TDMSci: A Specialized Corpus for Scientific Literature Entity Tagging of Tasks Datasets and Metrics

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin and Debasis Ganguly
IBM Research Europe, Ireland
{yhou|mgleize|fbonin|debasga1}@ie.ibm.com

Abstract

Tasks, *Datasets* and *Evaluation Metrics* are important concepts for understanding experimental scientific papers. However, most previous work on information extraction for scientific literature mainly focuses on the abstracts only, and does not treat datasets as a separate type of entity (Zadeh and Schumann, 2016; Luan et al., 2018). In this paper, we present a new corpus that contains domain expert annotations for *Task (T), Dataset (D), Metric (M)* entities on 2,000 sentences extracted from NLP papers. We report experiment results on TDM extraction using a simple data augmentation strategy and apply our tagger to around 30,000 NLP papers from the ACL Anthology. The corpus is made publicly available to the community for fostering research on scientific publication summarization (Erera et al., 2019) and knowledge discovery.

1 Introduction

The recent years have witnessed a significant growth in the number of scientific publications and benchmarks in many disciplines. As an example, in the year 2019 alone, more than 170k papers were submitted to the pre-print repository arXiv[1] and among them, close to 10k papers were classified as NLP papers (i.e., cs.CL). Each empirical field of science, including NLP, will benefit from the massive increase in studies, benchmarks, and evaluations, as they can provide ingredients for novel scientific advancements.

However, researchers may struggle to keep track of all studies published in a particular field, resulting in duplication of research, comparisons with old or outdated benchmarks, and lack of progress. In order to tackle this problem, recently there have been a few manual efforts to summarize the state-of-the-art on selected subfields of NLP in the form of leaderboards that extract tasks, datasets, metrics and results from papers, such as *NLP-progress*[2] or *paperswithcode*.[3] But these manual efforts are not sustainable over time for all NLP tasks.

Over the past few years, several studies and shared tasks have begun to tackle the task of entity extraction from scientific papers. Augenstein et al. (2017) formalized a task to identify three types of entities (i.e., *task, process, material*) in scientific publications (SemEval 2017 task10). Gábor et al. (2018) presented a task (SemEval 2018 task 7) on semantic relation extraction from NLP papers. They provided a dataset of 350 abstracts and reuse the entity annotations from Zadeh and Schumann (2016). Recently Luan et al. (2018) released a corpus containing 500 abstracts with six types of entity annotations. However, these corpora do not treat *Dataset* as a separate type of entity and most of them focus on the abstracts only.

In a previous study, we developed an IE system to extract {*task, dataset, metric*} triples from NLP papers based on a small, manually created task/dataset/metric (TDM) taxonomy (Hou et al., 2019). In practice, we found that a TDM knowledge base is required to extract TDM information and build NLP leaderboards for a wide range of NLP papers. This can help researchers quickly understand related literature for a particular task, or to perform comparable experiments.

As a first step to build such a TDM knowledge base for the NLP domain, in this paper we present a specialized English corpus containing 2,000 sentences taken from the full text of NLP papers which have been annotated by domain experts for three main concepts: *Task* (T), *Dataset* (D) and *Metric*

[1] https://arxiv.org/help/stats/2019_by_area

[2] https://github.com/sebastianruder/NLP-progress

[3] https://paperswithcode.com

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 707–714
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

(M). Based on this corpus, we develop a TDM tagger using a novel data augmentation technique. In addition, we apply this tagger to around 30,000 NLP papers from the ACL Anthology and demonstrate its value to construct an NLP TDM knowledge graph. We release our corpus at `https://github.com/IBM/science-result-extractor`.

2 Related Work

A lot of interest has been focused on information extraction from scientific literature. SemEval 2017-task 10 (Augenstein et al., 2017) proposed a new task for the identification of three types of entities (*Task, Method*, and *Material*) in a corpus of 500 paragraphs taken from open access journals. Based on Augenstein et al. (2017) and Gábor et al. (2018), Luan et al. (2018) created *SciERC*, a dataset containing 500 scientific abstracts with annotations for six types of entities and relations between them. Both SemEval 2017-task 10 and *SciERC* do not treat "*dataset*" as a separate entity type. Instead, their "*material*" category comprises a much larger set of resource types, including tools, knowledge resources, bilingual dictionaries, as well as datasets. In our work, we focus on "*datasets*" entities that researchers use to evaluate their approaches because dataset is one of the three core elements to construct leaderboards for NLP papers.

Concurrent to our work, Jain et al. (2020) develop a new corpus *SciREX* which contains 438 papers on different domains from *paperswithcode*. It includes annotations for four types of entities (i.e., *Task, Dataset, Metric, Method*) and the relations between them. The initial annotations were carried out automatically using distant signals from *paperswithcode*. Later human annotators performed necessary corrections to generate the final dataset. *SciREX* is the closest to our corpus in terms of entity annotations. In our work, we focus on TDM entities which reflect the collectively shared views in the NLP community and our corpus is annotated by five experts who all have 5-10 years NLP research experiences.

3 Corpus Creation

3.1 Annotation Scheme

We developed an annotation scheme for annotating Task, Dataset, and Evaluation Metric phrases in NLP papers. Our annotation guidelines[4] are based on the scientific term annotation scheme described in Zadeh and Schumann (2016). Different from previous corpora (Zadeh and Schumann, 2016; Luan et al., 2018), we only annotated **factual** and **content-bearing** entities. This is because we aim to build a TDM knowledge base in the future and non-factual entities (e.g., *a high-coverage sense-annotated corpus* in Example 1) do not reflect the collectively shared views of TDM entities in the NLP domain.

(1) In order to learn models for disambiguating a large set of content words, *a high-coverage sense-annotated corpus* is required.

Following the above guidelines, we also do not annotate *anonymous entities*, such as "*this task*" or "*the dataset*". These entities are anaphors and can not be used independently to refer to any specific TDM entities without contexts. In general, we choose to annotate TDM entities that normally have specific names and whose meanings usually are consistent across different papers. From this perspective, the TDM entities that we annotate are similar to named entities, which are self-sufficient to identify the referents.

3.2 Pilot Annotation Study

Data preparation. For the pilot annotation study, we choose 100 sentences from the NLP-TDMS corpus (Hou et al., 2019). The corpus contains 332 NLP papers which are annotated with triples of $\{$*Task, Dataset, Metric*$\}$ on the document level. We use string and substring match to extract a list of sentences from these papers which are likely to contain the document level *Task, Dataset, Metric* annotations. We then manually choose 100 sentences from this list following the criteria: 1) the sentence should contain the valid mention of *Task*, *Dataset*, or *Metric*; 2) the sentences should come from different papers as much as possible; and 3) there should be a balanced distribution of *task*, *dataset*, and *metric* mentions in these sentences.

Annotation agreement. Four NLP domain experts annotated the same 100 sentences for a pilot annotation study, following the annotation guidelines described above. All the annotations were conducted using BRAT (Stenetorp et al., 2012). The inter annotator agreement has been calculated with a pairwise comparison between annotators using *precision*, *recall* and *F-score* on the exact match of the annotated entities. In other words,

[4] Please see the appendix for the whole annotation scheme.

	Mean F-score (EM)	Fleiss' κ (Token)
Task	0.720	0.797
Dataset	0.752	0.829
Metric	0.757	0.896
Overall	0.743	0.842

Table 1: Inter-annotator agreement.

	Train	Test
# Sentences	1500	500
# Task	1219	396
# Dataset	420	192
# Metric	536	174

Table 2: Statistics of task/dataset/metric mentions in the training and testing datasets.

two entities are considered matching (true positive) if they have the same boundaries and are assigned to the same label. We also calculate Fleiss' kappa on a per token basis, comparing the agreement of annotators on each token in the corpus. Table 1 lists the mean F-score as well as the token-based Fleiss' κ value for each entity type. Overall, we achieve high reliability for all categories.

Adjudication. The final step of the pilot annotation was to reconcile disagreements among the four annotators to produce the final canonical annotation. This step also allows us to refine the annotation guidelines. Specifically, through the discussion of annotation disagreements we could identify ambiguities and omissions in the guidelines. For example, one point of ambiguity was whether a *task* must be associated with a dataset, or can we annotate higher level tasks, e.g., *sequence labeling*, which do not have a dedicated dataset but may include several tasks and datasets. This discussion also revealed the overlap in how we refer to tasks and datasets in the literature. As authors we frequently use these interchangeably, often with shared tasks, e.g., "*SemEval-07 task 17*" seems to more often refer to a dataset than a specific instance of the (Multilingual) Word Sense Disambiguation task, or the "*MultiNLI*" corpus is sometimes used as shorthand for the task. After the discussion, we agreed that we should annotate higher level tasks. In addition, we should assign labels to entities according to their actual referential meanings in contexts.

3.3 Main Annotation

After the pilot study, 1,900 additional sentences were annotated by five NLP researchers. Four annotators participated in the pilot annotation study, and all annotators joined the adjudication discussion. Note that every annotator annotate a different set of sentences. The annotator who designed the annotation scheme annotated 700 sentences, the other four annotators annotated 300 sentences each.[5]

In general, most sentences in our corpus are not from the abstracts. Note that the goal of developing our corpus is to automatically build an NLP TDM taxonomy and use them to tag NLP papers. Therefore, the inclusion of sentences from the whole paper other than the abstract section is important for our purpose. Because not all abstracts talk about all three elements. For instances, for the top ten papers listed in the {*sentiment analysis, IMDB, accuracy*} leaderboard in *paperswithcode*[6], only four abstracts mention the dataset "*IMDB*". If we only focus on the abstracts, we will miss the other six papers from the leaderboard.

4 A TDM Entity Tagger

Our final corpus *TDMSci* contains 2,000 sentences with 2,937 mentions of three entity types. We convert the original BRAT annotations to the standard CoNLL format using BIO scheme.[7] We develop a tagger to extract TDM entities based on this corpus.

4.1 Experimental Setup

To evaluate the performance of our tagger, we split *TDMSci* into training and testing sets, which contains 1,500 and 500 sentences, respectively. Table 2 shows the statistics of task/dataset/metric mentions in these two datasets. For evaluation, we report precision, recall, F-score on exact match for each entity type as well as micro-averaged precision, recall, F-score for all entities.

[5] Due to time constraints, we did not carry out another round of pilot study. Partially it is because we felt that the revised guidelines resulting from the discussion were sufficient for the annotators to decide ambiguous cases. So in the second stage annotators annotated disjoint sets of sentences. After this, the annotator who designed the annotation scheme went through the whole corpus again to verify the annotations.

[6] https://paperswithcode.com/sota/sentiment-analysis-on-imdb, search was carried out on November, 2020.

[7] Note that our BRAT annotation contains a small amount of embedded entities, e.g., *WSJ portion of Ontonotes* and *Ontonotes*. We only keep the longest span when we convert the BRAT annotations to the CoNLL format.

	CRF			*CRF w/ gazetteer*			*SciIE*			*Flair-TDM*		
	P	R	F	P	R	F	P	R	F	P	R	F
Original training data												
Task	63.79	46.72	53.94	61.86	45.45	52.40	69.23	54.55	61.02	61.54	54.55	57.83
Dataset	65.42	36.46	46.82	65.45	37.50	47.68	66.97	38.02	48.50	52.66	46.35	49.30
Metric	80.00	66.67	72.73	80.95	68.39	74.14	77.99	71.26	74.47	76.33	74.14	75.22
Micro-	68.45	48.69	56.90	67.70	48.69	56.64	71.21	54.20	61.55	62.99	56.96	59.79
Original Training data + Augmented masked training data												
Task	63.24	43.43	51.50	62.96	42.93	51.05	68.63	55.81	**61.56**	65.14	53.79	58.92
Dataset	62.38	32.81	43.00	64.71	34.38	44.90	55.43	50.52	52.86	59.15	50.52	**54.50**
Metric	80.15	62.64	70.32	79.29	63.79	70.70	76.83	72.41	74.56	79.63	74.14	**76.79**
Micro-	67.58	45.14	54.13	67.77	45.54	54.47	67.17	58.27	**62.40**	67.23	57.61	62.05

Table 3: Results of different models for *task/dataset/metric* entity recognition on *TDMSci* test dataset.

4.2 Models

We model the task as a sequence tagging problem. We apply a traditional CRF model (Lafferty et al., 2001) with various lexical features and a BiLSTM-CRF model for this task. To compare with the state-of-the-art entity extraction model on scientific literature, we also use *SciIE* from Luan et al. (2018) to train a *TDM* entity recognition model based on our training data. Below we describe all models in detail.

CRF. We use the Stanford CRF implementation (Finkel et al., 2005) to train a *TDM* NER tagger based on our training data. We use the following features: unigrams of the previous, current and next words, current word character n-grams, current POS tag, surrounding POS tag sequence, current word shape, surrounding word shape sequence.

CRF with gazetteers. To test whether the above CRF model can benefit from knowledge resources, we add two gazetteers to the feature set: one is a list containing around 6,000 dataset names which were crawled from LRE Map,[8] and another gazetteer comprises around 30 common evaluation metrics compiled by the authors.

SciIE. Luan et al. (2018) proposed a multi-task learning system to extract entities and relations from scientific articles. *SciIE* is based on span representations using ELMo (Peters et al., 2018) and here we adapt it for *TDM* entity extraction. Note that if *SciIE* predicts several embedded entities, we keep the one that has the highest confidence score. In practice we notice that this does not happen in our corpus.

Flair-TDM For BiLSTM-CRF model, we use the recent *Flair* framework (Akbik et al., 2018) based on the cased BERT-base embeddings (Devlin et al., 2018). We train our *Flair-TDM* model with a learning rate of 0.1, a batch size of 32, a hidden size of 768, and the maximum epochs of 150.

4.3 Data Augmentation

For TDM entity extraction, we expect that the surrounding context will play an important role. For instance, in the following sentence "we show that for X on the Y, our model outperforms the prior state-of-the-art", one can easily guess that X is a task entity while Y is a dataset entity. As a result, we propose a simple data augmentation strategy that generates the additional mask training data by replacing every token within an annotated TDM entity as `UNK`.

4.4 Results and Discussion

Table 3 shows the performance of different models for *task/dataset/metric* entity recognition on our testing dataset.

First, it seems that although adding gazetteers can help the CRF model detect *dataset* and *metric* entities better, the positive effect is limited. In general, both *SciIE* and *Flair-TDM* perform better than *CRF* models for detecting all three type of entities.

Second, augmenting the original training data with the additional masked data as described in Section 4.3 further improves the performance both for *SciIE* and *Flair-TDM*. However, this is not the case for the CRF models. We assume this is because CRF models heavily depend on the lexical features.

Finally, we randomly sampled 100 sentences from the testing dataset and compared the predicted TDM entities in *Flair-TDM* against the gold annotations. We found that most errors are from the boundary mismatch for task and dataset entities, e.g., *text summarization* vs. *abstractive text sum-*

[8] `http://www.elra.info/en/catalogues/lre-map/`

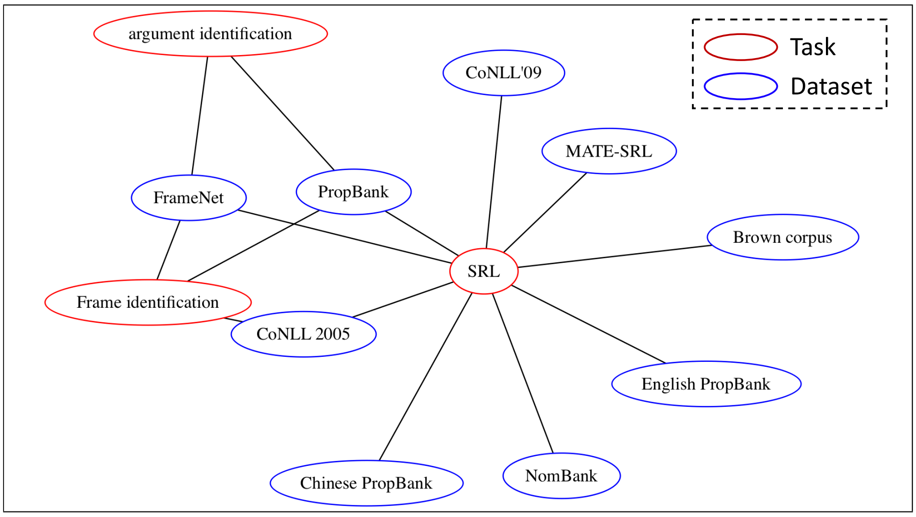

Figure 1: A subset of the *TDM* graph.

marization, or *Penn Treebank* vs. *Penn Treebank dataset*. The last error comes from the bias in the training data. A lot of researchers use "*Penn Treebank*" to refer to a dataset. So the model will learn this bias and only tag "*Penn Treebank*" as the dataset even though in a specific testing sentence, "*Penn Treebank dataset*" was used to refer to the same corpus.

In general, we think these mismatched predictions are reasonable in the sense that they capture the main semantics of the referents. Note that the numbers reported in Table 3 are based on exact match. Often, requiring exact match may be too restictive for downstreaming tasks. Therefore, we carried out an additional evaluation for the best *Flair-TDM* model using partial match from SemEval 2013-Task 9 (Segura-Bedmar et al., 2013), which gives us a micro-average F1 of 76.47 for type partial match.

5 An Initial TDM Knowledge Graph

In this section, we apply the *Flair-TDM* tagger to around 30,000 NLP papers from ACL Anthology to build an initial TDM knowledge graph.

We downloaded all NLP papers from the ACL Anthology[9] covering the period of 1974-2019. For each paper, we collect sentences from the title, the abstract/introduction/dataset/corpus/experiment sections, as well as from the table captions. We then apply the *Flair-TDM* tagger to these sentences. Based on the tagger results, we build an initial graph G using the following steps:

- add a *TDM* entity as a node into G if it appears at least five times in more than one paper;
- create a link between a *task* node and a *dataset/metric* node if they appear in the same sentence at least five times in different papers.

By applying the above simple process, we get a noisy *TDM* knowledge graph containing 180k nodes and 270k links. After checking a few dense areas, we find that our graph encodes valid knowledge about NLP task/dataset/metric. Figure 1 shows that in our graph, the task "SRL" (semantic role labelling) is connected to a few datasets such as "FrameNet", "PropBank", and "NomBank" that are standard benchmark datasets for this task.

Based on the tagged ACL Anthology and this initial noisy graph, we are exploring various methods to build a large-scale NLP TDM knowledge graph and to evaluate its accuracy/coverage in an ongoing work.

6 Conclusion

In this paper, we have presented a new corpus (*TDMSci*) annotated for three important concepts (*Task/Dataset/Metric*) that are necessary for extracting the essential information from an NLP paper. Based on this corpus, we have developed a *TDM* tagger using a simple but effective data augmentation strategy. Experiments on 30,000 NLP papers show that our corpus together with the *TDM* tagger can help to build *TDM* knowledge resources for the NLP domain.

[9] https://www.aclweb.org/anthology/

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

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. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, Guy Lev, Achiya Jerbi, Jonathan Herzig, Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, Francesca Bonin, and David Konopnicki. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 211–216, Hong Kong, China. Association for Computational Linguistics.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT, New Orleans, Louisiana, June 5-6, 2018*, pages 679–688.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5203–5213.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

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 the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Behrang Q. Zadeh and Anne-Kathrin Schumann. 2016. The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In *LREC*.

A TDM Entity Annotation Guidelines

A.1 Introduction

This scheme describes guidelines for annotating *Task*, *Dataset*, and *Evaluation Metric* phrases in NLP papers. We have pre-processed NLP papers in PDF format and chosen sentences that are likely to contain the above-mentioned entities for annotation. These sentences may come from different sections (e.g., Abstract, Introduction, Experiment, Dataset) as well as tables (e.g., table captions).

A.2 Entity Types

We annotate the following three entity types:

- Task: A task is a problem to solve (e.g., *information extraction*, *sentiment classification*, *dialog state tracking*, *POS tagging*, *NER*).

- Dataset: A dataset is a specific corpus or language resource. Datasets are often used to develop models or run experiments for NLP tasks. A dataset normally has a short name, e.g., *IMDB*, *Gigaword*.

- Metric: An evaluation metric explains the performance of a model for a specific task, e.g., *BLEU* (for machine translation), or *accuracy* (for a range of NLP tasks).

A.3 Notes and Examples

Entity spans. Particular attention must be paid to the entity spans in order to improve agreement. The following list indicates all the annotation directions that annotators have been given regarding entity spans. Table 4 shows examples of correct span annotation.

- Following the ACL RD-TEC 2.0 annotation guideline,[10] determiners should not be part of an entity span. For example, the string 'the text8 test set', only the span 'test8' is annotated as *dataset*.

- Minimum span principle: Annotators should annotate only the minimum span necessary to represent the original meaning of task/dataset/metric. See Table 4, rows 1,2,3,4.

- Include 'corpus/dataset/benchmark' when annotating dataset if these tokens are the head-noun of the dataset entity. For example: 'ubuntu corpus', 'SemEval-2010 Task 8 dataset'.

- Exclude the head noun of 'task/problem' when annotating task (e.g., only annotation "link prediction" for "the link prediction problem") unless they are the essential part of the task itself (e.g., CoNLL-2012 shared task, SemEval-2010 relation classification task).

- Conjunction: If the conjunction NP is an ellipse, annotate the whole phrase (see Table 4, rows 6,11); otherwise, annotate the conjuncts separately (see Table 4, row 5).

- Tasks can be premodifiers (see Table 4, rows 7,8,12)

- Embedded spans: Normally TDM entities do not contain any other TDM entities. A small number of *Task* and *Dataset* entities can contain other entities (see Table 4, row 12).

Anonymous entities. Do not annotate anonymous entities, which include anaphors. The following examples are anonymous entities:

- *this task*

- *this metric*

- *the dataset*

- *a public corpus for context-sensitive response selection* in the sentence "Experimental results in a a public corpus for context-sensitive response selection demonstrate the effectiveness of the proposed multi-vew model."

Abbreviation. If both the full name and the abbreviation are present in the sentence, annotate the abbreviation with its corresponding full name together. For instance, we annotate "20-newsgroup (20NG)" as a dataset entity in Example 2.

Factual entity. Only annotate "factual, content-bearing" entities. Task, dataset, and metric entities normally have specific names and their meanings are consistent across different papers. In Example 3, "*a high-coverage sense-annotated corpus*" is not a factual entity.

[10] https://github.com/languagerecipes/acl-rd-tec-2.0/blob/master/distribution/documents/acl-rd-tec-guidelines-ver2.pdf

Row	Phrase	Annotation	Entity
1	The public Ubuntu Corpus	Ubuntu Corpus	Dataset
2	the web portion of TriviaQA	web portion of TriviaQA	Dataset
3	sentiment classification of movie reviews	sentiment classification	Task
4	the problem of part-of-speech tagging for informal, online conversational text	part-of-speech tagging	Task
5	The FB15K and WN18 datasets	FB15K; WN18	Dataset
6	Hits at 1, 3 and 10	Hits at 1, 3 and 10	Metric
7	Link prediction benchmarks	Link prediction	Task
8	POS tagging accuracy	POS tagging; accuracy	Task, Metric
9	the third Dialogue State Tracking Challenge	Dialogue State Tracking, third Dialogue State Tracking Challenge	Task, Dataset
10	SemEval-2017 Task 9	SemEval-2017 Task 9	Task
11	temporal and causal relation extraction and classification	temporal and causal relation extraction and classification	Task
12	the SemEval-2010 Task 8 dataset	SemEval-2010 Task 8 dataset; SemEval-2010 Task 8	Dataset,Task

Table 4: Examples of entity span annotation guidelines

(2) We used four datasets: IMDB, Elec, RCV1, and 20-newsgrous (20NG) to facilitate direct comparison with DL15.

(3) In order to learn models for disambiguating a large set of content words, a high-coverage sense-annotated corpus is required.