Dataset Mention Extraction and Classification

Animesh Prasad[†] **Chenglei Si**[‡] **Min-Yen Kan**[†]
[†] School of Computing, National University of Singapore
[‡] River Valley High School, Singapore
[†]{animesh,kanmy}@comp.nus.edu.sg

Abstract

Datasets are integral artifacts of empirical scientific research. However, due to natural language variation, their recognition can be difficult and even when identified, can often be inconsistently referred across and within publications. We report our approach to the *Coleridge Initiative's Rich Context Competition*, which tasks participants with identifying dataset surface forms (dataset mention extraction) and associating the extracted mention to its referred dataset (dataset classification). In this work, we propose various neural baselines and evaluate these model on one-plus and zero-shot classification scenarios. We further explore various joint learning approaches – exploring the synergy between the tasks – and report the issues with such techniques.

1 Introduction

The modern scientific method hinges on replicability and falsifiability. Datasets are an essential aspect of enabling such analysis in much of modern empirical studies. Datasets themselves are varied — in size, complexity, substructure, and scope — and references to them are also varied — in naming convention and subsequent reference or citation, both within and across documents.

Dataset mention extraction and classification has thus become more critical not only to facilitate the identification of proper target datasets for testing hypotheses but also to benchmark incremental research by extension. In this work, we explore various neural approaches to identifying cited surface forms associated with a dataset and interlinking them. We benchmark our approach on the Coleridge Initiative's Rich Text Context Competition (RTCC), released in 2018, which we participated in, whose dataset comprises of social science publications exemplify such confusability problems with multiple surface dataset citations.

2 Related Work

The extraction of important scientific terms within full-text documents has been desiderata of scholarly document analyses extending back decades. In the early 90s, work by Liddy (Liddy, 1991) explored the possibility of promoting key scholarly document metadata into structured abstracts. Generic aspects of scholarly documents have been explored in (Gupta and Manning, 2011) where key aspects of publications namely *focus*, *domain* and *techniques* were identified using linguistic patterns. Domain-specific corpora with complex taxonomies such as the ACL RD-TEC (QasemiZadeh and Schumann, 2016) have also been employed to train systems to identify fine-grained aspects. In the field of nursing and primary care, the key metadata of *Patients*, *Intervention*, *Condition*, and *Outcome* characterize the acronym "PICO", which has also been the target of much work (Zhao et al., 2010; Wallace et al., 2016).

Recently, shared tasks concerning key generic metadata (inclusive of datasets) have been run in the community. The ScienceIE shared task (Augenstein et al., 2017) benchmarked techniques for identifying predefined entities matching *Process*, *Task* and *Materials*; where the definition of *Material* entities overlap with that of datasets. The task asked to extract such entities and identify the relations among them on short excerpts of scientific documents. State-of-the-art deep learning and feature-based sequential labeling models set the standard for approaches on such tasks, using Long Short-Term Memory (LSTM) (Ammar et al., 2017) and Conditional Random Field (CRF) (Prasad and Kan, 2017) models, respectively.

Though related to a general named entity recognition, we see the problem of dataset mention extraction as having particular challenges. In contrast to the related scientific document process-

Publication: ....Source: Monitoring the Future: National Survey on Drug Use, 1975-2009....Section 2 provides a brief summary of trends in adolescent drinking and smoking, using data for the US from the annual Monitoring the Future survey.....Trends in Adolescent Drinking and Smoking: Monitoring the Future.....Systematic annual data on the prevalence of underage drinking and smoking in the US are collected and tracked by several organizations. This section relies on data from the Monitoring the Future (MTF)....
Datasets (Present): [... **56:** Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1984; **101:** Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1989;...]
Datasets (Not Present): [... **100:** Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 1996; **108:** Current Population Survey, May 1973; ...] –

Figure 1: A text fragment from the training data. Highlighted text represent dataset mentions (citations). Note that a particular mention may refer to multiple datasets. In some examples, as highlighted here, there are many different datasets which closely resemble each other in their surface form.

ing tasks of keyphrase extraction (with 10–15 keywords within a document; i.e., (Kim et al., 2010)) or identify such semantic entities within a small excerpt (identifying which 5-10 tokens constitute entities over 30–40 tokens; i.e., (Augenstein et al., 2017)) dataset mentions within full-text documents exhibits a much higher ratio of sparsity. Further, coreference resolution techniques specific to linking the dataset mention to the dataset have yet to be well explored.

3 Background

We first formally define the task following the specification from the RTCC, as consisting of two sub-problems:

•**Dataset Mention Extraction:** Given a publication (d_i), identify fragments of the text that are mentions of a dataset.

•**Dataset Classification:** Classify the detection mention text fragment to a particular dataset in the knowledge base (D_i).

Corpus. The corpus is compiled by Coleridge Initiative Rich Context Competition[1] (see the example in Fig. 1) and consists of 5K publication sampled from various social studies, averaging 7K tokens in length. About half of the documents (2.5K) are annotated, featuring an average of 2.2 datasets and 7.5 different dataset mentions per document. Note that some documents do not mention datasets at all. Additionally, the RTCC makes a list of known datasets available (sized 10K), which is taken as an input knowledge base for resolution. Many of the 10K datasets do not appear in the corpus. Hence for these datasets, there is no mention–dataset pair. The corpus allows us to explore the dataset classification problem at three levels of complexity, from easiest to most challenging:

•**One-plus classification:** at least one dataset–mention pair is present in *training* for all the *testing* datasets.

•**Zero-shot classification:** no dataset–mention pairs are known in *training* data for the *testing* dataset, but the dataset is known to the provided knowledge base. The model knows the dataset description and has to do the classification subtask, but not discovery.

•**Zero-shot discovery:** the scenario where even the dataset (and by extension, dataset–mentions pairs) is unknown to the system (not present in the provided knowledge base). This is also the ultimate aim of a discovery system, which simultaneously needs to populate datasets and their mentions from an empty knowledge base. We do not address this scenario directly in this current work but discuss joint models that can potentially cater to this problem.

4 Model

As the RTCC corpus has only been recently released, there are no formally published approaches, nor public results. However, we have identified that the top performing systems in the competition treat the subtasks of mention extraction and dataset classification as two separate tasks. We explore various neural approaches for both the individual tasks and the look more closely the case of joint modeling. Correct extraction dataset mention is the direct prerequisite task of dataset classification. This motivates us to investigate joint model to perform both tasks. We examine two different realizations of such a joint model that supports multi-task learning.

Baselines. We model mention extraction as a sequence labeling task. This admits a range of neural models as sequence labeling baselines

[1] https://coleridgeinitiative.org/richcontextcompetition

for this task. We start with a Bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) ('BiLSTM') model that employs pre-trained word embeddings. We then incrementally increase the model's power of representation in other baselines. First, we incorporate a Convolutional Neural Network (CNN) over character embeddings ('CNN-BiLSTM'). Second, we add a CRF layer over the BiLSTM outputs ('CNN-BiLSTM-CRF'); and finally incorporate Bahdanau attention (Bahdanau et al., 2014) over the LSTM layer ('CNN-BiLSTM-Attn-CRF').

Our selection of these incremental components is motivated by the aspects of the problem. Applying a CNN over the character embeddings is introduced to tackle domain-specific terminology that may conserve internal character sequences, such as acronyms found in dataset names. Such names are generally out-of-vocabulary (OOV) with respect to generic word embeddings. The application of the CRF is motivated to reduce token-level noise by incorporating global (i.e., within a sentence input) decoding. The attention mechanism is similarly motivated to focus the model on the specific parts of the input sequence, as datasets and their mentions occur within specific contexts and are not uniformly distributed. The attention mechanism used is defined as follows: first, suppose the sequence output of the BiLSTM $H \in \mathbf{R}^{N \times T \times h}$, where N is the batch size, T is the sequence length and h is the hidden dimension of BiLSTM. Then the model performs the following operations:

$$\begin{aligned} A &= H^T \\ A &= Softmax(A) \\ S &= A^T \odot H \end{aligned} \tag{1}$$

where $W \in \mathbf{R}^{T \times T}$ is the weight matrix to be trained and $\odot$ represents the Hadamard product. For the third dataset discovery task, we use sentence classification models *i.e.* BiLSTM and CNN (Kim, 2014) as baselines, replacing the standard sigmoid final binary classification with a softmax layer to enable multilabel multiclass classification.

Shared Layer Extraction–Classification ('SL E–C'). The first joint system selects the best system for each of the individual subtasks, then unifies them by providing a common feature extraction base and optimization using joint losses over both subtasks. We start with the best overall baseline for the mention extraction subtask (*cf.* 5): CNN-BiLSTM-Attn-CRF. It uses the single CNN-BiLSTM-Attn to encode the textual content, followed by a CRF. For the dataset classification subtask, we share the output from the CNN-BiLSTM-Attn base, and substitute the CRF layer with a CNN layer for dataset classification, as from our empirical tuning, we found the CNN model provides the best performance for dataset classification.

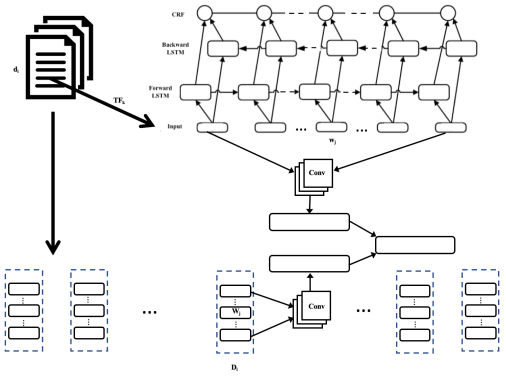

Figure 2: KBSL E–C model. Word embeddings for tokens in each text fragment TF_i (upper left) are translated to its hidden representation via BiLSTM-CNN-Attn-CRF trained with binary labels for mention tokens (upper right). Separately, we apply CNN on the text fragment and all j datasets to obtain datasets representation (individually at a time; bottom row). These are merged and passed to a dense layer, which we train with binary labels to establish which dataset is referenced.

KB Shared Layer Extraction–Classification ('KBSL E–C'). In this model (*cf.* Fig. 2), we leverage on the meta-information of the dataset knowledge base to better support zero-shot learning. There is a description (we experiment two configurations – *name* and *description*) of each dataset in the given knowledge base as part of the corpus. First, we use convolution followed by global max pooling to obtain a representation of each dataset's description text. We then apply convolutions to known mentions of the dataset D_i. Both representations are then merged and passed to a dense layer with a binary output such that $f(TF_k, D_i) = 1$ if TF_k mentions D_i, else 0. This step is repeated for all datasets ($i \in [0, m]$) during testing, and a few randomly, sampled datasets per text fragment during training.

Unlike SL E–C, KBSL E–C can incorporate new classes dynamically by creating a new class representation for predicted new class. Thus KBSL E–C represents an end-to-end zero-shot

dataset discovery model.

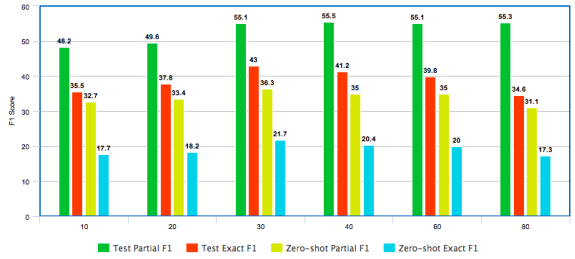

Figure 3: Token ngram-based CRF performance with differing segment lengths.

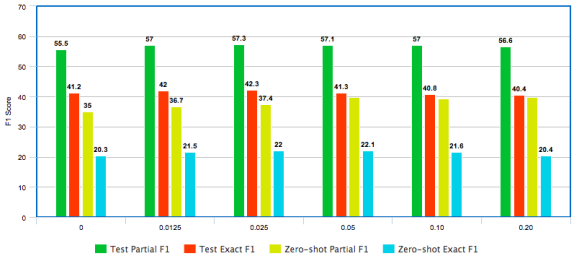

Figure 4: Token ngram-based CRF performance with different NSR, segment length 40.

4.1 Experiment

We elaborate on the complete experimental setup, which has the following configuration:

Hyper-parameters. As the documents in the corpus have 7K tokens on average, the sequence lengths are too long for any model to process directly. We split the documents into shorter text fragment (TF_i) for training and inference. Most fragments do not contain any dataset mentions; these segments we term "negative segments".

The document collection is thus highly skewed, with only 0.4% positive tokens (similarly for positive segments). We under-sample to lessen the effect of data skew, by only considering some of the negative segments during training. We sample all "positive segments", those with dataset mentions.

Our processing methods involve two hyper-parameters – the segment length and the sampling rate of negative segments. Both hyper-parameters affect the ratio of negative tokens sampled in the training set, which in turn impacts performance. We experiment with the CRF baseline model (trigram model, whose hand-tuned features include uppercasing and digits) to analyze the effect of these hyper-parameters and select optimal values (*cf.* Fig. 3 and Fig. 4). For example, a negative sampling rate (NSR) of 0.05 means that we sample 5% of the total number of negative segments from the original dataset for training; conversely, NSR=0 means every training segment contains at least one dataset mention. Note that even for NSR=0, there are still many negative tokens as each segment only contains a few short mention phrases (4.7 tokens per mention on average), with the rest negative.

From the table, we can see that the model generally works better when the negative token rate is small. We use the optimal segment length 40 and NSR=0.015 (1.5%) for all neural models in this paper.

Model Configuration. For all models, we use the 300-dimensional GloVe (Pennington et al., 2014) word embeddings. All models are trained with Adam optimizer.

For dataset mention extraction, the task-specific parameters are as follows. For the base BiLSTM, we use a hidden size of 100 and a dropout rate of 0.2 on word embeddings. We then used a dense layer with sigmoid activation to determine the probability of the input being part of a dataset mention. For the character embedding CNN, we use character embedding dimension 300, 1D convolution 300 filters, window size 6, and a dropout rate of 0.4. For the CNN-BiLSTM-CRF model, we add a CRF layer on top of the BiLSTM instead of a dense layer.

For dataset classification, the task-specific parameters are as follows. For the CNN model, we use 1D convolution with 256 kernels, with window size 6, followed by global max pooling, and a dense layer for the final classification output. For the LSTM based model, we use a BiLSTM with hidden dimension 100 to encode the input sequence and use a dense layer on the final state of the BiLSTM for the final dataset classification. We use a sigmoid for the final non-linear activation function. As explained earlier, the rationale to use sigmoid is to allow the model to associate a single mention to multiple datasets which appear commonly in the dataset (see the example in Fig. 1).

Evaluation Method. We evaluate our model on the **development set**, the **test set** and on the **zero-shot test set**. We first randomly held out 7% of the datasets from the corpus and select the publications (219 documents in total) containing these datasets to form the zero-shot test set. To be clear, the datasets in the zero-shot test set are not seen at all within the training set. We then ran-

	Development Set						Test Set						Zero-Shot Test Set					
	Partial			Exact			Partial			Exact			Partial			Exact		
Model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BiLSTM	71.4	64.4	67.7	31.3	34.0	32.6	29.4	32.1	30.7	11.2	12.8	12.0	25.3	20.0	22.4	6.3	6.3	6.3
CNN-BiLSTM	77.5	75.5	**76.5**	41.4	44.6	43.0	49.8	44.7	47.1	28.6	31.2	29.8	38.7	28.6	32.9	18.0	20.8	19.3
CNN-BiLSTM-CRF	79.1	71.1	74.9	42.7	44.6	**43.6**	54.1	44.6	48.9	35.6	33.8	34.7	41.6	27.9	33.4	23.2	22.7	**23.0**
CNN-BiLSTM-Attn-CRF	76.1	73.8	74.9	39.4	47.7	43.2	58.0	50.0	**53.7**	34.8	38.0	**36.4**	42.6	28.9	**34.4**	17.2	17.3	17.3
SL E–C	77.2	72.6	74.8	39.9	41.6	40.7	40.3	43.1	41.7	27.1	28.4	27.7	29.0	28.0	28.5	16.3	16.7	16.5

Table 1: Mention Extraction Subtask performance. Segment length 40, negative sampling rate: 0.015.

	Development Set			Test Set			Zero-Shot Test Set		
Model	P	R	F_1	P	R	F_1	P	R	F_1
BiLSTM	73.1	71.6	72.3	27.5	47.4	34.8	3.0	5.7	3.9
CNN	81.3	79.5	**80.4**	42.8	46.5	**44.6**	4.9	5.0	**5.0**
SL E–C	70.6	70.0	70.3	31.8	49.3	38.6	3.6	6.3	4.6
KBSL E–C	96.0	85.9	90.7	17.6	27.5	21.5	0.8	1.1	0.9
KBSL E–C descript	97.5	88.1	92.6	12.3	44.6	19.4	0.5	1.9	0.9

Table 2: Dataset Classification Subtask performance. Segment length 40, negative sampling rate 0.015.

domly hold out 225 publications to form the test set. The datasets mentioned in these testing documents may have other mentions in the training set as well. The dev set is split from the training set (5%) and has the same distribution and length as the training set.

Since the test set and zero-shot test set contain complete documents and do not have any sampling, the distribution is different from the sampled training set. During the evaluation, we do not sample. We first split the test documents into text segments of the same length as the training segments and perform inference with our trained model on these segments. We combine the predicted results as the prediction for the entire test document.

We employ **precision** (**P**), **recall** (**R**) and $\boldsymbol{F_1}$ **score** as our evaluation metrics. For dataset mention subtask, these metrics can be interpreted in a relaxed or strict manner, with respect to true token coverage. The relaxed, **partial** match metric attributes a true positive count if any of the ground truth tokens are correctly predicted by the model as a mention phrase. The strict, **exact** match metric attributes a true positive only when if every token in the mention is predicted correctly. We also report exact match P, R, F_1 at the document level.

5 Results

CNN-LSTM-Attn-CRF and CNN outperform all the other models in the single task setup for mention extraction and dataset discovery, respectively. We note that the performance of sequence labeling models is not very high even though when the task seems trivial. We attribute this to the high number of text fragments with no dataset mention, resulting in low accuracy. Similar to CRF (*cf.* Fig. 3 and Fig. 4), the precision-recall trade-off for smaller-to-bigger fragments does not allow for optimization by mere tuning of fragment size.

We further find that surprisingly the SL E–C model doesn't increase the performance of either of the tasks. The sequence labeling task is more sensitive to local information. Ideally, the output of mention extraction should be input to classification and hence prime signal for the classification task. But, we find that the classification benefit from more contextual information than just the mention (in fact we find using extracted mentions works even worse) and hence sharing layers causes mix-up of representations of the text input which isn't ideal for either task.

KBSL E–C model retain the trend of the decrease in performance on individual tasks. But surprisingly the model doesn't perform well on the zero-shot test set. On further analysis, we realize this is caused by the nature of dataset with multiple similar datasets making it easier for even simple classification model to achieve a partial score for classification even when the model has not seen an example of the dataset.

6 Conclusion

We explore the problem of identifying the mention of datasets in publications and associate the identified mention to a dataset. In our experiments we find CNN-BiLSTM-CRF and CNN models work best for dataset mention extraction and classification respectively. We identify that while mention extraction is primarily dependent on local signals the dataset classification uses a much wider context than just the mention.

References

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 the 11th International Workshop on Semantic Evaluation, SemEval 2017, Vancouver, Canada, August 3-4, 2017*, pages 592–596.

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, Vancouver, Canada, August 3-4, 2017*, pages 546–555.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Sonal Gupta and Christopher Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*, pages 1–9.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

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 the 5th International Workshop on Semantic Evaluation, SemEval 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 21–26.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*, pages 1746–1751.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*.

Animesh Prasad and Min-Yen Kan. 2017. WING-NUS at SemEval-2017 task 10: Keyphrase identification and classification as joint sequence labeling. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval 2017, Vancouver, Canada, August 3-4, 2017*, pages 973–977.

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 Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596.

Jin Zhao, Min-Yen Kan, Paula M Procter, Siti Zubaidah, Wai Kin Yip, and Goh Mien Li. 2010. Improving search for evidence-based practice using information extraction. In *AMIA Annual Symposium Proceedings*, volume 2010, page 937. American Medical Informatics Association.