

# 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,<sup>1</sup> Dirk Grunwald,<sup>1</sup> Chandra Bhagavatula,<sup>1</sup> Isabella Miles Crawford,<sup>1</sup> Doug Downey,<sup>1</sup> Jason Dombrowski,<sup>1</sup> Ahmed Elgohary,<sup>1</sup> Sergey Feldman,<sup>1</sup> Vu Ha,<sup>1</sup> Rodney Kinney,<sup>1</sup> Sebastian Kohlmeier,<sup>1</sup> Kyle Lo,<sup>1</sup> Tyler Murray,<sup>1</sup> Hsu-Han Ooi,<sup>1</sup> Matthew Peters,<sup>1</sup> Joanna Power,<sup>1</sup> Sam Skjonsberg,<sup>1</sup> Lucy Lu Wang,<sup>1</sup> Chris Wilhelmi,<sup>1</sup> Zheng Yuan,<sup>1</sup> Madeline van Zuylen,<sup>1</sup> and Oren Etzioni<sup>1</sup>  
[waleed@allenai.org](mailto:waleed@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 380M nodes, including papers, authors, entities, and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction to familiar NLP tasks (e.g., entity extraction) and 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 available at [www.semanticscholar.org](http://www.semanticscholar.org).

**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?
- Who is the most authors published one or more papers on protein localization?
- Which papers discuss the effects of Ramizanib 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., Lyt et al., 2017) or to improve results in academic search (e.g., Xiong et al., 2014).

2017). We describe methods used in a scalable deployed production system for organizing 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 questions listed earlier as well as many others.

For example, given a query 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. To do so, the system needs to check if all edges on the path are labeled “author”.

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and named entity recognition, and therefore the practical 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., 2016).

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



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 literature into a  
heterogeneous graph to facilitate  
algorithmic manipulation and discovery.  
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 Grunewald, Chandra Bhagavatula, Iz Bergia, Miles Crawford, Doug Downey\*, Jason Dunkelberger, Ahmed Elgashay, 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 Wilhelmi, Zheng Yuan\*, Madeline van Zuylen, and Oren Etzioni  
waleed@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 380M nodes, including papers, entities, and authors, 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), pointing 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 available at [www.semantic-scholar.org](http://www.semantic-scholar.org).

**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?
- Who are my co-authors published one or more papers on preference evolution?
- Which papers discuss the effects of Rambutan on the Retina?

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and entity extraction, and address some of the practical 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., 2017).

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).

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

84

**Figure 1: Part of the literature graph.**

**Construction of the Literature Graph in Semantic Scholar**

Waleed Ammar, Dirk Grunewald, Chandra Bhagavatula, Iz Bergia, Miles Crawford, Doug Downey\*, Jason Dunkelberger, Ahmed Elgashay, 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 Wilhelmi, Zheng Yuan\*, Madeline van Zuylen, and Oren Etzioni  
waleed@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 380M nodes, including papers, entities, and authors, 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), pointing 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 available at [www.semantic-scholar.org](http://www.semantic-scholar.org).

**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?
- Who are my co-authors published one or more papers on preference evolution?
- Which papers discuss the effects of Rambutan 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).

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

84

**Figure 1: Part of the literature graph.**

{

**"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 / PDF or Scans

Document Layout

Output / Metadata JSON

# Layout Parser

# Parse document layouts with Deep Learning

## Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Grunenwald, Chandra Bhagavatula, Le Retina, Miles Crawford, Doug Downey,<sup>†</sup> Jason Dunkelberger, Ahmed Elgashry, 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 Wilhelmi, Zheng Yuan,<sup>\*</sup> Madeleine Zuylen, and Oren Etzioni  
[waleed@allenai.org](mailto:waleed@allenai.org)

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA  
<sup>†</sup>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 380M nodes, including papers, entities, and mentions, and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction to familiar NLP tasks (e.g., entity extraction) but highlight points of 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 available in semantic features in [www.semanticscholar.org](http://www.semanticscholar.org).

### 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 protein phosphorylation?
- Which papers discuss the effects of Ramipril on the Retina?

We reduce literature graph construction into familiar NLP tasks such as sequence labeling, entity linking and relation extraction, and address some of the practical 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. ©2018 Association for Computational Linguistics

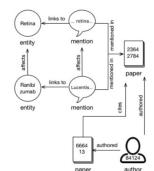


Figure 1: Part of the literature graph.

# L<sup>P</sup> Layout Parser

# Layout Detection Models

## Construction of the Literature Graph in Semantic Scholar

Waleed Ammar, Dirk Grunenwald, Chandra Bhagavatula, Le Retina, Miles Crawford, Doug Downey,<sup>†</sup> Jason Dunkelberger, Ahmed Elgashry, 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 Wilhelmi, Zheng Yuan,<sup>\*</sup> Madeleine Zuylen, and Oren Etzioni  
[waleed@allenai.org](mailto:waleed@allenai.org)

Allen Institute for Artificial Intelligence, Seattle WA 98103, USA  
<sup>†</sup>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, including papers, entities, and mentions, and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction to familiar NLP tasks (e.g., entity extraction) but highlight points of 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 available in semantic features in [www.semanticscholar.org](http://www.semanticscholar.org).

### 1 Introduction

The goal of our 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 protein phosphorylation?
- Which papers discuss the effects of Ramipril 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.,

84

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

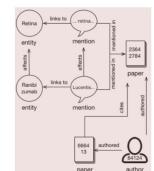


Figure 1: Part of the literature graph.

Input / PDF or Scans

Intermediate / Document Layout

# Layout Parser

## Support data extraction based on layouts

**Construction of the Literature Graph in Semantic Scholar**

Waleed Ammar, Dirk Grunwald, Chandra Bhagavatula, Iza Belegny, Miles Crawford, Doug Downey\*, Jason Dinkhaberger, Ahmed Elshafy, 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\*, Madeline van Zuylen, and Oren Etzioni  
waleed@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 380M nodes, including papers, authors, and entities, and various interactions between them (e.g., authorships, citations, entity mentions). We reduce literature graph construction to familiar NLP tasks (e.g., entity extraction) and highlight potential 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 available as semantic features in our [semanticscholar.org](http://www.semanticscholar.org).

**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?
- Who is one co-author published one or more papers on certificate resolution?
- Which papers discuss the effects of Ramibizumab 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., Lyt et al., 2017) or to improve ranks of results in academic search (e.g., Xiong et al., 2018).

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

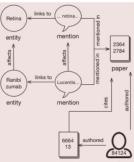


Figure 1: Part of the literature graph.

**lp Layout Parser**

**Layout Data APIs**

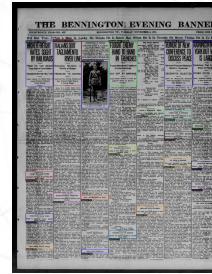
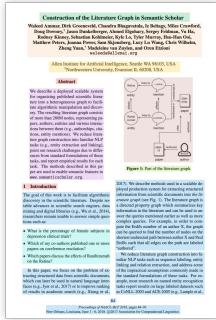
{  
"title": "Construction of the Literature Graph in Semantic Scholar",  
"authors": "Waleed Ammar et. al.",  
"abstract": "We describe a deployed scalable system for organizing published scientific literature into a heterogeneous graph to facilitate algorithmic manipulation and ...",  
"sections":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

**Intermediate / Document Layout**

**Output / Metadata JSON**

# Generalizability?

## How to adapt to new different data quickly?

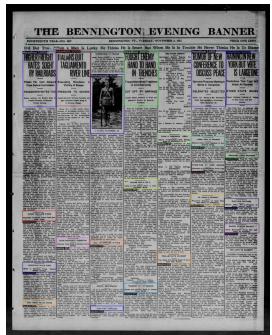


**Layout Model  
for Document Type A**

**New Data  
for Document Type B**

# *Layout Parser + Label Studio Annotation*

**Customizing models → better accuracy on new data**



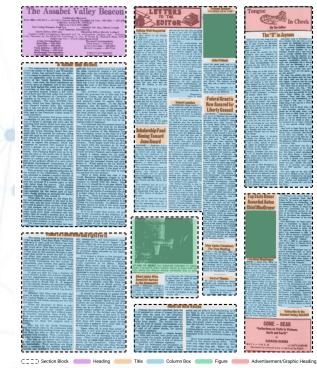
**Data Annotation**

**Train/Eval  
Dataset**



**Model Training**

**New Doc  
Data**



**Apply new  
models**

# 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](#)