Leveraging Citation Graph for Scientific Information Extraction

Mahendra Nandi
dept. of Computer Science
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur, Howrah, WB, India
mahendranandi.rkma@gmail.com

Under the Supervision of
Prof. Deborshi Kumar Sanyal
School of Mathematical & Computational Sciences
IACS, Jadavpur, WB, India

***Abstract*—The rapid expansion in published scientific knowledge has enormous potential for good, if it can only be harnessed correctly. Despite the value of huge quantity of focused research, it is infeasible for the scientific community to read many papers in a time-critical situation, and make accurate judgements to help separate signal from the noise. To this end, how can machines help researchers quickly identify relevant papers? One step in this direction is to automatically extract and organize scientific information (e.g. important concepts and their relations) from a collection of research articles, which could help researchers identify new methods or materials for a given task.**

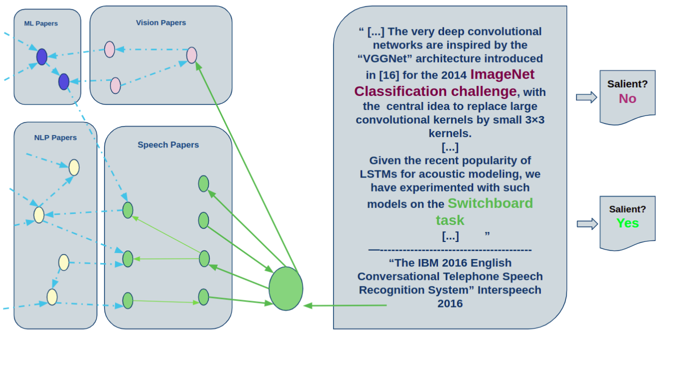
I. Introduction

Automatically extracting key information from scientific documents has the potential to help scientists work more efficiently and accelerate the pace of scientific progress. Most existing works on scientific information extraction (SciIE) consider extraction solely based on the content of an individual paper, without considering the paper's place in the broader literature. Scientific information extraction (SciIE) which aims to extract structured information from scientific articles, has seen growing interest recently, as reflected in the rapid evolution of systems and datasets. However, scientific papers do not exist in a vacuum they are part of a larger ecosystem of papers, related to each other through different conceptual relations. We can claim a better under-standing of a research article relies not only on its content but also on its relations with associated works, using both the content of related papers and the paper's position in the larger citation network. For example, in Fig. 1., if we consider the given article, we can see that we can't say whether the entities are salient or not, but if we see the place of the article in the broader place of literature then we can surely say that the 1st one can't be a salient one. (Where salient means whether the entity is key to the work described in a paper.)

II. Document-level Scientific IE

A. Task Definition

We consider the task of extracting document-level relations from scientific texts. Most work on scientific information extraction has used annotated datasets of scientific abstracts, such as those provided for SemEval 2017 and SemEval 2018 shared tasks (Augenstein et al., 2017; Gábor et al., 2018), the SciERC dataset (Luan et al., 2018), and the BioCreative V Chemical Disease Relation dataset (Wei et al., 2016). We here focus on the task of open-domain document level relation extraction from long, full-text documents. This is in contrast to the above methods that only use paper abstracts. Our setting is also different from works that consider a fixed set of candidate relations (Hou et al., 2019; Kardas et al., 2020) or those that only consider IE tasks other than relation extraction, such as entity recognition. We follow the paper CitaionE by Vijay Viswanathan and Graham Neubig and Pengfei Liu.



" [...] The very deep convolutional networks are inspired by the "VGGNet" architecture introduced in [16] for the 2014 ImageNet Classification challenge, with the central idea to replace large convolutional kernels by small 3×3 kernels.
[...]
Given the recent popularity of LSTMs for acoustic modeling, we have experimented with such models on the Switchboard task
[...] "

"The IBM 2016 English Conversational Telephone Speech Recognition System" Interspeech 2016

Fig. 1. Example of using the citation graph to improve the task of salient entity classification

Each document consists of sections D = $S_1, ..., S_N$, where each section contains a sequence of words $Si = w_{i,1}, ..., w_{i,N_i}$. Each document comes with annotations of entities, coreference clusters, cluster-level saliency labels, and 4-ary document-level relations. We break down the end-to-end information extraction process as a sequence of these four related tasks, with each task taking the output of the preceding tasks as input.

a) Mention Identification: For each span of text within a section, this task aims to recognize if the span describes a Task, Dataset, Method, or Metric entity, if any.

b) Coreference: This task requires clustering all entity mentions in a document such that, in each cluster, every mention refers to the same entity. The SciREX dataset which we have used for this project includes coreference annotations for each Task, Dataset, Method, and Metric mention

c) Salient Entity Classification: Given a cluster of mentions corresponding to the same entity, the model must predict whether the entity is key to the work described in a paper. We follow the definition from the SciREX dataset (Jain et al., 2020), where an entity in a paper is deemed salient if it plays a role in the paper's evaluation.

d) Relation Extraction: The ultimate task in our IE pipeline is relation extraction. We consider relations as 4-ary tuples of typed entities ($E_{Task}, E_{Dataset}, E_{Method}, E_{Metric}$), which are required to be salient entities. Given a set of candidate relations, we must determine which relations are contained in the main result of the paper.

III. Dataset

Although citation network information has been shown to be effective in other tasks, few works have recently tried using it in SciIE systems. One potential reason is the lack of a suitable dataset. We combine the rich annotations of SciREX with a source of citation graph information, S2ORC (Lo et al., 2020). For each paper, S2ORC includes parsed metadata about which other papers cite this paper, which other papers are cited by this paper, and locations in the body text where reference markers are embedded. our work only used the SciREX dataset, our methods can be readily extended to other SciIE datasets. We now describe our citation-aware scientific IE architecture, which incorporates citation information into mention identification, salient entity classification, and relation extraction. For each task, we consider two types of citation graph information, either separately or together: (1) structural information from the graph network topology and (2) textual information from the content of citing and cited documents.

IV. Baseline Model

We base our work on top of the model of Jain et al. (2020), which was introduced as a strong baseline accompanying the SciREX dataset which is also followed by CitationIE by Vijay Viswanathan. This multi-task model performs three of our tasks (mention identifica tion, saliency classification, and relation extraction) in a sequence, treating coreference resolution as an external black box. While word and span representations are shared across all tasks and updated to minimize multi-task loss, the model trains each task on gold input. Figure 2 summarizes the baseline model's end-to-end architecture, and highlights the places where we use citation information for improvements for the model.

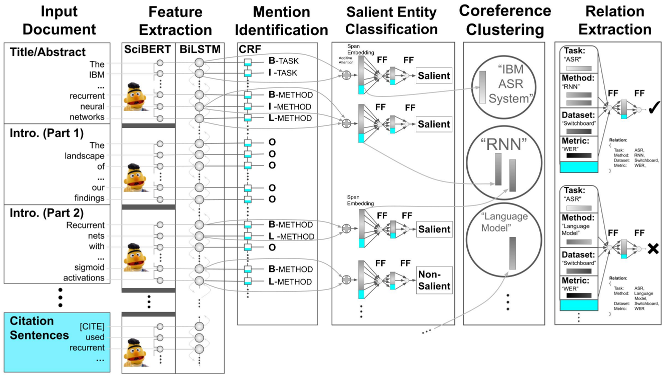


Fig. 2. Architecture of the model we use for neural information extraction. Light blue blocks indicate places where we can incorporate information from the citation graph for the citation-aware CitationIE architecture.

A. Feature Extraction

The model extracts features from raw text in two stages. First, contextualized word embeddings are obtained for each section by running SciBERT (Beltagy et al., 2019) on that section of text (up to 512 tokens). Then, the embeddings from all words over all sections are passed through a bidirectional LSTM (Graves et al., 2005) to contextualize each word's representation with those from other sections.

B. Mention Identification

The baseline model treats this named entity recognition task as an IOBES sequence tagging problem (Reimers and Gurevych, 2017). The tagger takes the SciBERT- BiLSTM (Beltagy et al., 2019; Graves et al., 2005) word embeddings (as shown in the Figure 2), feeds them through two feedforward networks (not shown in Figure 2), and produces tag potentials at each word. These are then passed to a CRF (Lafferty et al., 2001) which predicts discrete tags.

C. Span Embeddings

For a given mention span, its span embedding is produced via additive attention (Bahdanau et al., 2014) over the tokens in the span.

D. Coreference

Using an external model, pairwise coreference predictions are made for all entity mentions, forming coreference clusters.

E. Salient Entity Classification

Saliency is a property of entity clusters, but it is first predicted at the entity mention level. Each entity mention's span embedding is simply passed through two feedforward networks, giving a binary saliency prediction. To turn these mention-level predictions into cluster-level predictions, the predicted saliency scores are max-pooled over all mentions in a coreference cluster to give cluster-level saliency scores.

F. Relation Extraction

The model treats relation extraction as binary classification, taking as input a set of 4 typed salient entity clusters. For each entity cluster in the relation, per-section entity cluster representations are computed by taking the set of that entity's mentions in a given section, and max-pooling over the span embeddings of these mentions. The four entity-section embeddings (one for each entity in the relation) are then concatenated and passed through a feedforward network to produce a relation-section embedding. Then, the relation-section embeddings are averaged over all sections and passed through another feedforward network which returns a binary prediction.

V. Experiments

The ultimate product of our work is an end-to-end document-level relation extraction system, but we also measure each component of our system in isolation, giving end-to-end and per-task metrics. We evaluate mention identification with the average F1 score of classifying entities of each span type. Salient Entity Classification Similar to Jain et al. (2020) we evaluate this task at the mention level and cluster level. We evaluate both metrics on gold standard entity recognition inputs. Relation Extraction This is the ultimate task in our pipeline. We use its output and metrics to evaluate the end-to-end system, but also evaluate relation extraction separately from upstream components to isolate its performance.

For each task, we first train that component in isolation from the rest of the system to minimize. the task-specific loss. We then take the best performing modifications and use them to train end to-end IE models to minimize the sum of losses from all tasks. We train each model on a single GPU with batch size 4 for up to 20 epochs. We trained the model for 20 epoch with a batch size of 4 with other hyper-parameter as usual.

VI. Quantitative Results

- For mention identification, we observe no major performance difference from using citation graphs, and include full results in beow table.
- Using citation graph embeddings significantly improves the system with respect to the salient mention metric.
- (2) Graph embeddings do not improve cluster evaluation significantly
- Incorporating graph embeddings and citances simultaneously is no better than using either.
- Relation Extraction Table shows that using graph embeddings here gives an 11.5 point improvement in document-level F1 over the reported baseline, 11 and statistically significant gains on both corpus-level F1 metrics. Despite seemingly large gains on the document level F1 metric, these are not statistically significant due to significant inter-model variability and small test set size, despite the graph embedding model performing best at every seed we tried.

  Below results are gained after using citation information.

TABLE I
Mention Identification Results:using BiLSTM

F1 measures of ENTITIES	**How Citation information incorporated**	
	Embedding	Citances
Material	0.57	0.54
Metric	0.73	0.73
Task	0.70	0.68
Method	0.78	0.77
overall	0.74	0.73

TABLE II
Mention Identification Results:using BiGRU

F1 measures of ENTITIES	**How Citation information incorporated**	
	Embedding	Citances
Material	0.52	0.54
Metric	0.72	0.73
Task	0.70	0.69
Method	0.78	0.78
overall	0.74	0.74

TABLE III
Salient Mention Evaluation :: using BiLSTM

	How Citation information incorporated	
metric	***Embedding***	Citances
f1	0.59	0.59
p	0.54	0.51
r	0.66	0.68

TABLE IV
Salient Mention Evaluation :: using BiGRU

	How Citation information incorporated	
Metric	***Embedding***	Citances
f1	0.58	0.57
p	0.53	0.50
r	0.65	0.67

TABLE V
4-ary Relation Metrics using BiLSTM:

Information type	Metric	filtering on gold clusters	on gold clusters
Embedding	**f1**	0.64	0.79
	p	0.69	0.80
	r	0.65	0.87
Citances	**f1**	0.58	0.82
	p	0.60	0.80
	r	0.59	0.94

TABLE VI
4-ary Relation Metrics using BiGRU:

Information type	Metric	filtering on gold clusters	on gold clusters
Embedding	**f1**	0.61	0.76
	p	0.64	0.77
	r	0.62	0.84
Citances	**f1**	0.45	0.63
	p	0.48	0.63
	r	0.47	0.72

ACKNOWLEDGMENT

I am very happy to get the chnace to thank the Authors of refernece papers- Sarthak Jain, Vijay Viswanathan and others for there valuable work that helped me a lot. Next I want to thank my University (Ramakrishna Mission Vivekananda Educational and Research Institute), the Computer Science Department, HOD Swami Dhyanagamyananda Maharaj, my project guide Prof. Deborshi Kumar Sanyal for their support and for providing an ideal environment and Research Laboratory.

REFERENCES

[1] 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.

[2] 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, Brussels, Belgium. Association for Computational Linguistics.

[3] Vijay Viswanathan and Graham Neubig and Pengfei Liu, Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)," CitationIE: Leveraging the Citation Graph for Scientific Information Extraction