Layout Detection for Scientific Paper PDFs

with Layout Parser & Label Studio

Shannon Shen

Applied Research Scientist
Allen Institute for AI

Ben Lee

PhD Candidate
University of Washington

Michael Malyuk

CEO
Heartex

Outline

Introduction (5min)

Demo (40min)

Q & A (15min)

Outline

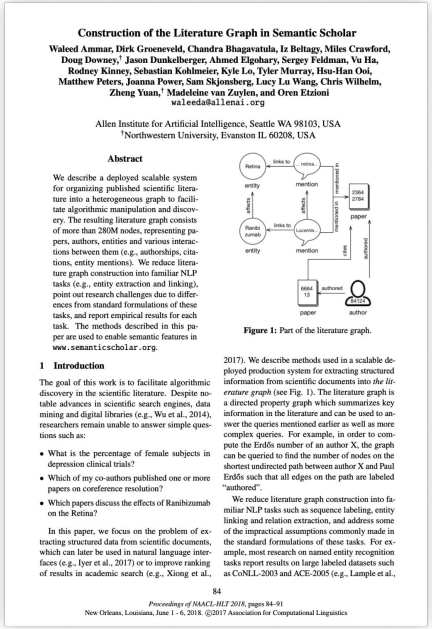
Introduction (5min)

Demo (40min)

Q & A (15min)

The Problem

Extract structured data from complex documents



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018.

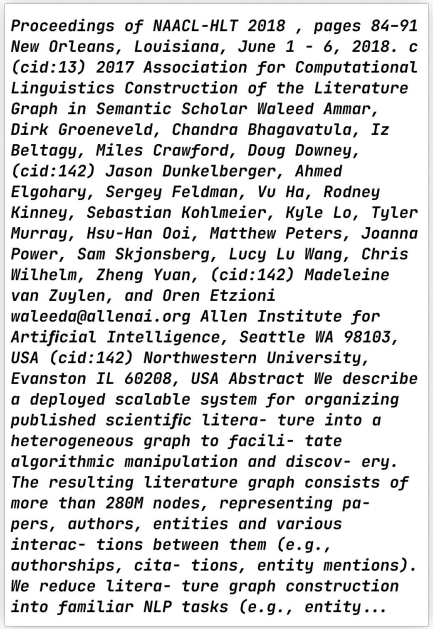
Input / **PDF or Scans**

```
{
"title": "Construction of
the Literature Graph in
Semantic Scholar",
"authors": "Waleed Ammar et.
al.",
"abstract": "We describe a
deployed scalable system for
organizing published
scientific literature into a
heterogeneous graph to
facilitate algorithmic
manipulation and ...",
"sections": ["..."]
}
```

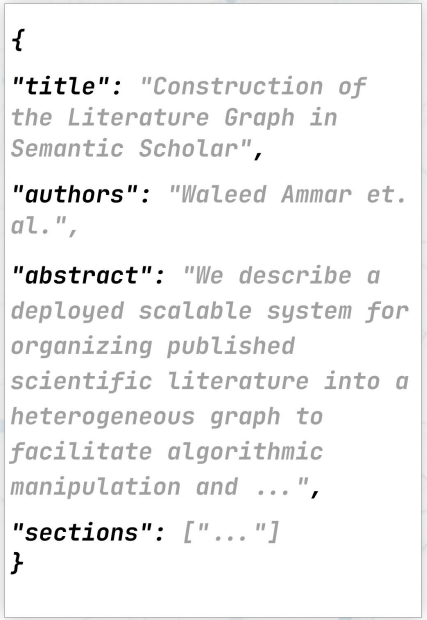
Output / **Metadata JSON**

Challenge

Difficult if using only string information



Proceedings of NAACL-HLT 2018 , pages 84–91 New Orleans, Louisiana, June 1 - 6, 2018. c (cid:13) 2017 Association for Computational Linguistics Construction of the Literature Graph in Semantic Scholar Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, (cid:142) Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, (cid:142) Madeleine van Zuylen, and Oren Etzioni waleeda@allenai.org Allen Institute for Artificial Intelligence, Seattle WA 98103, USA (cid:142) Northwestern University, Evanston IL 60208, USA Abstract We describe a deployed scalable system for organizing published scientific litera- ture into a heterogeneous graph to facili- tate algorithmic manipulation and discov- ery. The resulting literature graph consists of more than 280M nodes, representing pa- pers, authors, entities and various interac- tions between them (e.g., authorships, cita- tions, entity mentions). We reduce litera- ture graph construction into familiar NLP tasks (e.g., entity...



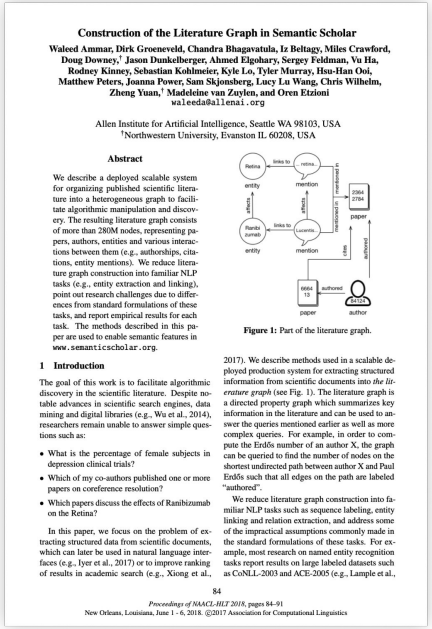
```
{
"title": "Construction of the Literature Graph in Semantic Scholar",
"authors": "Waleed Ammar et. al.",
"abstract": "We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and ...",
"sections": ["..."]
}
```

Input / **Text Strings**

Output / **Metadata JSON**

Solution

Utilize document layouts



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

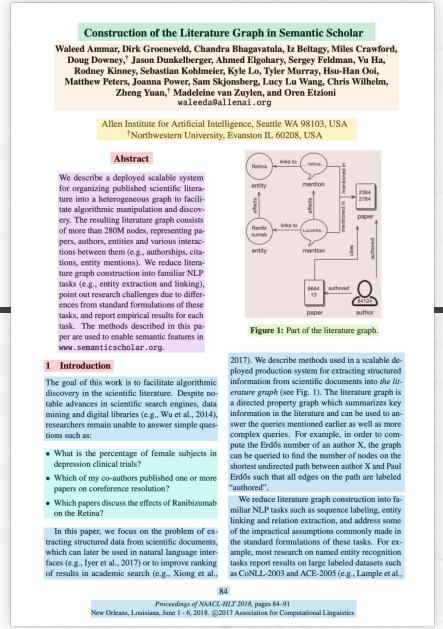
- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018.

Input / **PDF or Scans**



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018.

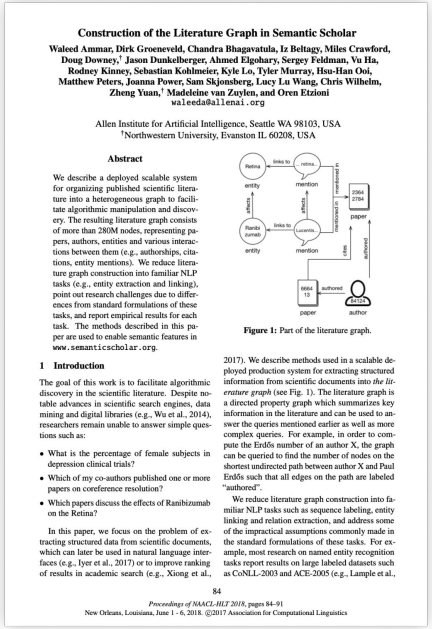
Document Layout


```
{
"title": "Construction of
the Literature Graph in
Semantic Scholar",
"authors": "Waleed Ammar et.
al.",
"abstract": "We describe a
deployed scalable system for
organizing published
scientific literature into a
heterogeneous graph to
facilitate algorithmic
manipulation and ...",
"sections": ["..."]
}
```


Output / **Metadata JSON**

Layout Parser

Parse document layouts with Deep Learning



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

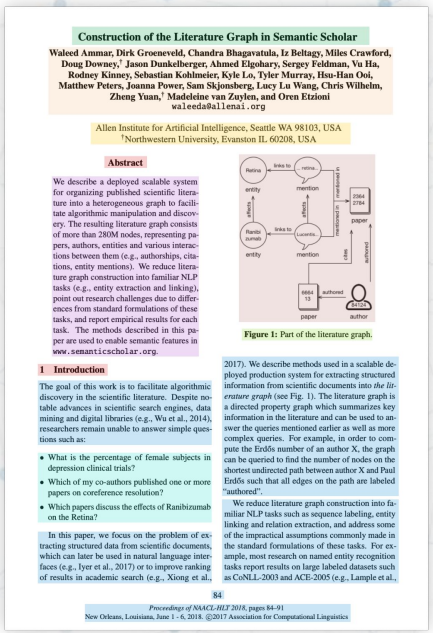
We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018.

Input / **PDF or Scans**

Layout Parser

Layout Detection Models



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

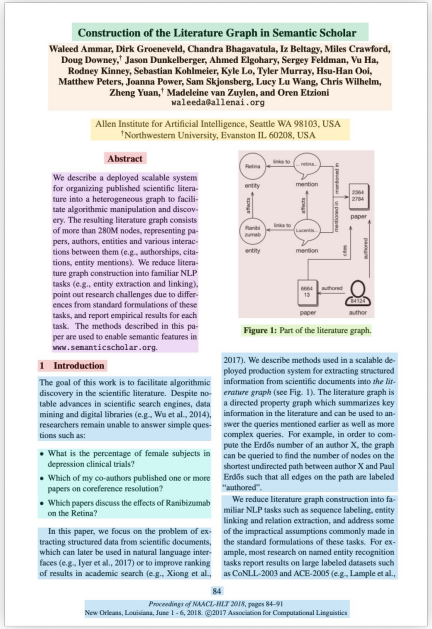
84

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018. ©2017 Association for Computational Linguistics

Intermediate / **Document Layout**

Layout Parser

Support data extraction based on layouts



Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey,† Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan,† Madeleine van Zuylen, and Oren Etzioni
waleeda@allenai.org

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA
†Northwestern University, Evanston IL 60208, USA

Abstract

We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and discovery. The resulting literature graph consists of more than 280M nodes, representing papers, authors, entities and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction into familiar NLP tasks (e.g., entity extraction and linking), point out research challenges due to differences from standard formulations of these tasks, and report empirical results for each task. The methods described in this paper are used to enable semantic features in www.semanticscholar.org.

Figure 1: Part of the literature graph.

1 Introduction

The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of extracting structured data from scientific documents, which can later be used in natural language interfaces (e.g., Iyer et al., 2017) or to improve ranking of results in academic search (e.g., Xiong et al., 2017). We describe methods used in a scalable deployed production system for extracting structured information from scientific documents into *the literature graph* (see Fig. 1). The literature graph is a directed property graph which summarizes key information in the literature and can be used to answer the queries mentioned earlier as well as more complex queries. For example, in order to compute the Erdős number of an author X, the graph can be queried to find the number of nodes on the shortest undirected path between author X and Paul Erdős such that all edges on the path are labeled "authored".

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the impractical assumptions commonly made in the standard formulations of these tasks. For example, most research on named entity recognition tasks report results on large labeled datasets such as CoNLL-2003 and ACE-2005 (e.g., Lample et al.,

Proceedings of NAACL-HLT 2018, pages 84–91
New Orleans, Louisiana, June 1 - 6, 2018.

Intermediate / **Document Layout**

Layout Parser

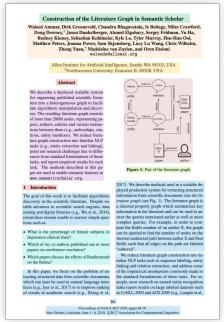
Layout Data APIs

```
{
"title": "Construction of
the Literature Graph in
Semantic Scholar",
"authors": "Waleed Ammar et.
al.",
"abstract": "We describe a
deployed scalable system for
organizing published
scientific literature into a
heterogeneous graph to
facilitate algorithmic
manipulation and ...",
"sections": ["..."]
}
```

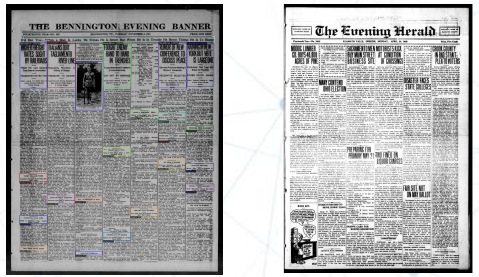
Output / **Metadata JSON**

Generalizability?

How to adapt to new different data quickly?



Layout Model for Document Type A

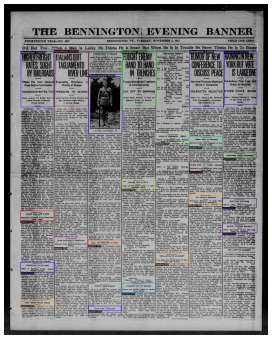


New Data for Document Type B

Image Credit: from the Library of Congress, Chronicling America: Historic American Newspapers site

Layout Parser + Label Studio Annotation

Customizing models → better accuracy on new data



New Doc Data

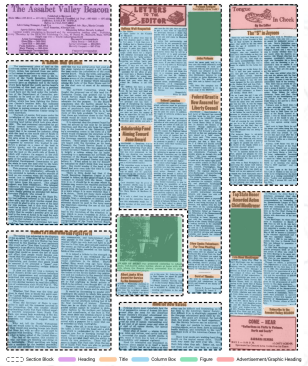
Label Studio

Data Annotation

Train/Eval Dataset

Layout Parser

Model Training



Apply new models

Credit: from the Library of Congress, Chronicling America: Historic American Newspapers site

Outline

Introduction (5min)

Demo (40min)

Q & A (15min)

Outline

Introduction (5min)

Demo (40min)

Q & A (15min)

Layout Parser

Parsing Complex Documents with DL

layout-parser/layout-parser

@layoutparser

layout-parser.slack.com

arXiv: 2103.15348

Label Studio

Data Annotation for Training Better Models

heartexlabs/label-studio

@heartexlabs

slack.labelstud.io.s3-website-us-east-1.amazonaws.com