predications in semantic space. In *Proceedings of AMIA Annual Symposium*, pages 114-118.

D. A. Dahl, M. S. Palmer, R. J. Passonneau. 1987. Nominalizations in PUNDIT. In *Proceedings of ACL*, pages 131-139.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA.

M. Fiszman, D. Demner-Fushman, H. Kilicoglu, T. C. Rindflesch. 2009. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. *Journal of Biomedical Informatics,* 42(5):801-813.

M. Fiszman, T. C. Rindflesch, H. Kilicoglu. 2004. Abstraction summarization for managing the biomedical research literature. In *Proceedings of HLT/NAACL Workshop on Computational Lexical Semantics,* pages 76-83.

C. Friedman, P. Kra, A. Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics,* 35:222–235.

J. Grimshaw. 1990. *Argument Structure.* MIT Press, Cambridge, MA.

J. Grimshaw and E. Williams. 1993. Nominalizations and predicative prepositional phrases. In J. Pustejovsky (ed.) *Semantics and the Lexicon.* Dordrecht: Kluwer Academic Publishers, pages 97-106.

O. Gurevich and S. A. Waterman. 2009. Mining of parsed data to derive deverbal argument structure. In *Proceedings of the 2009 Workshop on Grammar Engineering Across Frameworks*. pages 19-27.

D. Hristovski, A. Kastrin, B. Peterlin, T. C. Rindflesch. 2010. Combining semantic relations and DNA microarray data for novel hypothesis generation. In C. Blaschke, H. Shatkay (Eds.) *ISMB/ECCB2009, Lecture Notes in Bioinformatics*, Heidelberg: Springer-Verlag, pages 53-61.

R. D. Hull and F. Gomez. 1996. Semantic interpretation of nominalizations. In *Proceedings of AAAI*, pages 1062-1068.

Z. P. Jiang and H. T. Ng. 2006. Semantic role labeling of NomBank: A maximum entropy approach. In *Proceedings of EMNLP'06*, pages 138–145.

H. Kilicoglu and S. Bergler. 2009. Syntactic dependency based heuristics for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 119-127.

H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, A. M. Ripple, T. C. Rindflesch. 2008. Semantic MEDLINE: A Web application to manage the results of PubMed searches. In *Proceedings of SMBM'08*, pages 69-76.

J-D. Kim, T. Ohta, S. Pyysalo, Y. Kano, J. Tsujii. 2009. Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1-9.

J-D. Kim, T. Ohta, J. Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(1):10.

Y. Kogan, N. Collier, S. Pakhomov, M. Krauthammer. 2005. Towards semantic role labeling & IE in the medical literature. In *Proceedings of AMIA Annual Symposium*, pages 410–414.

S. Kulick, A. Bies, M. Liberman, M. Mandel, R. McDonald, M. Palmer. A. Schein, L. Ungar. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of BioLINK: Linking Biological Literature, Ontologies and Databases*, pages 61–68.

G. Leroy and H. Chen. 2005. Genescene: An ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. *Journal of the American Society for Information Science and Technology*, 56(5): 457–468.

C. Liu and H. Ng. 2007. Learning predictive structures for semantic role labeling of NomBank. In *Proceedings of ACL*, pages 208–215.

C. Macleod, R. Grishman, A. Meyers, L. Barrett, R. Reeves. 1998. NOMLEX: A lexicon of nominalizations. In *Proceedings of EURALEX'98*.

M. Masseroli, H. Kilicoglu, F-M. Lang, T. C. Rindflesch. 2006. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. *BMC Bioinformatics*, 7:291.

A. T. McCray, S. Srinivasan, A. C. Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of 18th Annual Symposium on Computer Applications in Medical Care*, pages 235–239.

A. Meyers, C. Macleod, R. Yanbarger, R. Grishman, L. Barrett, R. Reeves. 1998. Using NOMLEX to produce nominalization patterns for information extraction. In *Proceedings of the Workshop on Computational Treatment of Nominals (COLING/ACL)*, pages 25-32.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, R. Grishman. 2004a. Annotating noun argument structure for NomBank. In *Proceedings of LREC*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, R. Grishman. 2004b. The NomBank project: An interim report. In *Proceedings of HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31.

N. Okazaki, S. Ananiadou, J. Tsujii. 2010. Building a high quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*: btq129+.

S. Padó, M. Pennacchiotti, C. Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of CoLing'08*, pages 665-672.

S. Pradhan, H. Sun, W. Ward, J. Martin, D. Jurafsky. 2004. Parsing arguments of nominalizations in English and Chinese. In *Proceedings of HLT/NAACL*, pages 141–144.

R. Quirk, S. Greenbaum, G. Leech, J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

T. C. Rindflesch and M. 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):462-77.

J. Schuman and S. Bergler. 2006. Postnominal prepositional phrase attachment in proteomics. In *Proceedings of BioNLP Workshop on Linking Natural Language Processing and Biology*, pages 82–89.

L. Smith, T. C. Rindflesch, W. J. Wilbur. 2004. MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320-2321.

T. Wattarujeekrit, P. K. Shah, N. Collier. 2004. PASBio: Predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5:155.