SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications

Isabelle Augenstein[1], Mrinal Das[2], Sebastian Riedel[1], Lakshmi Vikraman[2], and Andrew McCallum[2]
[1]Department of Computer Science, University College London (UCL), UK
[2]College of Information and Computer Sciences, University of Massachusetts Amherst, USA

Abstract

We describe the SemEval task of extracting keyphrases and relations between them from scientific documents, which is crucial for understanding which publications describe which processes, tasks and materials. Although this was a new task, we had a total of 26 submissions across 3 evaluation scenarios. We expect the task and the findings reported in this paper to be relevant for researchers working on understanding scientific content, as well as the broader knowledge base population and information extraction communities.

1 Introduction

Empirical research requires gaining and maintaining an understanding of the body of work in specific area. For example, typical questions researchers face are which papers describe which tasks and processes, use which materials and how those relate to one another. While there are review papers for some areas, such information is generally difficult to obtain without reading a large number of publications.

Current efforts to address this gap are search engines such as Google Scholar,[1] Scopus[2] or Semantic Scholar,[3] which mainly focus on navigating author and citations graphs.

The task tackled here is mention-level identification and classification of keyphrases, e.g. Keyphrase_Extraction (TASK), as well as extracting semantic relations between keywords, e.g. Keyphrase_Extraction HYPONYM-OF Information_Extraction. These tasks are related to the tasks of named entity recognition, named entity classification and relation extraction. However, keyphrases are much more challenging to identify than e.g. person names, since they vary significantly between domains, lack clear signifiers and contexts and can consist of many tokens. For this purpose, a double-annotated corpus of 500 publications with mention-level annotations was produced, consisting of scientific articles of the Computer Science, Material Sciences and Physics domains.

Extracting keyphrases and relations between them is of great interest to scientific publishers as it helps to recommend articles to readers, highlight missing citations to authors, identify potential reviewers for submissions, and analyse research trends over time. Note that organising keyphrases in terms of synonym and hypernym relations is particularly useful for search scenarios, e.g. a reader may search for articles on information extraction, and through hypernym prediction would also receive articles on named entity recognition or relation extraction.

We expect the outcomes of the task to be relevant to the wider information extraction, knowledge base population and knowledge base construction communities, as it offers a novel application domain for methods researched in that area, while still offering domain-related challenges.

Since the dataset is annotated for three tasks dependent on one another, it could also be used as a testbed for joint learning or structured prediction approaches to information extraction (Kate and Mooney, 2010; Singh et al., 2013; Augenstein et al., 2015; Goyal and Dyer, 2016).

Furthermore, we expect the task to be interesting for researchers studying tasks aiming at understanding scientific content, such as keyphrase extraction (Kim et al., 2010b; Hasan and Ng, 2014; Sterckx et al., 2016; Augenstein and Søgaard, 2017), semantic relation extraction (Tateisi et al.,

[1] https://scholar.google.co.uk/
[2] http://www.scopus.com/
[3] https://www.semanticscholar.org/

Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017), pages 546–555,
Vancouver, Canada, August 3 - 4, 2017.

2014; Gupta and Manning, 2011; Marsi and Öztürk, 2015), topic classification of scientific articles (Ó Séaghdha and Teufel, 2014), citation context extraction (Teufel, 2006; Kaplan et al., 2009), extracting author and citation graphs (Peng and McCallum, 2006; Chaimongkol et al., 2014; Sim et al., 2015) or a combination of those (Radev and Abu-Jbara, 2012; Gollapalli and Li, 2015; Guo et al., 2015).

The expected impact of the task is an interest of the above mentioned research communities beyond the task due to the release of a new corpus, leading to novel research methods for information extraction from scientific documents. What will be particularly useful about the proposed corpus are annotations of hypernym and synonym relations on mention-level, as existing hypernym and synonym relation resources are on type-level, e.g. WordNet.[4] Further, we expect that these methods will directly impact industrial solutions to making sense of publications, partly due to the task organisers' collaboration with Elsevier.[5]

2 Task Description

The task is divided into three subtasks:

A) Mention-level keyphrase identification

B) Mention-level keyphrase classification. Keyphrase types are PROCESS (including methods, equipment), TASK and MATERIAL (including corpora, physical materials)

C) Mention-level semantic relation extraction between keyphrases with the same keyphrase types. Relation types used are HYPONYM-OF and SYNONYM-OF.

We will refer to the above subtasks as Subtask A, Subtask B, and Subtask C respectively.

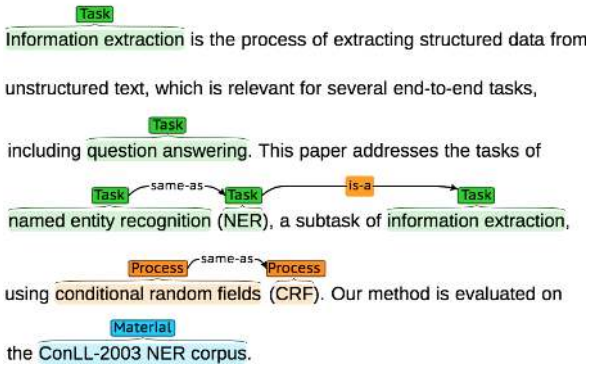
A shortened (artificial) example of a data instance for the Computer Science area is displayed in Example 1, examples for Material Science and Physics are included in the appendix. The first part is the plain text paragraph (with keyphrases in italics for better readability), followed by stand-off keyphrase annotations based on character offsets, followed relation annotations.

Example 1.

Text: *Information extraction* is the process of extracting structured data from unstructured text, which is relevant for several end-to-end tasks, including *question answering*. This paper addresses the tasks of *named entity recognition* (*NER*), a subtask of *information extraction*, using *conditional random fields* (*CRF*). Our method is evaluated on the *ConLL-2003 NER corpus*.

ID	Type	Start	End
0	TASK	0	22
1	TASK	150	168
2	TASK	204	228
3	TASK	230	233
4	TASK	249	271
5	PROCESS	279	304
6	PROCESS	306	309
7	MATERIAL	343	364

ID1	ID2	Type
2	0	HYPONYM-OF
2	3	SYNONYM-OF
5	6	SYNONYM-OF



3 Resources for SemEval-2017 Task

3.1 Corpus

A corpus for the task was built from ScienceDirect[6] open access publications and was available freely for participants, without the need to sign a copyright agreement. Each data instance consists of one paragraph of text, drawn from a scientific paper.

Publications were provided in plain text, in addition to xml format, which included the full text of the publication as well as additional metadata. 500 paragraphs from journal articles evenly distributed among the domains Computer Science, Material Sciences and Physics were selected.

[4] https://wordnet.princeton.edu/
[5] https://www.elsevier.com/
[6] http://www.sciencedirect.com/

The training data part of the corpus consists of 350 documents, 50 for development and 100 for testing. This is similar to the pilot task described in Section 5, for which 144 articles were used for training, 40 for development and for 100 testing.

We present statistics about the dataset in Table 1. Notably, the dataset contains many long keyphrases. 22% of all keyphrases in the training set consist of words of 5 or more tokens. This contributes to making the task of keyphrase identification very challenging. However, 93% of those keyphrases are noun phrases[7], which is valuable information for simple heuristics to identify keyphrase candidates. Lastly, 31% of keyphrases contained in the training dataset only appear in it once, systems will have do generalise to unseen keyphrases well.

3.2 Annotation Process

Mention-level annotation is very time-consuming, and only a handful of semantic relations such as hypernymy and synonymy can be found in each publication. We therefore only annotate paragraphs of publications likely to contain relations.

We originally intended to identify suitable documents by automatically extracting a knowledge graph of relations from a large scientific dataset using Hearst-style patterns (Hearst, 1991; Snow et al., 2005), then using those to find potential relations in a distinct set of documents, similar to the distant supervision (Mintz et al., 2009; Snow et al., 2005) heuristic. Documents containing a high number of such potential relations would then be selected. However, this requires automatically learning to identify keyphrases between which those potential relations hold, and requires relations to appear several times in a dataset for such a knowledge graph to be useful.

In the end, this strategy was not feasible due to the difficulty of learning to detect keyphrases automatically and only a small overlap between relations in different documents. Instead, keyphrase-dense paragraphs were detected automatically using a coarse unsupervised approach (Mikolov et al., 2013) and those likely to contain relations were selected manually for annotation.

For annotation, undergraduate student volunteers studying Computer Science, Material Science or Physics were recruited using UCL's student newsletter, which reaches all of its students. Students were shown example annotations and the annotation guidelines, and if they were still interested in participating in the annotation exercise, afterwards asked to select beforehand how many documents they wanted to annotate. Approximately 50% of students were still interested, having seen annotated documents and read annotation guidelines. They were then given two weeks to annotate documents with the BRAT tool (Stenetorp et al., 2012), which was hosted on an Amazon EC2 instance as a web service. Students were compensated for annotations per document. Annotation time was estimated as approximately 12 minutes per document and annotator, on which basis they were paid roughly 10 GBP per hour. They were only compensated upon completion of all annotations, i.e. compensation was conditioned on annotating all documents. The annotation cost was covered by Elsevier. To develop annotation guidelines, a small pilot annotation exercise on 20 documents was performed with one annotator after which annotation guidelines were refined.[8]

We originally intended for student annotators to triple annotate documents and apply majority voting on the annotations, but due to difficulties with recruiting high-quality annotators we instead opted to double-annotate documents, where the second annotator was an expert annotator. Where annotations disagreed, we opted for the expert's annotation. Pairwise inter-annotator agreement between the student annotator and the expert annotator measured with Cohen's kappa is shown in Table 2. The * indicates annotation quality decreased over time, ending with the annotator not completing annotating all documents. To account for this, documents for which no annotations are given are excluded from computing inter-annotator agreement. Out of the annotators completing the annotation exercise, Cohen's kappa ranges between 0.45 and 0.85, with half of them having a substantial agreement of 0.6 or higher. For future iterations of this task, we recommend to invest significant efforts into recruiting high-quality annotators, perhaps with more pre-annotation quality screening.

[7] Parts of speech are determined automatically, using the nltk POS tagger

[8] Annotation guidelines were available to task participants, they can be found here: https://scienceie.github.io/resources.html

Characteristic	
Labels	Material, Process, Task
Topics	Computer Science, Physics, Material Science
Number all keyphrases	5730
Number unique keyphrases	1697
% singleton keyphrases	31%
% single-word mentions	18%
% mentions, word length >= 3	51%
% mentions, word length >= 5	22%
% mentions, noun phrases	93%
Most common keyphrases	'Isogeometric analysis', 'samples', 'calibration process','Zirconium alloys'

Table 1: Characteristics of SemEval 2017 Task 10 dataset, statistics of training sets

Student Annotator	IAA
1	0.85
2	0.66
3	0.63
4	0.60
5	0.50
6	0.48
7	0.47
8	0.45
9*	0.25
10*	0.22
11*	0.20
12*	0.15
13*	0.06

Table 2: Inter-annotator agreement between the student annotator and the expert annotator, measured with Cohen's Kappa

4 Evaluation

SemEval 2017 Task 10 offers three different evaluation scenarios:

1) Only plain text is given (Subtasks A, B, C).

2) Plain text with manually annotated keyphrase boundaries are given (Subtasks B, C).

3) Plain text with manually annotated keyphrases and their types are given (Subtask C).

We refer to the above scenarios as Scenario 1, Scenario 2, and Scenario 3 respectively.

4.1 Metrics

Keyphrase identification (Subtask A) has traditionally been evaluated by calculating the exact matches with the gold standard. There is existing work for capturing semantically similar keyphrases (Zesch and Gurevych, 2009; Kim et al., 2010a), however since these are captured using relations, similar to the pilot task on keyphrase extraction (Section 5) we evaluate keyphrases, keyphrase types and relations with exact match criteria. The output of systems is matched exactly against the gold standard. The traditionally used metrics of precision, recall and F1-score are computed and the micro-average of those metrics across publications of the three genres are calculated. These metrics are also calculated for Subtasks B and C. In addition, for Subtasks B and C, participants are given the option of using text manually annotated with keyphrase mentions and types.

5 Pilot Task

A pilot task on keyphrase extraction from scientific documents was run by other organisers at SemEval 2010 (Kim et al., 2010b). The task was to extract a list of keyphrases representing key topics from scientific documents, i.e. similar to the first part of our proposed Subtask A, only on type-level. Participants were allowed to submit up to 3 runs and were required to submit a list of 15 keyphrases for each document, ranked by the probability of being reader-assigned phrases. Data was collected from the ACM Digital Library for the research areas Distributed Systems, Information Search and Retrieval, Distributed Artificial Intelligence Multiagent Systems and Social and Behavioral Sciences Economics. Participants were provided with 144 training, 40 development and 100 test articles, each set containing a mix of articles of the different research areas. The data was provided in plain text, converted from pdf with pdftotext. Publications were annotated with keyphrases by 50 Computer Science students and added to author-provided keyphrases required by the journals they were published in. Guidelines were for the keyphrases to exactly appear

anywhere in the text of the paper, in reality 15% of annotator-provided keyphrases did not, as well as 19% of author-provided keyphrases. The number of author-specified keywords was 4 on average, whereas annotators identified 12 on average. Returned phrases are considered correct if they are exact matches of either the annotator- or author-assigned keyphrases, allowing for minor syntactic variations (A of B → B A ; A's B → A B). Precision, recall and F1 is calculated for the top 5, top 10 and all keywords. 19 systems were submitted to the task, the best one achieving an F1 of 27.5% on the combined author-assigned and annotator-assigned keywords.

Lessons learned from the task were that performance varies depending on how many keywords are to be extracted, the task organisers recommend against fixing a threshold for a number of keyphrases to extract lead. They further recommend a more semantically-motivated task, taking into account synonyms of keyphrases instead of requiring exact matches. Both of those recommendations will be taken into account for future task design. To fulfill the latter, we will ask annotator to assign types to the identified keywords (process, task, material) and identify semantic relations between them (hypernym, synonym).

6 Existing Resources

As part of the FUSE project with IARPA, we created a small annotated corpus of 100 noun phrases generated from the titles and abstracts derived from the Web Of Science corpora[9] of the domains Physics, Computer Science, Chemistry and Computer Science. These corpora cannot be distributed publicly and were made available by the IARPA funding agency. Annotation was performed by 3 annotators using 14 fine-grained types, including PROCESS.

We measured inter-annotator agreement among the three annotators for the 14 categories using Fleiss' Kappa. The k value was found to be 0.28 which implies that there was fair agreement between them, however distinguishing between the fine-grained types added significantly to the annotation time. Therefore we only use three main types for the SemEval 2017 Task 10.

There are some existing keyphrase extraction corpora, however, they are not similar enough to the proposed task to justify reuse. Below is a description of existing corpora.

The SemEval 2010 Keyphrase Extraction corpus (Kim et al., 2010b)[10] consists of a handful of document-level keyphrases per article. In contrast to the task proposed, the keyphrases are annotated on type-level and not further classified as process, task or material and semantic relations are not annotated. Further, the domains considered are different and mostly sub-domains of Computer Science.

The corpus released by Tateisi et al. (2014)[11] contains sentence-level fine-grained semantic annotations for 230 publication abstracts in Japanese and 400 in English. In contrast to what we propose, the annotations are more fine-grained and annotations are only available for abstracts.

Gupta and Manning (2011) studied keyphrase extraction from ACL Anthology articles, applying a pattern-based bootstrapping approach based on 15 016 documents and assigning the types FOCUS, TECHNIQUE and DOMAIN. Performance was evaluated on 30 manually annotated documents. Although the latter corpus is related to what we propose, manual annotation is only available for a small number of documents and only for the Natural Language Processing domain.

The ACL RD-TEC 2.0 dataset (QasemiZadeh and Schumann, 2016) consists of 300 ACL Anthology abstracts annotated on mention-level with seven different types of keyphrases. Unlike our dataset, it does not contain relation annotations. Note that this corpus was created at the same time as the one SemEval 2017 Task 10 dataset and thus we did not have the chance to build on it. A more in-depth comparison between the two datasets as well as keyphrase identification and classification methods evaluated on them can be found in Augenstein and Søgaard (2017).

6.1 Baselines

We frame the task as a sequence-to-sequence prediction task. We preprocess the files by splitting documents into sentences and tokenising them with nltk, then aligning span annotations from .ann files to tokens. Each sentence is regarded as one sequence. We then split the task into the

[9] http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html

[10] https://github.com/snkim/AutomaticKeyphraseExtraction

[11] https://github.com/mynlp/ranis

three subtasks, keyphrase boundary identification, keyphrase classification and relation classification and add three output layers. We predict the following types, for the three subtasks respectively:
Subtask A: $t_A = O, B, I$ for tokens being outside, at the beginning, or inside a keyphrase
Subtask B: $t_B = O, M, P, T$ for tokens being outside a keyphrase, or being part of a material, process or task
Subtask C: $t_C = O, S, H$ for Synonym-of and Hyponym-of relations. For Subtask A and B, we predict one output label per input token. For Subtask C we predict a vector for each token, that encodes what the relationship between that token and every other token in the sequence is for the first token in each keyphrase. After predictions for tokens are obtained, these are converted back to spans and relations between them in a post-processing step.

We report results for two simple models: one to estimate the *upper bound*, that converts .ann files into instances, as described above, then converts them back into .ann files. Next, to estimate a lower bound, a *random baseline*, that for each token assigns a random label for each of the subtasks.

The *upper bound* span-token-span round-trip conversion performance, an F1 of 0.84, shows that we already lose a significant amount of performance due to sentence splitting and tokenisation alone. The *random* baseline further shows hard especially the keyphrase boundary identification task is and as a result the overall task, since the subtasks depend on one another. For Subtask A, a random baseline achieves an F1 of 0.03. The overall tasks gets easier if keyphrase boundaries are given, resulting in F1 of 0.23 for keyphrase classification, and if keyphrase types are given, an F1 of 0.04 are achieved with the random baseline for Subtask C.

7 Summary of Participating Systems

In this section, we summarise the outcome of the competition. For more details please refer to the respective system description papers and the task website https://scienceie.github.io/.

We had three subtasks, described in Sec 2, which were grouped together in three evaluation scenarios, described in Sec 4. The competition was hosted in CodaLab[12] in two phases: (i) development phase and (ii) testing phase. Fifty four teams participated in the development phase, and out of them twenty six teams participated in the final competition. One of the major success of the competition is due to such wide participation and application of various different techniques starting from neural networks, supervised classification with careful feature engineering to simple rule based methods. We present a summary of approaches used by task participants below.

7.1 Evaluation Scenario 1

In this scenario teams need to solve all three subtasks A, B, and C; where no annotation information was given. Some teams participated only in Subtask A, or B; but the overall micro F1 performance across subtasks is considered for the ranking of the teams. Seventeen teams participated in this scenario. The F1 scores range from 0.04 to 0.43. Complete results are given in Table 3.

Various different types of methods have been applied by different teams with various levels of supervision. The best three teams TTI_COIN, TIAL_UW, and s2_end2end have used recurrent neural network (RNN) based approaches to obtain F1 scores of 0.38, 0.42 and 0.43 respectively. However, TIAL_UW, and s2_end2end, by using a conditional random fields (CRF) layer on top of RNNs achieve a higher F1 in Subtask A compared to TTI_COIN.

The fourth team PKU_ICL with an F1 of 0.37 found classification models based on random forest and support vector machines (SVM) useful with carefully engineered feature such as TF-IDF over a very large external corpus, IDF weighted word-embeddings etc, along with an existing taxonomy. SciX on the other hand used noun phrase chunking and trained an SVM classifier on provided training data to classify phrases, and used a CRF to predict labels of the phrases. CRF based methods with parts-of-speech (POS) tagging and orthographic features such as presence of symbols and capitalisation have been tried by several teams (NTNU, SZTE-NLP, WING-NUS) and they leading to a reasonable performance (F1: 0.23, 0.26, and 0.27, respectively).

Noun phrase extraction with length constraint by HCC-NLP, and using a global list of keyphrases by NITK_IT_PG are found not to perform satisfactorily (F1: 0.16 and 0.14 respectively). The

[12] https://competitions.codalab.org/competitions/15898

Teams	Overall	A	B	C
s2_end2end (Ammar et al., 2017)	**0.43**	0.55	**0.44**	**0.28**
TIAL_UW	0.42	**0.56**	**0.44**	
TTI_COIN (Tsujimura et al., 2017)	0.38	0.5	0.39	0.21
PKU_ICL (Wang and Li, 2017)	0.37	0.51	0.38	0.19
NTNU-1 (Marsi et al., 2017)	0.33	0.47	0.34	0.2
WING-NUS (Prasad and Kan, 2017)	0.27	0.46	0.33	0.04
Know-Center (Kern et al., 2017)	0.27	0.39	0.28	
SZTE-NLP (Berend, 2017)	0.26	0.35	0.28	
NTNU (Lee et al., 2017b)	0.23	0.3	0.24	0.08
LABDA (Segura-Bedmar et al., 2017)	0.23	0.33	0.23	
LIPN (Hernandez et al., 2017)	0.21	0.38	0.21	0.05
SciX	0.2	0.42	0.21	
IHS-RD-BELARUS	0.19	0.41	0.19	
HCC-NLP	0.16	0.24	0.16	
NITK_IT_PG	0.14	0.3	0.15	
Surukam	0.1	0.24	0.1	0.13
GMBUAP (Flores et al., 2017)	0.04	0.08	0.04	
upper bound	0.84	0.85	0.85	0.77
random	0.00	0.03	0.01	0.00

Table 3: F1 scores of teams participating in Scenario 1 and baseline models for Overall, Subtask A, Subtask B, and Subtask C. Ranking of the teams is based on overall performance measured in Micro F1.

former is surprising, as keyphrases are with an overwhelming majority noun phrases, the latter not as much, many keyphrases only appear once in the dataset (see Table 1). GMBUAP further tried using empirical rules obtained by observing the training data for Subtask A, and a Naive Bayes classifier trained on provided training data for Subtask B. Such simple methods on their own prove not to be accurate enough. Attempts of such give us additional insight about the hardness of the problem and applicability of simple methods to the task.

7.2 Evaluation Scenario 2

In this scenario teams needed to solve sub-tasks B, and C. Partial annotation was provided to the teams, that is, solution to the Subtask A. Four teams participated in this scenario with F1 cores ranging from 0.43 to 0.64. Please refer to Table 4 for complete result.

Except MayoNLP, other three teams participated only in Subtask B. Although ranking is done based on overall performance, but in this scenario rankings are consistent in each category. BUAP with the worst F1 score for Subtask B (0.45), is still better than the best team in Scenario 1 s2_end2end for Subtask B (0.44). Partial annotation or accuracy for Subtask A proves to be critical, reinforcing again that identifying keyphrase boundaries is the most difficult part of the shared task.

Unlike the Scenario 1, in this case the top two teams used classifiers with lexical features (F1: 0.64) as well as neural networks (F1: 0.63). The first team MayoNLP used SVM with rich feature sets like n-grams, lexical features, orthographic features, whereas the second team UKP/EELECTION used used three different neural network approaches and subsequently combined them via majority voting. Both these methods perform quite similarly. However, a CRF based approach and an SVM with simpler feature sets attempted by the two teams LABDA and BUAP are found to be less effective in this scenario.

MayoNLP applied a simple rule based method for synonym-of relation extraction, and Hearst patterns for hyponym-of relation detection. The rules for synonym-of detection is based on presence of phrases such as *in terms of*, *equivalently*,

[13] After the end of the evaluation period, team UKP/EELECTION discovered those results were based on training on the development set. For training on the training set, their results are: 0.69 F1 overall and 0.72 F1 for Subtask B only

Teams	Overall	B	C
MayoNLP (Liu et al., 2017)	**0.64**	**0.67**	**0.23**
UKP/EELECTION (Eger et al., 2017)[13]	0.63	0.66	
LABDA (Segura-Bedmar et al., 2017)	0.48	0.51	
BUAP (Alemán et al., 2017)	0.43	0.45	
upper bound	0.84	0.85	0.77
random	0.15	0.23	0.01

Table 4: F1 scores of teams participating in Scenario 2 and baseline models for Overall, Subtask B, and Subtask C. Ranking of the teams is based on overall performance measured in Micro F1. Teams participating in Scenario 2 received partial annotation with respect to Subtask A.

Teams	Overall
MIT (Lee et al., 2017a)	**0.64**
s2_rel (Ammar et al., 2017)	0.54
NTNU-2 (Barik and Marsi, 2017)	0.5
LaBDA (Suárez-Paniagua et al., 2017)	0.38
TTI_COIN_rel (Tsujimura et al., 2017)[15]	0.1
upper bound	0.84
random	0.04

Table 5: F1 scores of teams participating in Scenario 3 and baseline models. Teams participating in Scenario 3 received partial annotation with respect to Subtask A, and Subtask B. Ranking of the teams is based on overall performance measured in Micro F1.

which are called etc in the text between two keyphrases. Interestingly, the RNN based approach of s2_end2end in Scenario 1 performs better than MayoNLP without using partial annotation of Subtask A.

7.3 Evaluation Scenario 3

In this scenario, teams need to solve only Subtask C. Partial annotations were provided to the teams for Subtask B and C. Five teams participated in this scenario, and F1 scores ranged from 0.1 to 0.64. Please refer to Table 5 for complete result.

Neural network (NN) based models are found to perform better than other methods in this scenario. The best method by MIT uses a convolutional NN (CNN). The other method uses two phases of NN and found to be reasonably effective (F1: 0.54).

On the other hand, application of supervised classification with five different classifiers (SVM, decision tree, random forest, multinomial naive Bayes and k-nearest neighbour) using three different feature selection techniques (chi square, decision tree, and recursive feature elimination) found close accuracy (F1: 0.5) with the top performing ones.

LaBDA also use a CNN based method. However, the rule based post-processing and argument ordering strategy applied by MIT seemed to give additional advantage as also observed by them.

However most of the teams in this scenario outperform, all teams from other scenarios (who did not have access to partial information for Subtask B, and C) in relation prediction. This also asserts the significance of accuracy on Subtask A, and B in order to perform accurately on Subtask C.

8 Conclusion

In this paper, we present the setup and discuss participating systems of SemEval 2017 Task 10 on identifying and classifying keyphrases and relations between them from scientific articles, to which 26 systems were submitted. Successful systems vary in their approaches. Most of them use RNNs, often in combination with CRFs as well as CNNs, however the system performing best for evaluation scenario 1 uses an SVM with a well-engineered lexical feature set. Identifying keyphrases is the most challenging subtask, since the dataset contains many long and infrequent keyphrases, and systems relying on remembering them do not perform well.

Acknowledgments

We would like to thank Elsevier for supporting this shared task. Special thanks go to Ronald Daniel Jr. for his feedback on the task setup and Pontus Stenetorp for his advice on brat and shared task organisation.

[15] After the end of the evaluation period, team TTI_COIN_rel discovered a bug in preprocessing, leading to low results. Their overall result after having corrected for that error is a Macro F1 of 0.48.

References

Yuridiana Alemán, Darnes Vilariño, and Josefa Somodevilla. 2017. BUAP at SemEval-2017 Task 10. In *Proceedings of SemEval*.

Waleed Ammar, Matthew Peters, Chandra Bhagavatula, and Russell Power. 2017. The AI2 system at SemEval-2017 Task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction. In *Proceedings of SemEval*.

Isabelle Augenstein and Anders Søgaard. 2017. Multi-Task Learning of Keyphrase Boundary Classification. In *Proceedings of ACL*.

Isabelle Augenstein, Andreas Vlachos, and Diana Maynard. 2015. Extracting Relations between Non-Standard Entities using Distant Supervision and Imitation Learning. In *Proceedings of EMNLP*.

Biswanath Barik and Erwin Marsi. 2017. NTNU-2@ScienceIE: Identifying Synonym and Hyponym Relations among Keyphrases in Scientific Documents. In *Proceedings of SemEval*.

Gábor Berend. 2017. SZTE-NLP at SemEval-2017 Task 10: A High Precision Sequence Model for Keyphrase Extraction Utilizing Sparse Coding for Feature Generation. In *Proceedings of SemEval*.

Panot Chaimongkol, Akiko Aizawa, and Yuka Tateisi. 2014. Corpus for Coreference Resolution on Scientific Papers. In *Proceedings of LREC*.

Steffen Eger, Erik-Lân Do Dinh, Ilia Kuznetsov, Masoud Kiaeeha, and Iryna Gurevych. 2017. EELECTION at SemEval-2017 Task 10: Ensemble of nEural Learners for kEyphrase ClassificaTION. In *Proceedings of SemEval*.

Gerardo Flores, Mireya Tovar, and José A. Reyes-Ortiz. 2017. GMBUAP at SemEval-2017 Task 10: A First Approach Based Empirical Rules and n-grams for the Extraction and Classification of Keyphrases. In *Proceedings of SemEval*.

Sujatha Das Gollapalli and Xiaoli Li. 2015. EMNLP versus ACL: Analyzing NLP research over time. In *Proceedings of EMNLP*.

Kartik Goyal and Chris Dyer. 2016. Posterior regularization for Joint Modelling of Multiple Structured Prediction Tasks with Soft Constraints. In *Proceedings of EMNLP*.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2015. Unsupervised Declarative Knowledge Induction for Constraint-Based Learning of Information Structure in Scientific Documents. *Transactions of the Association for Computational Linguistics* 3:131–143.

Sonal Gupta and Christopher Manning. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of IJCNLP*.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. In *Proceedings of ACL*.

Marti Hearst. 1991. Noun Homograph Disambiguation Using Local Context in Large Text Corpora. *Using Corpora* pages 185–188.

Simon David Hernandez, Davide Buscaldi, and Thierry Charnois. 2017. LIPN at SemEval-2017 Task 10: Selecting Candidate Phrases with Part-of-Speech Tag Sequences from the training data. In *Proceedings of SemEval*.

Dain Kaplan, Ryu Iida, and Takenobu Tokunaga. 2009. Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach. In *Proceedings of NLPIR4DL workshop at ACL-IJCNLP*.

Rohit J Kate and Raymond J Mooney. 2010. Joint Entity and Relation Extraction using Card-Pyramid Parsing. In *Proceedings of CoNLL*.

Roman Kern, Stefan Falk, and Andi Rexha. 2017. Know-Center at SemEval-2017 Task 10: CODE annotator. In *Proceedings of SemEval*.

Su Nam Kim, Timothy Baldwin, and Min-Yen Kan. 2010a. Evaluating n-gram based evaluation metrics for automatic keyphrase extraction. In *Proceedings of Coling*.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010b. SemEval-2010 Task 5 : Automatic Keyphrase Extraction from Scientific Articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, Uppsala, Sweden, pages 21–26.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2017a. MIT at SemEval-2017 Task 10: Relation Extraction with Convolutional Neural Networks. In *Proceedings of SemEval*.

Lung-Hao Lee, Kuei-Ching Lee, and Yuen-Hsien Tseng. 2017b. The NTNU System at SemEval-2017 Task 10: Extracting Keyphrases and Relations from Scientific Publications Using Multiple Conditional Random Fields. In *Proceedings of SemEval*.

Sijia Liu, Shen Feichen, Vipin Chaudhary, and Hongfang Liu. 2017. MayoNLP at SemEval 2017 Task 10: Word Embedding Distance Pattern for Keyphrase Classification in Scientific Publications. In *Proceedings of SemEval*.

Erwin Marsi and Pinar Öztürk. 2015. Extraction and generalisation of variables from scientific publications. In *Proceedings of EMNLP*.

Erwin Marsi, Utpal Kumar Sikdar, Cristina Marco, Biswanath Barik, and Rune Sætre. 2017. NTNU-1@ScienceIE at SemEval-2017 Task 10: Identifying and Labelling Keyphrases with Conditional Random Fields. In *Proceedings of SemEval*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS*.

M. Mintz, S. Bills, R. Snow, and D. Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*.

Diarmuid Ó Séaghdha and Simone Teufel. 2014. Unsupervised learning of rhetorical structure with untopic models. In *Proceedings of Coling*.

Fuchun Peng and Andrew McCallum. 2006. Information extraction from research papers using conditional random fields. *Information Processing and Management* 42(4):963–979.

Animesh Prasad and Min-Yen Kan. 2017. WING-NUS at SemEval-2017 Task 10: Keyphrase Extraction and Classification as Joint Sequence Labeling. In *Proceedings of SemEval*.

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 LREC*.

Dragomir Radev and Amjad Abu-Jbara. 2012. Rediscovering ACL Discoveries Through the Lens of ACL Anthology Network Citing Sentences. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*.

Isabel Segura-Bedmar, Cristóbal Colón-Ruiz, and Paloma Martínez. 2017. LABDA at SemEval-2017 Task 10: Extracting Keyphrases from Scientific Publications by combining the BANNER tool and the UMLS Semantic Network. In *Proceedings of SemEval*.

Yanchuan Sim, Bryan Routledge, and Noah A. Smith. 2015. A Utility Model of Authors in the Scientific Community. In *Proceedings of EMNLP*.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint Inference of Entities, Relations, and Coreference. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC)*.

R. Snow, D. Jurafsky, and A.Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems* 17:1297–1304.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of EACL System Demonstrations*.

Lucas Sterckx, Cornelia Caragea, Thomas Demeester, and Chris Develder. 2016. Supervised Keyphrase Extraction as Positive Unlabeled Learning. In *Proceedings of EMNLP*.

Víctor Suárez-Paniagua, Isabel Segura-Bedmar, and Paloma Martínez. 2017. LaBDA Team at SemEval-2017 Task C: Relation Classification between keyphrases via Convolutional Neural Network. In *Proceedings of SemEval*.

Yuka Tateisi, Yo Shidahara, Yusuke Miyao, and Akiko Aizawa. 2014. Annotation of Computer Science Papers for Semantic Relation Extraction. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncin Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of LREC*.

Simone Teufel. 2006. Argumentative Zoning for Improved Citation Indexing. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text*, Springer, volume 20 of *The Information Retrieval Series*, pages 159–169.

Tomoki Tsujimura, Makoto Miwa, and Yutaka Sasaki. 2017. TTI-COIN at SemEval-2017 Task 10: Investigating Embeddings for End-to-End Relation Extraction from Scientific Papers. In *Proceedings of SemEval*.

Liang Wang and Sujian Li. 2017. PKU_ICL at SemEval-2017 Task 10: Keyphrase Extraction with Model Ensemble and External Knowledge. In *Proceedings of SemEval*.

Torsten Zesch and Iryna Gurevych. 2009. Approximate Matching for Evaluating Keyphrase Extraction. In *Proceedings of RANLP*.