SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang Qasemizadeh, Haïfa Zargayouna, Thierry Charnois

SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers

Kata Gábor[1], **Davide Buscaldi**[1], **Anne-Kathrin Schumann**[2], **Behrang QasemiZadeh**[3], **Haïfa Zargayouna**[1], **Thierry Charnois**[1]
[1] LIPN, CNRS (UMR 7030), Université Paris 13
firstname.lastname@lipn.univ-paris13.fr
[2] ProTechnology GmbH, Dresden
annek_schumann@gmx.de
[3] DFG SFB 991, Heinrich-Heine University, Dsseldorf
zadeh@phil.uni-duesseldorf.de

Abstract

This paper describes the first task on semantic relation extraction and classification in scientific paper abstracts at SemEval 2018. The challenge focuses on domain-specific semantic relations and includes three different subtasks. The subtasks were designed so as to compare and quantify the effect of different pre-processing steps on the relation classification results. We expect the task to be relevant for a broad range of researchers working on extracting specialized knowledge from domain corpora, for example but not limited to scientific or bio-medical information extraction. The task attracted a total of 32 participants, with 158 submissions across different scenarios.

1 Introduction

One of the emerging trends of natural language technologies is their use for the humanities and sciences. Recent works in the semantic web (Osborne and Motta, 2015; Wolfram, 2016) and natural language processing (Tsai et al., 2013; Luan et al., 2017; Augenstein and Søgaard, 2017; Kim et al., 2010) aimed to improve the access to scientific literature, and in particular to respond to information needs that are currently beyond the capabilities of standard search engines. Such queries include finding all papers that address a problem in a specific way, or discovering the roots of a certain idea. This ambition involves the identification and classification of concepts, and the relations connecting them.

The purpose of the task is to automatically identify relevant domain-specific semantic relations in a corpus of scientific publications. In particular, we search for and classify relations that provide snippets of information such as "a (new) method is proposed for a task", or "a phenomenon is found in a certain context", or "results of different experiments are compared to each other". Identifying such semantic relations between domain-specific concepts allows us to detect research papers which deal with the same problem, or to track the evolution of results on a certain task.

2 Related Work

SemEval 2010 Task 8 (Hendrickx et al., 2010) proposed a discrete classification of word pairs into 9 semantic relations, however, this task was not tailored to the needs of scientific text analysis as neither relation types nor the vocabulary were domain-specific. SemEval 2012 Task 2 (Jurgens et al., 2012) proposed a gradual notion of relational similarity: the task was to quantify the similarity between examples of relation instances. The data set was aimed at evaluating specific semantic representations for relational similarity, but does not fit our task: in this task, entity pairs were treated as static class instances; in particular, they were presented without any context. However, the relation types we deal with are contextual: e.g., a specific *machine learning method* is trained on a specific *data set* to perform an *NLP task* in the context of a given experiment reported by a paper. Finally, the most closely related to our task is SemEval 2017 Task 10 (Augenstein et al., 2017), which responds to the growing interest towards the semantic analysis of scientific corpora. This task focuses mostly on keyword extraction and categorization. The subtask concerned with relation classification proposes 3 categories of taxonomic relations (synonym, hypernym, unrelated). Our task goes a step further by proposing a more fine-grained and, thus, more informative set of semantic relations (see Table 1). The relation types were selected and annotated based on a careful corpus study and are intended to represent the major re-

Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 679–688
New Orleans, Louisiana, June 5–6, 2018.

lations that define the information content of the abstract of a scientific paper.

3 Task description

The task consists in identifying and classifying instances of semantic relations between concepts in a set of 6 discrete categories. The relations are specific to the science domain and their instances can frequently be found in the abstract/introduction of scientific papers. The task is split into three subtasks. This is done to provide a framework for the systematic evaluation of the steps that are necessary for full information extraction from scientific text, i. e. relation extraction and relation classification. Two of the subtasks focus solely on the *classification* of relation instances into 6 relation categories. Another subtask includes both the *extraction* of relation instances and their *classification*. The data we provide is presented as complete abstracts of scientific papers. An abstract contains about 100 words on average. Entities are annotated in both the training and the test data. Furthermore, in the classification subtasks, the relation instances (entity pairs that belong to one of the relation classes) as well as the directionality of the relation (argument1, argument2) are given in the training and test data. In the extraction subtask, relation instances are not provided in the test data. The training data for each subtask contains 350 annotated abstracts with the corresponding relation instances and their categories[1]. The test data consists of 150 abstracts[2]. Participants were allowed three submissions/subtask/team.

3.1 Relation classification scenario

Given a pair of entities in an abstract, the task consists in classifying the semantic relation between them. A pre-defined list of relations is given (see Table 1), together with training examples for each relation.

- **Subtask 1.1 : Relation classification on clean data.**
  Entity occurrences are *manually* annotated in both the training and the test data. In the training data, semantic relations are manually annotated between entities. In the test data, only entity annotations and unlabeled relation instances are given. The task is to predict the semantic relation between the entities. The following example shows a text snippet with the information provided in the test data :

  > Korean, a <entity id="H01-1041.10">verb final language</entity>with <entity id="H01-1041.11">overt case markers</entity>(...)

  A relation instance is identified by the unique identifier of the entities in the pair, e.g. (H01-1041.10, H01-1041.11). The information to be predicted is the relation class label: MODEL-FEATURE(H01-1041.10, H01-1041.11).

- **Subtask 1.2 : Relation classification on noisy data.**
  Entity occurrences are *automatically* annotated in both the training and the test data. Delimitation errors may occur in the entity annotation. In the training data, semantic relations are manually annotated between the entities. In the test data, only automatic entity annotations and unlabeled relation instances are given. The task is to predict the semantic relation between the entities. The following example shows a text snippet with the information provided in the test data:

  > This <entity id="L08-1203.8"> paper </entity> introduces a new <entity id="L08-1203.9">architecture</entity>(...)

  The relation instance is (L08-1203.8, L08-1203.9). The information to be predicted is the relation class label: TOPIC(L08-1203.8, L08-1203.9)

3.2 Relation extraction and classification scenario

Given an abstract with annotated entities, the subtask consists in:

- identifying instances of semantic relations between entities in the same sentence,
- assigning class labels, i.e. one of six predefined relation types (see Table 1), to the relation instances.

[1]The training data for subtask 1.1 and subtask 2 were identical.

[2]After the end of the competition, the complete dataset was published at https://lipn.univ-paris13.fr/~gabor/semeval2018task7/

RELATION TYPE	Explanation	Example
USAGE	*Methods, tasks, and data are linked by usage relations.*	
used_by	ARG1: *method, system* ARG2: *other method*	approach – model
used_for_task	ARG1: *method/system* ARG2: *task*	approach – parsing
used_on_data	ARG1: *method* applied to ARG2: *data*	MT system – Japanese
task_on_data	ARG1: *task* performed on ARG2: *data*	parse – sentence
RESULT	*An entity affects or yields a result.*	
affects	ARG1: *specific property of data* ARG2: *results*	order – performance
problem	ARG1: *phenomenon* is a problem in a ARG2: *field/task*	ambiguity – sentence
yields	ARG1: *experiment/method* ARG2: *result*	parser – performance
MODEL	*An entity is a analytic characteristic or abstract model of another entity.*	
char	ARG1: *observed characteristics* of an observed ARG2: *entity*	order – constituents
model	ARG1: *abstract representation* of an ARG2: *observed entity*	interpretation – utterance
tag	ARG1: *tag/meta-information* associated to an ARG2: *entity*	categories – words
PART_WHOLE	*Entities are in a part-whole relationship.*	
composed_of	ARG2: *database/resource* ARG1: *data*	ontology – concepts
datasource	ARG1: *information* extracted from ARG2: *kind of data*	knowledge – domain
phenomenon	ARG1: *entity, a phenomenon* found in ARG2: *context*	expressions – text
TOPIC	*This category relates a scientific work with its topic.*	
propose	ARG1: *paper/author* presents ARG2: *an idea*	paper – method
study	ARG1: *analysis* of a ARG2: *phenomenon*	research – speech
COMPARISON	*An entity is compared to another entity.*	
compare	ARG1: *result, experiment* compared to ARG2: *result, experiment*	result – standard

Table 1: Semantic relation typology. The six major relation types result from a finer grained classification which was used in manual annotation.

The training data we provide contains the same information as in the classification scenario, i.e. manually annotated entities, and labeled semantic relations holding between entities. The test data contains only abstracts with annotated entities: both the entity pairs and their relation type are to be predicted.

3.3 Evaluation

Submissions are evaluated differently for the individual subtasks. A dedicated gold standard containing entity and relation annotations is used.

3.3.1 Metrics for the classification scenario (subtasks 1.1 and 1.2)

Submissions for scenario 1 are assessed by means of standard metrics:

- **Class-wise evaluation**: Precision, recall, and F1 ($\beta = 1$) for each relation type.
- **Global evaluation**:
  - Macro-average of the F1 scores obtained for every relation type.
  - Micro-average of the F1 scores obtained for every relation type.

The official ranking of submissions is performed according to the **macro-average F1 score**.

3.3.2 Metrics for the extraction and classification scenario (Subtask 2)

Evaluation of submissions for scenario 2 is carried out in two steps:

- **Evaluation of relation extraction**: Extraction evaluation assesses the quality of identified relation instances. Relation labels and directionality are ignored in this step. *Precision* is calculated as the percentage of correctly connected entity pairs. *Recall* is calculated as the percentage of gold entity pairs found by the system. The official **F1 score** is calculated as the harmonic mean of precision and recall.

- **Evaluation of relation classification**: Classification evaluation considers only correctly identified relation instances as per step 1. For these instances, the same evaluation metrics are calculated as for task 1. The official score for this task is **macro-average F1**.

4 Data Preparation

The task is carried out on abstracts from published research papers in computational linguistics. Two existing high-quality corpora were used as starting points for data creation, namely ACL RD-TEC 2.0 (QasemiZadeh and Schumann, 2016) and ACL-RelAcS (Gábor et al., 2016a). Both resources are based on the ACL Anthology Reference Corpus (Bird et al., 2008). In ACL RD-TEC 2.0 entities were annotated manually, and it was used for the "clean" subtasks (subtasks 1.1 and 2). In ACL-RelAcS, entities were annotated fully automatically, and it was used for the "noisy" subtask (1.2).

4.1 Entity annotation

Manual ("clean") entity annotations were carried out in accordance with the ACL RD-TEC annotation guidelines (Schumann and QasemiZadeh, 2015). Thus, for subtasks 1.1 and 2 (training data) termhood is defined by a combination of semantic, linguistic, and formal criteria. The formal criteria, for instance, aim at making the annotations maximally useful for real-world extraction scenarios by accounting for various contextual usage patterns of terminological units in scientific prose. Therefore, annotators are instructed to annotate maximal noun phrases, abbreviations, and their contextual variants, including variants with incorrect spelling. Still, entity annotation proves to be a non-trivial task even for human expert annotators: Qasemi-Zadeh and Schumann (2016) show that agreement scores are satisfactory (e.g., $\kappa > 0.7$) only after a thorough annotation training phase and the subsequent refinement of the annotation guidelines.

To extend the set of abstracts that were already available in ACL RD-TEC with double entity annotations, expert annotators were recruited from amongst the task organizers. Annotators were asked to read the ACL RD-TEC annotation guidelines. A training phase was carried out, during which each annotator carried out test annotations on unseen data. To facilitate annotations, abstracts were pre-annotated automatically using the automatic entity annotator of the ACL-RelAcS corpus (see below). Annotators were asked to correct the automatic annotations, in particular, to correct the boundary of the identified entity. Individual feedback was provided to novice annotators and annotation difficulties were clarified. Annotations were consistently monitored and potential causes for disagreement discussed and corrected.

The ACL RD-TEC already provided 171 double-annotated and 129 single-annotated abstracts. While double-annotations could directly be passed over to manual relation annotation, more single-pass annotations had to be performed to create a fully double-annotated training set. The remaining 150 abstracts for the test set of subtask 1.1 were single-annotated. It should be noted that, due to their origin from ACL RD-TEC, abstracts for subtask 1.1 contain not only entity annotations, but also information about the the semantic class of the annotated entity. This information was not explicitly included in the provided data, but was accessible to participants through the original ACL RD-TEC corpus.

The "noisy" subtask (1.2) was carried out on data coming from the ACL-RelAcS corpus 1.0 (Gábor et al., 2016a). The corpus consists of 4.2 million words from the abstract and introduction sections of papers in the ACL Anthology Corpus, with an automatic annotation of entities. This automatic annotation is based on a gazetteer which, in turn, was created using a combination of terminology extraction tools and ontological resources. As a domain specific resource, the domain models and topic hierarchies in the NLP domain from Saffron Knowledge Extraction Framework[3] (Bordea, 2013; Bordea et al., 2013) were included. Terminology extraction was performed with TermSuite (Daille et al., 2013) and the resulting list of terms was filtered by part of speech and looked up in BabelNet (Navigli and Ponzetto, 2012). The extracted terms that were found in BabelNet were added to the gazetteer and used for automatic annotation.

4.2 Relation annotation

The work was divided as 1) defining the typology of semantic relations, 2) validation of the typology and of the annotation guidelines and 3) annotation. A data-driven approach was adopted to identify the relation types and define a typology (Gábor et al., 2016b). Domain experts studied the abstracts with entity annotation and were instructed to read the text and indicate the semantic relations that are explicit and relevant for the understanding of the abstract. They annotated entity pairs and the text span between them which explicitly indicates the relation.

Instances of explicit relations were thus discovered and manually annotated in a sample of 100 abstracts from ACL-RelAcS. A fine-grained typology of domain-specific relations was set up. The fine-grained relation types (see Table 1) were defined very precisely and specifically, e.g. using strict constraints on which types of entities the relations take as argument. The manual annotation used this typology; the relations were then automatically converted to the 6 types used in the classification tasks.

Only explicit relations were annotated, between already annotated entities. Entity annotation itself is never modified or corrected manually during the relation annotation phase. On the textual level, a semantic relation is conceived as a text span link-

[3] http://saffron.insight-centre.org/

ing two annotated instances of concepts within the same sentence. On the semantic level, relation types need to be specific enough to be easily distinguished from each other by a domain expert. Annotation was carried out by one of the organizers and two NLP student annotators who were subjected to a training of three weeks during which they annotated 100 abstracts under supervision. This training material was not included in the future dataset. Weekly feedback was given and difficult instances were discussed. If the annotation quality in the 100 abstracts was judged satisfactory, the annotator was allowed to carry on, and their subsequent annotations were included in the dataset (two out of three annotator candidates passed this phase).

Inter-annotator agreement was calculated using a double annotation on a sample of 150 abstracts from subtask 1.1 by two annotators. The overall class label agreement rate on these annotations was 90.8%. We also calculated the macro-averaged F1 score across classes, taking one of the annotators as "gold standard". The result was 0.91 (the performance of the best ranking system on this task is 0.81). When comparing agreement for individual relations, it turns out that the relation with the lowest agreement (F1=0.83) is PART_WHOLE, followed by RESULT (F1=0.89).

5 Results

5.1 Baseline system

As a baseline, we created a simple memory-based k-nearest neighbor (k-nn) search (Daelemans and van den Bosch, 2005) which relies on a small set of hand-crafted features.

Given a sentence s annotated with an ordered set of $e_1 \dots e_n$ entities appearing in it, we first pull out all tuples (e_i, e_j), in which $j - i \leq 2$. For each tuple (e_i, e_j), we encode their co-occurrence context using a set of 5 vectors of low dimensionality ($n = 100$). These vectors encode information about (a) tokens that appear before e_i in s (we use simple white-space tokenization), (b) tokens that appear between e_i and e_j, (c) tokens appearing after e_j, as well as (d) two additional vectors that capture the context of e_i and e_j occurrences in the ACL Anthology Reference Corpus (Bird et al., 2008). To encode information about these context-token occurrences into low-dimensional vectors, we use positive-only random projections (Qasemi-Zadeh and Kallmeyer, 2016). Additionally, feature vectors in each of the above-mentioned categories are weighted using positive pointwise mutual information with respect to the collected co-occurrence information in vectors for each category for all the tuples in the training and test data (for each subtask). Finally, the weighted vectors are concatenated to form a 500 dimensional feature vector for each entity pair.

For each subtask, all the (e_i, e_j) extracted from the sentences in the training set are added to the k-nn's training instance memory T: if (e_i, e_j) is annotated with a relation, then the fetched label is assigned to it, otherwise it is marked as a negative example. Given the feature vector $\vec{v}$ for a tuple (e_x, e_y) in the test set, the similarity between $\vec{v}$ and all the training instances $t_i \in T$ is computed using the Pearson's correlation to find the k *most similar* t_i. Finally, we assign (e_x, e_y) to the relation category l_y using a majority voting.

Results obtained from this baseline system are listed in Tables 5, 7, 6, and 8 in the Appendix. We choose $k = 5$ based on the observed performances over the development dataset.

5.2 Summary of participating systems and results

The task attracted 32 participants altogether who took part in at least one subtask. The most popular subtask was the classification on clean data (subtask 1.1) with 28 participants; 19 of them also participated in the classification on noisy data (subtask 1.2). One participant chose to compete only in subtask 1.2. Subtask 2 attracted 11 teams. The scenario allowed to compete only in relation extraction, without classifying the extracted instances; only one team used this opportunity. The complete results and rankings are available in the Appendix section. Most participants opted for the use of deep learning methods, with a clear preference for Convolutional Neural Networks (CNN) which were used by 10 systems, and Long Short Term Memory (LSTM) networks, used by 5 systems. Support Vector Machines (SVM) were the preferred non-DL method, used by 5 systems. One participant (Bf3R) opted for a combination of existing tools in Subtask 2. In Figure 1 and 2 we show an overview of the number of methods chosen by participants and the average results obtained by each family of methods for each subtask. The average was calculated on all submissions. The number of systems doesn't necessarily match the number of participants (some participants tested different methods). Most par-

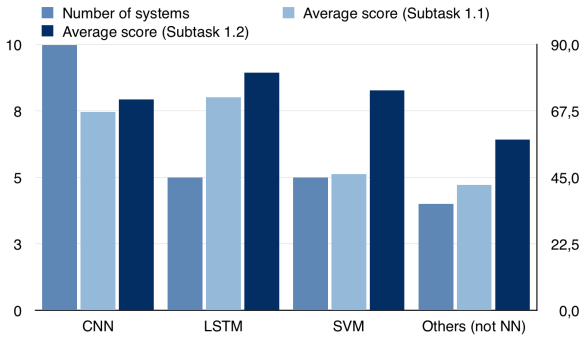

Figure 1: Popularity of methods chosen by participants (as number of systems that used the method, left) and average F1 score obtained for each method (right) in Subtask 1.1 and 1.2.

ticipants exploited the possibility of aggregating training data from subtask 1.1 and subtask 1.2.

Word embeddings were used as features by the majority of systems (13 systems). Some participants chose to calculate the embeddings on domain-specific corpora, such as ACL (4 systems) and arXiv (3 systems), sometimes in combination with pre-trained embeddings. Pre-trained embeddings alone were used by a minority of participants, with TakeLab highlighting some problems in dealing with out-of-vocabulary words. Apart from the corpora dedicated to training the embeddings, participants didn't use external resources, with the exception of one system which employed VerbNet and two systems that used WordNet synonyms and hypernyms.

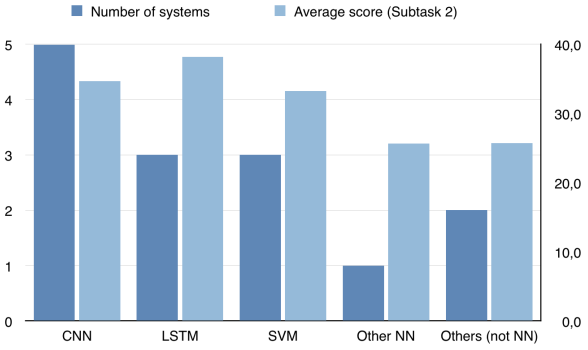

Figure 2: Popularity of methods chosen by participants (as number of systems that used the method, left) and average F1 score obtained for each method (right) in Subtask 2.

Among the chosen features, positional embeddings were quite popular (5 systems), to account for the relative position of the left and right entities. Only three participants recurred to syntactic features, in particular dependency trees, despite their apparent relevance for the task.

SpaCy[4] and CoreNLP[5] were the most popular tools to analyze and preprocess text, with a slight preference for the first one (4 participants vs. 2).

6 Analysis of Results

6.1 Which processing step is the most difficult?

From the overall task results provided in the Appendix (Tables 5 – 8), it seems straightforward to conclude that the reliable identification of semantic relation instances is by far the most difficult step in the complete processing pipeline: Whereas systems reached an average F1 score of 47.28 in subtask 1.1 and 62.51 in subtask 1.2, performance scores drop rather sharply in scenario 2, namely to an average F1 of 30.8 for the extraction task and 20.34 for the extraction+classification task.

6.2 Which relation types are the most difficult to classify?

We examined whether there were relation types that were more difficult for the systems to classify, and whether it is possible to relate this to the semantics of the relations. For instance, the class MODEL-FEATURE is broad because it encompasses relatively different sub-classes: models, parts of models (such as a representation, a tag used for a word), or attributes (frequency of a phenomenon). To analyze this, we calculated the average recall by relation type over a sample of submissions to subtask 1.1 (70 submissions) and 1.2 (42 submissions) and the characteristic prediction error types by relation, if any (Table 2). We also calculated the average F1 score by relation type of the five top scoring systems from different participants (Tables 3 and 4).

Our analysis suggests that rather than the semantics of the relation types, it is their distribution in the data that poses difficulties. Class distribution is very imbalanced. Moreover, the distribution of classes in training and test data of subtask 1.1 and 1.2 is different. This difference is due to the nature of entities annotated automatically and those annotated manually. Because of the terminology extraction process and the resources that were used

[4] https://spacy.io/

[5] https://stanfordnlp.github.io/CoreNLP/

Relation	Average recall	Frequently mistaken for	Training frequency	Test frequency
Subtask 1.1				
USAGE	73%	MODEL-FEATURE	483	175
TOPIC	66%	-	18	3
MODEL-FEATURE	51%	USAGE, PART_WHOLE	326	66
PART_WHOLE	44%	USAGE, MODEL-FEATURE	234	70
COMPARE	42%	USAGE	95	21
RESULT	40%	USAGE	72	20
Subtask 1.2				
TOPIC	77%	-	243	69
USAGE	72%	PART_WHOLE	470	123
COMPARE	66%	-	41	3
RESULT	65%	USAGE	123	29
PART_WHOLE	64%	USAGE	196	56
MODEL-FEATURE	52%	USAGE, PART_WHOLE	175	75

Table 2: Relations: results and distribution.

for annotation, entities in subtask 1.2 are typically shorter terms with an intermediate level of specificity. On the other hand, entities in the clean scenario are more complex and more specific to the NLP domain. For instance, the TOPIC relation is more frequent in 1.2 than in 1.1 because entities like ”paper” or ”article” were annotated by the automated process, but not in the manual annotation.

Another aspect is that certain classes are lexically less varied than others and this might well affect the ”difficulty” of the classification task. For instance, the TOPIC class has the lowest type-token ratio of all classes in subtask 1.2[6]. This does not seem surprising. Neither does it seem surprising that in subtask 1.2, TOPIC has gained the best average recall (2) and the highest F1 score among the top-5 systems (4). TOPIC is also much more frequent in subtask 1.2 than in subtask 1.1 and this effect is one likely cause for the difference in performance achieved over subtasks 1.1 and 1.2.

Relation	Top 5 Average F1
USAGE	0.85
RESULT	0.75
PART_WHOLE	0.73
TOPIC	0.71
MODEL-FEATURE	0.69
COMPARE	0.59

Table 3: Relations: Task 1.1 average results top 5 systems.

6.3 The effects of entity annotation

Entity annotation has a demonstrable effect on system performance. As stated earlier, annotation decisions have direct consequences for the distribution of certain types in the data and thus influence measurable system performance.

[6]In this analysis, a tuple of two entities pertaining to a certain relation class was counted as a ”type”.

Relation	Top 5 Average F1
TOPIC	0.97
RESULT	0.91
USAGE	0.87
COMPARE	0.80
PART_WHOLE	0.80
MODEL-FEATURE	0.79

Table 4: Relations: Task 1.2 average results top 5 systems.

A maybe rather surprising result of this task is the difference in system performance for subtasks 1.1 and 1.2. While ”clean” entities can, with some plausibility, be considered more useful for a potential human user of the extracted information, ”noisy” entity annotations seem to be more machine-friendly. The difference in the distribution of the TOPIC relation between subtasks 1.1 and 1.2 has already been pointed out as one potential cause for this effect. Moreover, the complexity of clean entities in subtask 1.1 could also have contributed to the performance gap. Manually annotated entities, in most cases, are long noun phrases, whereas automatically annotated entities in subtask 1.2 are generally shorter, partial (and therefore less specific!) entity matches. This also means that more training examples are likely to be found for automatically annotated entities. Moreover, some instances of automatic annotations in subtask 1.2 included explicit verbal relation cues. These cues sometimes explicitly state the type of the semantic relation, but they were not annotated in subtask 1.1. Verbal cues (e. g. the well-known Hearst patterns (Hearst, 1992)) have typically been used in earlier work on relation classification and, in fact, several teams participating in the task describe recurrent verbal elements between relation arguments.

The role of the specialized lexicon in relation extraction and classification is a topic that de-

serves further exploration for the following reasons: Firstly, highly specialized, complex terminological units are the main units of knowledge representation in specialized domains. Secondly, task results clearly show that a careful handling of lexical information improves performance: many successful systems in the task used domain-specific training data. The only system that treated complete specialized entities as semantic units, UWNLP, ranked first in the relation extraction task. None of the systems participating in subtasks 1.1 or 2 used semantic class information available for annotated entities from ACL RD-TEC, although it may be hypothesized that this feature helps to generalize lexical instance information.

7 Conclusion and Future Work

We presented the setup and results of SemEval 2018 Task 7: Semantic relation extraction and classification in scientific papers. The task is divided into three subtasks: classification on clean data, classification on noisy data, and a combined extraction and classification scenario. We also presented the dataset used for the challenge: a subset of abstracts of published papers in the ACL Anthology Reference Corpus, annotated for domain specific entities and semantic relations.

32 participants submitted to one or more subtasks. The most popular methods include Convolutional Neural Networks and Long Short Term Memory networks, with word embedding based features, often calculated on domain-specific corpora. Although it was allowed, only a minority of the participants used external knowledge resources. The results show that while good results can be obtained on the supervised multi-class classification of relation instances, the extraction of such instances remains very challenging. Moreover, the quality and type of entity annotation also plays an important role in determining relation extraction and classification results.

Knowledge extraction from a special domain poses specific challenges, such as working with a smaller corpus, dealing with specialized vocabularies, and the scarcity of annotated data and available domain-specific resources. One of the important future directions is to explore domain adaptation techniques to address these issues.

Acknowledgments

This work was generously supported by the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL). Behrang QasemiZadeh is funded by the Deutsche Forschungsgemeinschaft through the "Collaborative Research Centre 991 (CRC 991): The Structure of Representations in Language, Cognition, and Science".

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 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, Vancouver, Canada. Association for Computational Linguistics.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association.

Georgeta Bordea. 2013. *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. Phd thesis, National University of Ireland, Galway.

Georgeta Bordea, Paul Buitelaar, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge University Press.

Beatrice Daille, Christine Jacquin, Laura Monceaux, Emmanuel Morin, and Jerome Rocheteau. 2013. TTC TermSuite : Une chaine de traitement pour la fouille terminologique multilingue. In *Proceedings of the Traitement Automatique des Langues Naturelles Conference (TALN)*.

Kata Gábor, Hafa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016a. Semantic annotation of the ACL Anthology Corpus

for the automatic analysis of scientific literature. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 3694–3701.

Kata Gábor, Hafa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois. 2016b. A typology of semantic relations dedicated to scientific literature analysis. In *SAVE-SD Workshop at the 25th World Wide Web Conference. Lecture Notes in Computer Science 9792*, pages 26–32.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING '92*, pages 539–545.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Saghdha, Sebastian Pad, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations (SemEval-2010)*.

David A. Jurgens, Peter D. Turney, Saif M. Mohammad, and Keith J. Holyoak. 2012. SemEval-2012 Task 2: Measuring degrees of relational similarity. In *Proceedings of the Workshop on Semantic Evaluations*.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of SemEval 2010*.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *EMNLP 2017*.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

F. Osborne and E. Motta. 2015. Klink-2: Integrating multiple web sources to generate semantic topic networks. In *Proceedings of the 14th International Conference on The Semantic Web - ISWC 2015 - Volume 9366*, pages 408–424, New York, NY, USA. Springer-Verlag New York, Inc.

Behrang QasemiZadeh and Laura Kallmeyer. 2016. Random Positive-Only Projections: PPMI-enabled incremental semantic space construction. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 189–198, Berlin, Germany. Association for Computational Linguistics.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Anne-Kathrin Schumann and Behrang QasemiZadeh. 2015. The ACL RD-TEC Annotation Guideline: A Reference Dataset for the Evaluation of Automatic Term Recognition and Classification. Technical report.

Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-based analysis of scientific literature. In *ACM Conference on Information and Knowledge Management ACM*, pages 1733–1738.

Dietmar Wolfram. 2016. Bibliometrics, information retrieval and natural language processing: Natural synergies to support digital library research. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)*, pages 6–13.

Appendix: Competition Results

Rank	Participant	Macro-F1 Score
1	ETH-DS3Lab	81.7
2	UWNLP	78.9
3	SIRIUS-LTG-UiO	76.7
4	ClaiRE	74.9
5	Talla	74.2
6	MIT-MEDG	72.7
7	TakeLab	69.7
8	Texterra	64.9
9	GU IRLAB	60.9
10	sbuNLP	49.7
11	IRCMS	49.1
12	OhioState	48.1
13	NTNU	47.4
14	danish037	45.7
15	HeMu	45.2
16	UniMa	44.0
17	LaSTUS/TALN	43.2
18	LIGHTREL	39.9
19	LTRC	37.3
N/A	*Baseline*	*34.4*
20	BIT_NLP	32.9
21	likewind_1234	29.3
22	Vitk	29.0
23	hccl	28.1
24	xingwang	27.8
25	SciREL	20.3
26	UKP	19.3
27	NEUROSENT-PDI	18.0
28	angelocsc	15.0

Table 5: Results for subtask 1.1.

Rank	Participant	F1 Score
1	UWNLP	50.0
2	ETH-DS3Lab	48.8
3	SIRIUS-LTG-UiO	37.4
4	UC3M-NII	35.4
5	NTNU	33.9
6	Bf3R	33.4
7	UniMa	28.4
N/A	*Baseline*	*26.8*
8	NEUROSENT-PDI	25.6
9	Texterra	15.6
10	xingwang	15.3
11	danish037	15.0

Table 6: Results for subtask 2: Extraction.

Rank	Participant	Macro-F1 Score
1	ETH-DS3Lab	90.4
2	Talla	84.8
3	SIRIUS-LTG-UiO	83.2
4	MIT-MEDG	80.6
5	GU IRLAB	78.9
6	ClaiRE	78.4
7	TakeLab	75.7
8	OhioState	74.7
9	Texterra	74.4
10	IRCMS	71.1
11	LaSTUS/TALN	69.5
12	LIGHTREL	68.2
13	NTNU	66.0
14	LTRC	65.7
N/A	*Baseline*	*53.5*
15	likewind_1234	45.8
16	BIT_NLP	40.7
17	hccl	38.0
18	xingwang	26.7
19	NEUROSENT-PDI	21.8
20	UKP	15.3

Table 7: Results for subtask 1.2.

Rank	Participant	Macro-F1 Score
1	ETH-DS3Lab	49.3
2	UWNLP	39.1
3	SIRIUS-LTG-UiO	33.6
4	Bf3R	20.3
5	UC3M-NII	18.5
6	NTNU	17.0
N/A	*Baseline*	*12.6*
7	Texterra	9.6
8	xingwang	8.3
9	danish037	4.6
10	NEUROSENT-PDI	3.1

Table 8: Results for subtask 2: Extraction + Classification.