# Goal-Oriented Representation of Scientific Papers

**Jumana Nassour[1] , Michael Elhadad[1], Arnon Sturm[2]**
[1]Dept. of Computer Science, [2]Dept. of Software and Information Systems Engineering
Ben-Gurion University of the Negev
{jumanan,elhadad}@cs.bgu.ac.il, sturm@bgu.ac.il

**Abstract**

Scientific papers constitute an essential part of the research process. Thus, there is a need for a way to better access and analyze them. In this paper, we present a generalized goal-oriented schema for creating an abstract visual representation of scientific papers. We evaluate the schema by means of suitability for various domains, inter-annotator agreement, and usage on various parts of the paper.

**Keywords:** Term Extraction, Relation Extraction, Evaluation Corpus, Language Resource, Knowledge Mapping, Concept Maps.

## 1. Introduction

Scientific papers are an essential source of information for research. They are written in a "semi-structured" manner, that is, they follow a standardized rhetorical structure (abstract, introduction, motivation, related work, citations), but are presented in free text. The high variability of the form of scientific papers complicates the task of automatically extracting information from them. Learning a new domain and staying informed requires intense reading, especially since the amount of scientific publications is increasing rapidly (their number was estimated at 50 million in 2010 and accelerating (Jinha, 2010)).

Consider a case where a researcher working on Text Generation decides to use discourse structure to improve abstract generation, but she does not know which methods to use. She comes across Saggion (2009) paper: "A classification algorithm for predicting the structure of summaries," and wants to find out methods for predicting discourse structure, ways of testing them, evaluation datasets, and the methods that give the best results. This type of goal-oriented discovery of a research domain is the task we are investigating. We aim at producing a formal representation of a research domain that can help discover techniques that solve tasks and their qualities.

Choi et al. (2016) present a graphical representation that models experimental results in the specific domain of Machine Translation. Their model is designed to answer similar research questions in that domain only. In this paper, we present a generalization of their model, using a more general goal-oriented annotation schema. Such an annotation schema should help researchers answer questions requiring inference, and comparison between methods according to different criteria. Our eventual goal is to produce such maps automatically.

We evaluate our schema in terms of: generalization, inter-annotator agreement, and expressiveness when considering different sections of the paper for annotation. By generalization, we mean usability in multiple domains for the purpose of domain exploration. In inter-annotation agreement, we check whether the schema can be reliably used to annotate scientific papers. To assess expressiveness, we compare between annotations received when considering only abstracts vs. considering more sections of the paper.

Our generalized representation is based on a goal-oriented annotation schema with two types of concepts: TASK and QUALITY, and four types of relations: ACHIEVED-BY, CONSISTS-OF, ASSOCIATED-WITH, and CONTRIBUTES-TO. It is visually represented using concept maps (Novak and Gowin, 1984). [1]

We start by describing related work on Concept and Relation Extraction[2], then we explain the annotation schema and the annotation process where we evaluate our schema in terms of generalization, inter-annotator agreement and expressiveness, and report statistics about the dataset.

## 2. Related Work

The task of annotating scientific papers has been addressed by multiple recent projects, mostly directed towards indexing (Tateisi et al., 2016; QasemiZadeh and Schumann, 2016; Augenstein et al., 2017). Our approach, inspired by Choi et al. (2016), is aimed towards automatically producing readable representations to help readers explore a domain. Potential uses of this representation and its experimental evaluation are presented in Sturm et al. (2017b). Table 4 contains a comparison of the different approaches we survey.

Choi et al. (2016) represent the experimental results of a scientific paper in a graphical model. The graph includes: DATASETS — name, size, and language; EXPERIMENT TYPE — goals and methods to achieve them (9 predefined goals and 27 methods); RESULTS — value, metric, and the system that achieved it. Their dataset consists of 67 papers in the field of Machine Translation.

QasemiZadeh and Schumann (2016) created a dataset annotating over 300 abstracts from ACL with the following term types: *method*, *tool*, *Language resource (LR)*, *LR product*, *model*, *measures and measurements*, and *other*. No relations were annotated - but the model is quite similar to that of (Choi et al., 2016), and is specific to ACL literature.

Tateisi et al. (2016) defined an annotation schema to identify technical concepts and the roles they play in the work

[1]Concept maps are labeled graphs with concepts as nodes and relations as edges between them.

[2]Concept Extraction is also referred to in the literature as Entity Extraction, Term Extraction, or key-phrase Extraction.

described by the paper. The annotation schema consists of 12 concept types (*e.g.*, *quantity*, *modality*, *organization*), not counting subtypes, and 20 relations (*e.g.*, APPLY-TO, RESULT, AGENT). Their dataset consists of annotated abstracts of scientific papers both in English (250 papers from ACL and 140 papers from ACM) and Japanese (230 papers from from IPSJ). Our annotation schema is more specific, focusing on goal-oriented modeling, and including less concept and relation types.

Gábor et al. (2016) automatically extracted concepts using entity linking to generic and domain specific ontologies for the purpose of creating an automated analysis of scientific corpora (abstracts and introductions from 100 papers taken from ACL). In addition, they propose a typology of semantic relations between concepts consisting of 18 domain-specific and 3 generic relations. Entities in ontologies are usually noun phrases. In contrast, our schema is based on tasks, which tend to appear as verbal phrases and clauses.

Augenstein et al. (2017) published a similar task as part of the SemEval effort. The task included extracting both key-phrases and relations between them from scientific documents in Computer Science, Material Sciences and Physics (500 paragraphs). Three types of terms were annotated (*process*, *task*, *material*), and two types of relations HYPONYM-OF and SYNONYM-OF.

Most tasks consider only abstracts (Tateisi et al., 2016; QasemiZadeh and Schumann, 2016) for annotation, since they supposedly constitute a short summary of the paper, while others consider more paragraphs (Augenstein et al., 2017; Choi et al., 2016; Gábor et al., 2016). Westergaard et al. (2018) show that mining full-text articles outperforms mining abstracts solely, in the task of extracting published protein-protein, disease-gene, and protein sub-cellular associations. Hence, we decided to examine their claim by comparing the represntations received when considering only abstracts (abstract-based) vs. considering additional relevant parts (paper-based).

## 3. Annotation Schema

Scientific papers dealing with technology domains, contain information about problems in the domain, and suggest solutions together with practical methods, and evaluations of these methods. In order to model scientific papers by these elements, we adopted Means-Ends maps (ME-MAP), a goal-oriented annotation schema introduced in Sturm et al. (2017a) and Maraee and Sturm (2017) that is based on means-end relations.

The ME-MAP annotation schema includes two types of concepts: TASK — describes problems and solutions (*e.g.*, entity matching, multi-document summarization, crowdsourcing); and QUALITY — describes qualities the tasks have (*e.g.*, performance, evaluation metrics). The schema specifies four types of relations: ACHIEVED-BY — describes a task (end) that can be achieved by another task (means), including instance-of relations; CONSISTS-OF/SUBTASK-OF — describes the decomposition of a main task into its subtasks; ASSOCIATED-WITH — associates a task with a certain quality (explicitly mentioned or implied); CONTRIBUTES-TO — describes a contribution being made to a quality by a task or another quality (contributor) with a contribution value. Table 1 contains an example of each of the relations. For example, the contribution in the example is made by *our method* to the quality *performance* associated with the task *the two problems*.

| Relation | Example |
|---|---|
| ACHIEVED-BY | Our method (means) can be used for cleaning existing datasets from duplicates (end) ... |
| CONSISTS-OF | Matching natural language sentences (subtask) is central for many applications such as information retrieval (main task) and question answering (main task). |
| CONTRIBUTES-TO QUALITY | Experimental results on real-world datasets show our method (contributor) achieves high (contribution) performance (quality) on the two problems (associated task). |

Table 1: Examples of the different relations of ME-MAP annotation schema.

## 4. Annotation Process

The annotation is performed using an on-line questionnaire, where annotators are asked to provide the following: main tasks (problems/solutions), relations between the tasks (ACHIEVED-BY/SUBTASK-OF), datasets used for evaluation, attributes/evaluation metrics, and comparison/results. The annotations are then automatically converted to the annotation schema, and then to concept maps using Cytoscape.js (Franz et al., 2015) for visualization. To facilitate application of key-phrase/relation extraction methods on our dataset, the annotations are also made available in BRAT's (Stenetorp et al., 2012) format, which is used by most annotation schemas including Augenstein et al. (2017).

The annotation was done by a computer science PhD student, and reviewed by an expert in natural language processing and a knowledge mapping expert, with the exception of 7 full papers that were annotated by 2 annotators to measure inter-annotator agreement. Annotators were instructed to extract tasks (problems/solutions) exactly as they appear in the text, and add relations between them based on the text, whether implied or explicitly specified.

### 4.1. Multi-Domain Generalization

To check how well our generalized schema applies to different domains, we started by first applying it to the machine translation domain using the same cluster of papers used by Choi et al. (2016), then we tried to apply it to two other domains: Summarization, and Ontology Alignment, which does not necessarily contain experimental results.

The machine translation cluster was automatically converted into our schema. We then manually re-annotated 18 papers from this cluster. Fig.4 contains the transformation of Marcu et al. (2006)'s paper representation from Choi et al. (2016)'s (Fig. 4a), to the automatically converted version (Fig. 4b), followed by the result of the re-annotation (Fig. 4c). From the automatically converted representation, Fig. 4b, we learn that the paper suggests a method called *SPMY-Comb* that improves *Machine Translation* by addressing its subtask, *add better rule context*, and that it performs better on *BLEU* than another method called *PBMT*

when tested against two datasets: *NIST 2003(< 20 words)* and *NIST 2003*. Questions about goals, techniques, names of datasets, method names, metrics, and results which can be answered from the Choi et al. (2016)'s representation, can be easily answered by our automatically generated representation as well; however, our representation does not contain information about the dataset: size, languages, and purpose (train/test/dev). We opted for a simpler visual representation.

The re-annotated representation in Fig. 4c gives a more comprehensive abstraction of the domain and the paper. We learn from it that the paper suggests a *statistical translation model that uses syntactified target language*, which incorporates the following subtasks: *target language submodels*, *a Kneser-Ney (1995) smoothed trigram language model*, *a rule extraction algorithm*, and *submodels (feature functions) developed in phrase-based systems* that are used for *choosing target translations of source language phrases*. We also learn that different features were tested; explanation of the different features is not included in this figure for simplicity. In addition, we learn that the results were compared against a *phrase-based system* called *PBMT - Och and Ney, 2004*.

The re-annotated version offers a more complete abstraction of the paper, since it contains all of the methods that have been tested with their results, and the subtasks that the paper uses/addresses and the technologies it uses.

The domain of Ontology Alignment is not related to Machine Translation, and does not follow a similar empirical data-driven methodology. To assess to which extent the approach suggested by Choi et al. (2016) applies to it in comparison to our schema, we took the paper of Severo et al. (2014) from the Ontology Alignment cluster, and annotated it using both approaches. Choi et al. (2016)'s approach focuses on goals, techniques, and experimental results, and it specifies a single technique per goal. Taking these considerations in mind, we would choose the following goals: *better visualize ontology alignments*, *better manipulate ontology alignments* and their respective techniques: *use a web-based environment*, *provide more manipulation options*. Fig. 1 contains a simplified version of the paper-annotation. The map gives a representation of *alignment tool*, the tasks that connect this concept to *ontology alignment*: *ontology alignment manipulation*, *ontology alignment visualization*, and the qualities that can be used to assess an *alignment tool*: *evaluation*, *alignment manipulation*, *correspondences edition*, *external matching*, *alignment visualization*, and *facilitate access*.

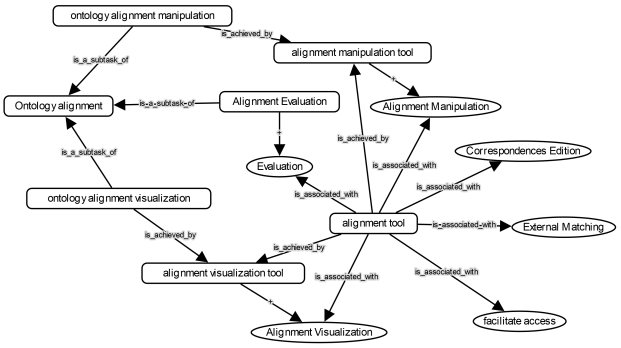



Figure 1: A simplified version of a paper-annotation of an Ontology Alignment paper that does not contain experimental results.

## 4.2. Inter-Annotator Agreement

We calculated inter-annotator agreement between two annotators on 7 papers from the abstract-based dataset using two measures: Jaccard score and F-score as calculated in QasemiZadeh and Schumann (2016). Table 2 contains the results. The tasks were checked for exact string match agreement (*Task*), which resulted in a low agreement, and after manually aligning them (*Task + Alignment*). For example, in Fig.1, the task "alignment visualization tool" could be annotated as "visualization tool" by a 2nd annotator. 23% of the tasks needed alignment, leading to a high increase in the agreement (F-score of 52% vs 85% – higher than the 76% agreement reported by QasemiZadeh and Schumann (2016)). The major difference between the aligned tasks was different choice of boundaries. The rest of the disagreements were caused by: completely different labels given to the same task/edge (47%), tasks that the annotator forgot to annotate (33%), and tasks which described a conjunction of simpler tasks (20%).

Our relations are transitive *e.g.*, if T1 CONSISTS-OF T2 and T2 CONSISTS-OF T3, we infer that T1 CONSISTS-OF T3. We report agreement results with and without considering inferred relations. As expected, transitive relations improve agreement, F-Score of 63% vs 73% over all relations and tasks. ACHIEVED-BY relations, which are the basis of our goal-oriented schema, received the highest agreement (71%). We analyzed disagreement cases over both ACHIEVED-BY and CONSISTS-OF relations: tasks not annotated by one of the annotators(68%), different labels (29%).

| Type | Jaccard | F-Score |
|---|---|---|
| Task | 35.14% | 52.00% |
| Task + Alignment | 74.14% | 85.15% |
| ASSOCIATED-WITH | 20.00% | 33.33% |
| ASSOCIATED-WITH + Transitivity | 40.00% | 57.14% |
| CONTRIBUTES-TO | 11.11% | 20.00% |
| CONTRIBUTES-TO + Transitivity | 38.89% | 56.00% |
| ACHIEVED-BY | 47.37% | 64.29% |
| ACHIEVED-BY + Transitivity | 55.26% | 71.19% |
| CONSISTS-OF | 45.45% | 62.50% |
| CONSISTS-OF + Transitivity | 50.00% | 66.67% |
| Total | 46.67% | 63.64% |
| Total + Transitivity | 57.62% | 73.11% |

Table 2: Inter-annotation agreement for abstract-based annotation. The relations were checked for agreement with and without transitive relations after aligning the tasks.

## 4.3. Considering Different Sections for Annotation

The annotation process included two different parts: abstract-based annotation where we consider only abstracts

for annotation, and paper-based annotation, where we consider additional relevant sections. In paper-based annotation, we first consider the abstract where the main problem is usually mentioned, then consider table/s with the main results, which contain specific mentions of methods/tools and their qualities, and end with looking for these specific mentions in the paper, to figure out which specific problems they address, and how they are connected to the general problem. This usually includes: titles, the last/before last paragraph of the introduction, and some paragraphs from the methodology and the experiment sections.

Fig. 2 contains the concept map of the abstract-based annotation of Saggion (2009) paper, Fig. 3 is the paper-based annotation of the same paper. From the first figure, we learn that the paper addresses the problem of *Abstract Generation*, by suggesting a *supervised machine learning* algorithm that *predicts the discourse structure*, a subtask of *Abstract Generation*, and improves its *accuracy* (*60%*). The algorithm uses both *local feature* and *contextual features*.

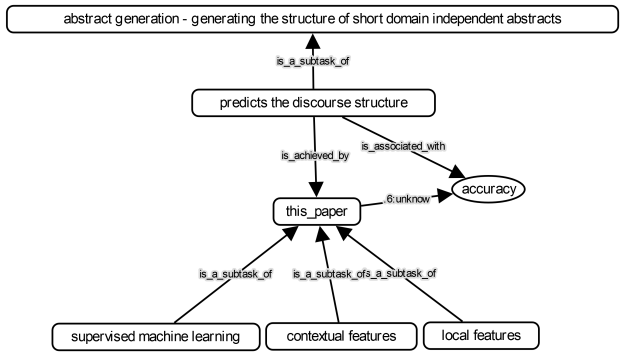



Figure 2: The visual abstraction of an abstract-based annotation.

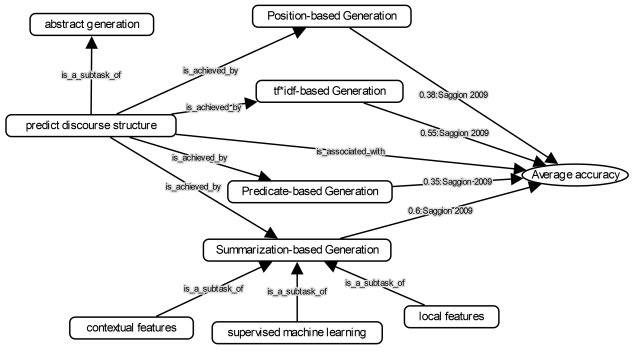



Figure 3: The visual abstraction of a paper-based annotation.

The paper-based map contains a name for the suggested approach, *Summarization-based Generation*, together with names of the different methods against which the suggested method is compared, the results of the comparison and that datasets against which it was tested, in addition to the information provided by the abstract-based concept map, which supports the findings of Westergaard et al. (2018). The annotation process of the paper-based maps is time consuming and requires understanding what the paper is about, but it provides a deeper structural representation of the paper and the way it connects to the main problem.

## 5. Annotated Dataset Statistics

The paper-based dataset contains 2 clusters of papers: 18 papers on Machine Translation taken from Choi et al. (2016)'s cluster, and 12 papers on summarization taken from Jha et al. (2015)'s cluster on Summarization. The abstract-based dataset consists of 42 papers in Ontology Alignment. Table 3 contains statistics about the annotations in comparison to similar datasets.

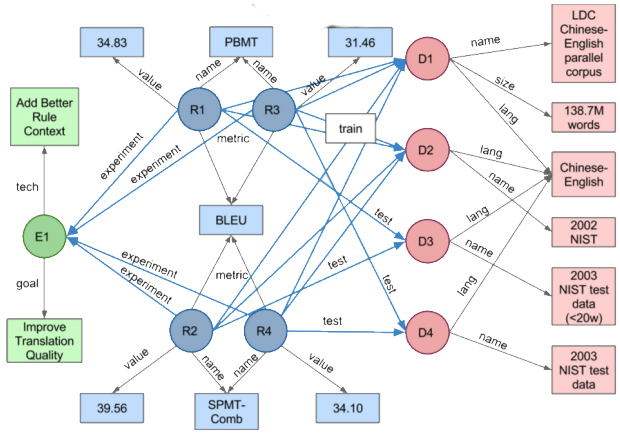
| Dataset | #Concepts | #Unique Concepts | #Relations |
|---|---|---|---|
| Tateisi et al. (2016) | 27350 | - | 24511 |
| Augenstein et al. (2017) | 5730 | 1697 | 643 |
| QasemiZadeh and Schumann (2016) | 4849 | 3318 | - |
| Choi et al. (2016) | 1063 | - | ∼959 |
| Gábor et al. (2016) | - | - | 100 |
| Abstract-based Dataset | 487 | 487 | 532 |
| Abstract-based Dataset BRAT Format | 729 | 487 | 431 |
| Paper-based Dataset | 540 | 540 | 1109 |
| Paper-based Dataset BRAT Format | 1277 | 540 | 455 |
| Total | 1027 | 1027 | 1641 |
| Total BRAT format | 2006 | 1027 | 886 |

Table 3: Statistics regarding the annotations of the dataset in comparison to similar datasets.

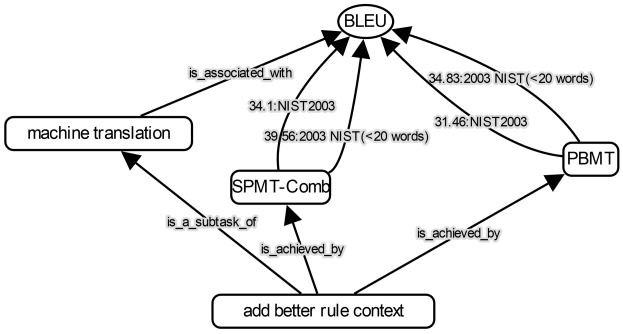
## 6. Conclusion

In this paper, we present a generalized multi-domain goal-oriented representation for scientific papers that models problems, solutions, methods, results, and trade-offs described in scientific papers and connects them to the general domain addressed by papers. Considering the full paper when looking for these elements gives a fuller more precise representation of the paper than abstract-based annotation. The schema leads to acceptable inter-annotator agreement and has been tested over 3 distinct domains to assess its generality. The annotated dataset is available online `http://jumanan.github.io/`.

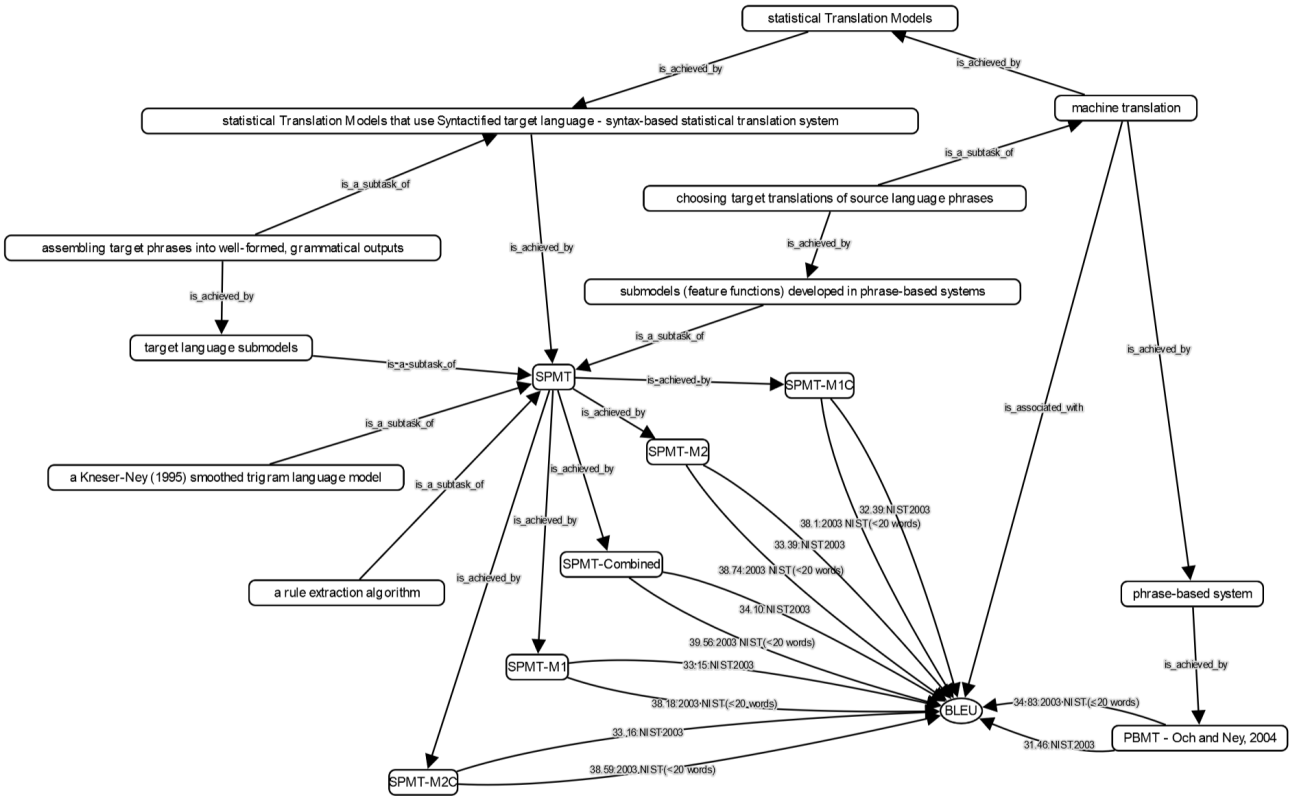
## Appendix




(a) A graph representation for a paper as created by Choi et al. (2016).




(b) The result of automatically converting the same paper presented in 4a to our schema.




(c) The same paper represented in our schema after a paper-based manual annotation.

Figure 4: Different representation of the same paper for comparison.

| Dataset | #Documents | Document Type | #Concepts | #Relations |
|---|---|---|---|---|
| Tateisi et al. (2016) English | 390 | abstracts | 12 | 20 |
| Augenstein et al. (2017) | 500 | paragraphs | 3 | 2 |
| QasemiZadeh and Schumann (2016) | 300 | abstracts | 7 | - |
| Choi et al. (2016) | 67 | experimental results | 8 | 3 |
| (Gábor et al., 2016) | 100 | abstracts and introductions | - | 21 |
| Abstract-based Dataset | 42 | abstracts | 2 | 4 |
| Paper-based Dataset | 30 | relevant paragraphs | 2 | 4 |

Table 4: A comparative table of the different annotation schemes.

## 7. Bibliographical References


Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555.

Choi, E., Horvat, M., May, J., Knight, K., and Marcu, D. (2016). Extracting structured scholarly information from the machine translation literature.

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2015). Cytoscape. js: a graph theory library for visualisation and analysis. *Bioinformatics*, 32(2):309–311.

Gábor, K., Zargayouna, H., Buscaldi, D., Tellier, I., and Charnois, T. (2016). Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *LREC*.

Jha, R., Coke, R., and Radev, D. (2015). Surveyor: A system for generating coherent survey articles for scientific topics. *Ann Arbor*, 1001:48109.

Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263.

Maraee, A. and Sturm, A. (2017). Formal semantics and analysis tasks for me-map models. In *Research Challenges in Information Science (RCIS), 2017 11th International Conference on*, pages 234–243. IEEE.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). Spmt: Statistical machine translation with syntactified target language phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52. Association for Computational Linguistics.

Novak, J. D. and Gowin, D. B. (1984). *Learning how to learn*. Cambridge University Press.

QasemiZadeh, B. and Schumann, A.-K. (2016). The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *LREC*.

Saggion, H. (2009). A classification algorithm for predicting the structure of summaries. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 31–38. Association for Computational Linguistics.

Severo, B., Trojahn, C., and Vieira, R. (2014). Voar: A visual and integrated ontology alignment environment.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Sturm, A., Gross, D., Wang, J., and Yu, E. S. K. (2017a). Means-ends based know-how mapping. *J. Knowledge Management*, 21(2):454–473.

Sturm, A., Yu, E., and Abrishamkar, S. (2017b). Know-how mapping–a goal-oriented approach and evaluation. In *Enterprise, Business-Process and Information Systems Modeling*, pages 272–286. Springer.

Tateisi, Y., Ohta, T., Pyysalo, S., Miyao, Y., and Aizawa, A. (2016). Typed entity and relation annotation on computer science papers. In *LREC*.

Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L. J., and Brunak, S. (2018). A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS computational biology*, 14(2):e1005962.