A Typology of Semantic Relations Dedicated to Scientific Literature Analysis

Kata Gábor[1], Haïfa Zargayouna[1], Isabelle Tellier[2],
Davide Buscaldi[1], Thierry Charnois[1]

[1] LIPN, CNRS (UMR 7030), Université Paris 13 Sorbonne Paris Cité
{gabor,haifa.zargayouna,davide.buscaldi,thierry.charnois}@lipn.univ-paris13.fr
[2] LaTTiCe, CNRS (UMR 8094), ENS Paris, Université Sorbonne Nouvelle - Paris 3
PSL Research University, Université Sorbonne Paris Cité
isabelle.tellier@univ-paris3.fr

Abstract We propose a method for improving access to scientific literature by analyzing the content of research papers beyond citation links and topic tracking. Our model relies on a typology of explicit semantic relations. These relations are instantiated in the abstract/introduction part of the papers and can be identified automatically using textual data and external ontologies. Preliminary results show a promising precision in unsupervised relationship classification.

1 Introduction

Compiling a state of the art is a fundamental activity for the understanding of any scientific research field. This activity requires the analysis of the existing literature to identify the involved concepts and actors and track relevant topics. Chavalarias and Cointet [3], Herrera et al. [8] or Skupin [17] work on analyzing and visualizing the evolution of topics over time. Citation links are extensively used to explore scientific communities [12,13]. [6] provides a list of usual tasks on bibliographies for different classes of users. The Citation Typing ontology (CiTo) [16] presents a typology of citations according to the relation between the research papers they express.

Citations alone, however, are not enough to fully understand the evolution of a research field: researchers need to analyze the contribution of individual papers. Such an analysis is focused on specific concepts and relations, for instance to identify that a *method* has been developed to *tackle* some *problem*, that a *refinement* of an *existing solution* has been *developed*, etc. For this purpose, we need to (i) identify the entities and concepts that describe a scientific field (*method, problem*) and (ii) identify the semantic relations between these entities (*tackle, developed*). Doing so, we will build semantic links between articles that go much beyond explicit citations. Hence, the definition and identification of the relevant semantic relations is at the core of our approach.

We combine natural language processing techniques with statistical term extractors and external ontological resources. Ontologies allow a fine-tuned semantic

analysis, as opposed to Open Information Extraction [5] or general-purpose approaches exploiting terminology extraction on the fly [11,3,12,17]. In particular, systems using an ontology can benefit from various typed relations. As a corpus of scientific texts, we use the ACL Anthology Corpus [15] and we focus on the "abstract" and "introduction" sections, as they provide the most informative description of the content of a paper. However, our approach does not rely on manually annotated data and aims to be domain-independent: the semantic relations considered are generic for any scientific field.

Section 2 introduces the main architecture of our model for semantic analysis. Section 3 proposes a typology of the semantic relations in the scientific domain, while section 4 briefly exposes the methodology by which this model has been instantiated. Finally, section 5 draws some conclusions and perspectives for future work.

2 General model

Our purpose is to automatically extract relevant semantic relations in the science/engineering domain, as they appear in texts such as "a (new) method is proposed for a task", or "a phenomenon is found in a certain context". By identifying concepts and semantic relations between concepts, we can detect research papers which deal with the same problem, or track the evolution of results on a certain task. Our model of the scientific domain contains scientific articles linked to typed relations whose arguments are mapped to existing ontologies (see Figure 1).

The process used to implement the model consists of three sub-tasks: entity annotation, concept mapping and relation classification. Entity annotation is the task of recognizing instances of domain concepts in the text . Concept mapping consists in finding mappings with external ontologies or vocabularies. Relation classification uses the entity-annotated text as input and aims to identify the relations between two entities based on a combination of two information sources: the text sequence between the two entities (extracted from the corpus), and the semantic type of the entities (extracted from the ontology). We are currently experimenting with an iterative process: after annotating concepts in the corpus, we extract sequential patterns, which are used to identify instances of known relations and to discover new *types* of relations. A further goal is to enrich ontologies with new relation types [14].

For example, from the text: *"This database contains recorded, transcribed and annotated read speech"* we want to be able to extract the following relation:

```
<relation type="composed_of">
<arg1><entity majorType="BabelNet" synset="bn:00025333n">
database</entity></arg1>
contains (...)
<arg2><entity majorType="BabelNet" synset="bn:00049911n">
speech</entity></arg2></relation>
```

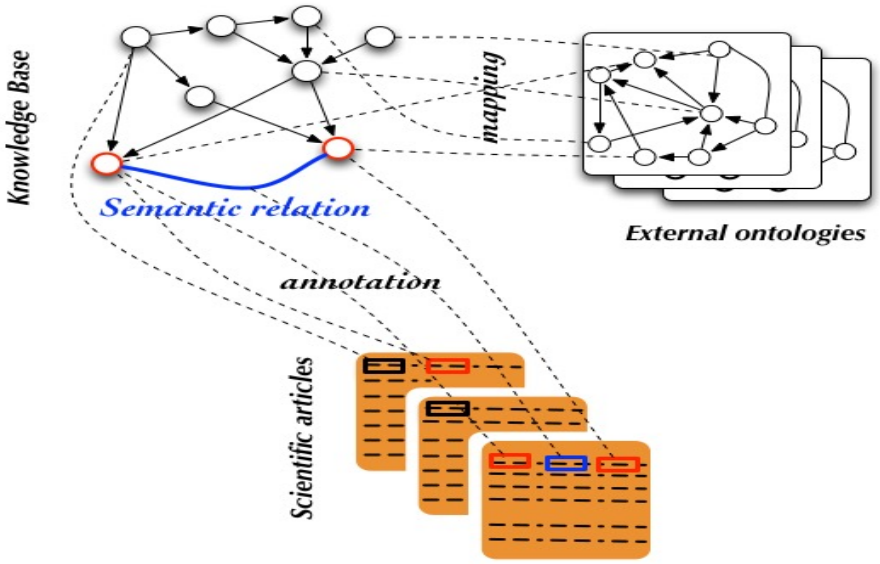

Figure 1. The general model

3 Semantic relations in scientific literature

A data-driven approach was adopted to discover relation types represented in the corpus. Our corpus contains 4.200.000 words from 11.000 papers (abstracts and introductions) in the ACL Anthology Corpus, pre-processed by E. Omodei [11]. A sample of 100 abstracts was extracted and instances of explicit semantic relations were discovered and manually annotated on these data.

Pattern-based approaches of relationship extraction [1] and classification [18] rely on the hypothesis that the context of occurrence of entity mention pairs is characteristic of the semantic relation between the two concepts. Thus, only linguistically explicit relations were taken into account. On the textual level, a semantic relation is conceived as a text span linking two annotated instances of concepts within the same sentence (see the example above). On the semantic level, relation types need to be specific enough to be easily distinguished from each other by a domain expert. Argument types are very informative when specifying a relation. Instances of arguments are typically domain-specific (e.g. the kind of *data* or *resources* are different across domains), hence, the link is ensured by mapping the entities to external ontologies. Table 1 provides the typology of relations that has been defined, together with argument type specifications. Table 2 shows how the various types of semantic relations are represented in this corpus.

The typology was set up on the basis of examples from the 100 abstracts. As a next step, we selected a sample of 500 abstracts to be manually annotated using the current typology. The agreement rate between the annotators will indicate

affects	ARG1: *specific property of data* ARG2: *results*
based_on:	ARG1: *method, system* based on ARG2: *other method*
char	ARG1: *observed characteristics* of an observed ARG2: *entity*
compare	ARG1: *result (of experiment)* compared to ARG2: *result2*
composed_of	ARG1: *database/resource* ARG2: *data*
datasource	ARG1: *information* extracted from ARG2: *kind of data*
method_applied	ARG1: *method* applied to ARG2: *data*
model	ARG1: *abstract representation* of an ARG2: *observed entity*
phenomenon	ARG1: *entity, a phenomenon* found in ARG2: *context*
problem	ARG1: *phenomenon* is a problem in a ARG2: *field/task*
propose	ARG1: *paper/author* presents ARG2: *an idea*
study	ARG1: *analysis* of a ARG2: *phenomenon*
tag	ARG1: *tag/meta-information* associated to an ARG2: *entity*
task_applied	ARG1: *task* performed on ARG2: *data*
used_for	ARG1: *method/system* ARG2: *task*
uses_information	ARG1: *method* relies on ARG2: *information*
yields	ARG1: *experiment/method* ARG2: *result*
wrt	ARG1 *a change* in/with respect to ARG2: *property*

Table 1. Semantic relation typology based on 100 abstracts

whether the relations are well defined on the semantic level (possible to classify), and whether they are indeed explicit on the textual level (possible to annotate). In case of a successful validation, these data will serve to evaluate relationship extraction and classification experiments.

Relation	Frequency in corpus
used_for	27%
composed_of	16%
propose	11%
yields	6%
study	6%
task_applied	5%
uses_information	4%
affects	4%

Table 2. Most frequent relations in manually annotated abstracts

4 Model instantiation

Entity annotation was applied to the corpus of 4.2 million words in two steps. First, candidates were generated with the terminology extraction tool TermSuite [4]. The list of extracted terms was then mapped to different ontological resources: the knowledge base of Saffron Knowledge Extraction Framework [2], and the

BabelNet ontology [10]. If a term was validated as a domain concept (i.e., found in at least one of the resources), it was annotated in the text. The complete process of entity annotation is described in [7].
A first set of unsupervised, pattern-based clustering experiments was performed to detect semantic relations. First, entity mention pairs were extracted from the corpus, together with the sequences. A co-occurrence matrix was built from entity pairs and sequences. The matrix rows (entity pairs) were clustered using CLUTO's [9] divisive algorithm with repeated bisections. As we are experimenting with completely unsupervised methods, the number of clusters to detect was not fixed to the number of relations in our typology. The reported evaluation was carried out on a set of 700 entity pairs, manually classified to one or more of the semantic relations we defined. Table 3 summarizes the results compared to a random clustering baseline. Precision and recall are calculated in terms of pairs of items correctly or falsely assigned to the same cluster or to different clusters.

Input	#clusters	Precision	Recall	F-measure
baseline	100	0.095	0.009	0.017
baseline	50	0.103	0.019	0.033
baseline	25	**0.104**	**0.041**	**0.058**
Sequences	100	**0.490**	0.046	0.084
Sequences	50	0.378	0.079	0.132
Sequences	25	0.313	**0.140**	**0.193**

Table 3. Evaluation of clustering

5 Conclusion

We presented a model for the analysis of scientific papers in order to automatically extract states of the art of a research field. The core of the model is a typology of semantic relations in the scientific domain, which was defined while manually annotating data from a corpus of natural language processing papers. These relationscan be identified automatically using a combination of pattern mining and natural language processing techniques. The first results on recognizing relations between unseen concepts are already very encouraging: a precision of 0.5 means that one out of two pairs of concepts assigned to the same category belong to the same semantic relation (among 18 distinct possible ones). Experiments are currently being carried out with bi-clustering algorithms, where text sequences and concept pairs are clustered at the same time.

Acknowledgments This work is part of the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL).

References

1. A. Auger and C Barrière. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 2008.
2. G. Bordea, P. Buitelaar, and T. Polajnar. Domain-independent term extraction through domain modelling. In *10th TIA Conference*, 2013.
3. D. Chavalarias and J-P. Cointet. Phylomemetic patterns in science evolution - the rise and fall of scientific fields. *PLOS ONE*, 8(2), 2013.
4. B. Daille, C. Jacquin, L. Monceaux, E. Morin, and J. Rocheteau. TTC TermSuite : Une chaîne de traitement pour la fouille terminologique multilingue. In *Proceedings of the TALN Conference*, 2013.
5. L. Del Corro and R. Gemulla. Clausie: Clause-based open information extraction. In *Proceedings of the 22nd WWW Conference*, pages 355–366, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
6. A. Di Iorio, R. Giannella, F. Poggi, and F. Vitali. Exploring bibliographies for research-related tasks. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 1001–1006. International World Wide Web Conferences Steering Committee, 2015.
7. K. Gabor, H. Zargayouna, D. Buscaldi, I. Tellier, and T. Charnois. Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *Proceedings of the LREC2016 Conference*, Portoroz, Slovenia, 2016.
8. M. Herrera, D. C. Roberts, and N. Gulbahce. Mapping the evolution of scientific fields. *PLOS ONE*, 5, 2010.
9. G. Karypis. Cluto: A clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002.
10. R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
11. E. Omodei, J-P. Cointet, and T. Poibeau. Mapping the natural language processing domain : Experiments using the acl anthology. In *Proceedings of the LREC 2014 Conference*, 2014.
12. F. Osborne and E. Motta. Mining semantic relations between reserach areas. In *International Semantic Web Conference, Boston (MA)*, 2012.
13. F. Osborne and E. Motta. Rexplore: unveiling the dynamics of scholarly data. *Digital Libraries*, 8(12), 2014.
14. G. Petasis, V. Karkaletsis, G. Paliouras, A. Krithara, and E. Zavitsanos. Ontology population and enrichment: State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution*, pages 134–166. Springer-Verlag, 2011.
15. D.R. Radev, P. Muthukrishnan, and V. Qazvinian. The acl anthology network corpus. In *Proceedings of the 2009 ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 2009.
16. D. Shotton. Cito, the citation typing ontology. *J. Biomedical Semantics*, 1(S-1):S6, 2010.
17. A. Skupin. The world of geography: Visualizing a knowledge domain with cartographic means. *Proceedings of the National Academy of Sciences*, 101, 2004.
18. P. D. Turney. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, 2006.