Semantic Scholar

Document Analysis at Scale

Miles Crawford, Director of Engineering
Outline

- Introduction to the Allen Institute for Artificial Intelligence and Semantic Scholar
- Research at Semantic Scholar
- Creating www.semanticscholar.org
- Other resources for researchers
Introduction to AI2 and S2
Allen Institute for Artificial Intelligence

Mosaic
Common Sense Knowledge and Reasoning

Aristo
Machine Reading and Question Answering

Euclid
Math and Geometry Comprehension

AllenNLP
Deep Semantic NLP Platform

PRIOR
Visual Reasoning

Semantic Scholar
AI-Based Academic Knowledge
Semantic Scholar: Vision & Strategy

Semantic Scholar makes the world's scholarly knowledge easy to survey and consume.
Semantic Scholar: Vision & Strategy

**Differentiation:**
S2 is dramatically better at surveying, extracting, and helping researchers consume the most relevant information from the world's research.

**Scale:**
Attract and retain a significant and sustainable share of academic search traffic.

**Impact on research with AI:**
Work towards a “Wright Brothers” moment for research through research on novel AI techniques that are prototyped with millions of active users.
Research at Semantic Scholar
Research at S2: Three Levels of Analysis

Paper

Relationships

Macro
The addition of MbPA reaches a test perplexity of 29.2 which is, to the authors’ knowledge, state-of-the-art at time of writing.
Extract meaningful structures

Attention Is All You Need
Ashish Vaswani, Noam Shazeer, ..., Illia Polosukhin - Published 2017 in NIPS

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks. These models also connect the input and output through a single network, which can handle the dependencies with recurrence and non-linearity. However, these models can be super-sensitive to errors in the input sequence.

1. Our model achieves 26.4 BLEU on the WMT 2014 English-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.3 after training for 4.8 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Japanese Dependency Analysis using Cascaded Chunking
Taku Kudo, Yujia Matusov - Published 2012 in CoNLL

In this paper, we propose a new statistical Japanese dependency parser using a cascaded chunking model. Conventional Japanese statistical dependency parsers are mainly based on a probabilistic model, which is not always efficient or scalable. We propose a new method that is simple and efficient, since it parses a sentence deterministically only deciding whether the current segment modifies the segment on its immediate right hand side. Experiments using the Kyoto University Corpus show that the... CONTINUE READING

From This Paper
Figures, tables, and topics from this paper:

Figure 1
Table 1
Figure 2
Table 2
Figure 3
Table 3
Figure 4
Table 4
Relationships: Establishing Connections

- Ammar et al. NAACL 2018 -- Construction of the Literature Graph in Semantic Scholar
- Wang et al. BioNLP 2018 -- Ontology Alignment in the Biomedical Domain Using Entit...
- Bhagavatula et al. NAACL 2018 -- Content-Based Citation Recommendation
Establishing Connections

Although dependency accuracy of our CFG cannot reach the level it would exceed if disambiguation in the subsequent processing.

Since Cabocha uses the same tagset as the RWC corpus, we use this.

Dependency analysis is preferred in order to analyze Japanese (Kurohashi and Nagao, 1998; Uchimoto et al., 2014) have been conducted.

As seen from Table 1, accuracy is still lower than KNP and CTA correctly in the subsequent processing.

In this case, “NEAREST” means only PGLR model was used for Segmentation Dependency Sentence NEAREST 65.68%, 87.88%, 92.88%, 64.48% 0 10 20 30 40 50 60 70 80 90 100 Rank Accuracy (NEAREST) Dependency Accuracy (BEST) Sentence Accuracy (BEST) etc.) On the other hand, “BEST” means only disambiguation...
Relationships: Establishing Connections

Neural machine translation

Neural machine translation (NMT) is the approach to machine translation in which a large neural network is trained to maximize translation. (More)

Papers overview

Semantic Scholar uses AI to extract papers important to this topic.

Six Challenges for Neural Machine Translation
Philipp Koehn, Rebecca Knowles • NMT@ACL • 2017
We explore six challenges for neural machine translation: domain mismatch, amount of training data, rare words, long sentences... (More)

Nematus: a Toolkit for Neural Machine Translation
Rico Sennrich, Orhan Firat, +8 authors • Naïdeide • EACL • 2017
We present Nematus, a toolkit for Neural Machine Translation. The toolkit prioritizes high translation accuracy, usability, and... (More)

Related topics

- Artificial neural network
- Recurrent neural network
- Statistical machine translation

Broader (3)

- Computational linguistics
- Computer-assisted translation
- Machine translation

Related mentions per year

 topic mentions per year

2014 2018

Associated With Malfunction
Of Gene Product

AMINO ACID, PEPTIDE, OR PROTEIN
alpha-Synuclein
Apolliprotein E
Apolliprotein E4
APP protein, human
caspase-3
Cathepsin B
CLU protein, human

MANIAC protein, human
Microtubule-Associated Protein Tau
MT3 protein, human
Phosphorylated-Binding
Clathrin Assembly Protein
Presenlin-1
Presenlin-2

Manifestation Of
DISEASE OR SYNDROME
Down Syndrome

May Be Treated By

CLINICAL DRUGS
24 HR rivastigmine 0.192 MG/Hr
Transdermal System
alpha-Tocopherol Acetate
Dipiridamol Sodium
Donepezil Hydrochloride 18 MG Oral Tablet
donepezil
Galantamine
Galantamine 8 MG Oral Tablet
rivastigmine
rivastigmine 1.5 MG Oral Capsule
rivastigmine tartrate
Selegiline
selgline Hydrochloride
Sodium Valproate

CHEMICAL
Donepezil hydrochloride
galantamine hydrobromide
tacrine
Tacrine Hydrochloride
Valproic Acid

VITAMIN
alpha-Tocopherol Acetate
tocopherols
Vitamin E
Macro: Trends in research

- pre-publishing vs. citation rates
- understanding peer reviews
- quantify demographic bias in clinical trials

Kang et al. NAACL 2018 -- A Dataset of Peer Reviews (PeerRead): Collection, Insights...
Feldman et al. arXiv 2018 -- Citation Count Analysis for Papers with Preprints
## 2018 Publications

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Conference/Preprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications</td>
<td>Kang, Ammar, Dalvi, van Zuylen, Kohlmeier, Hovy, Schwartz</td>
<td>NAACL 2018</td>
</tr>
<tr>
<td>Content-Based Citation Recommendation</td>
<td>Bhagavatula, Feldman, Power, Ammar.</td>
<td>NAACL 2018</td>
</tr>
<tr>
<td>Construction of the Literature Graph in Semantic Scholar</td>
<td>Ammar, Groeneveld, Bhagavatula, Etzioni</td>
<td>NAACL 2018</td>
</tr>
<tr>
<td>Ontology Alignment in the Biomedical Domain Using Entity Definitions and Context</td>
<td>Wang, Bhagavatula, Neumann, Lo, Wilhelm, Ammar</td>
<td>BioNLP 2018</td>
</tr>
<tr>
<td>Extracting Scientific Figures with Distantly Supervised Neural Networks</td>
<td>Siegel, Lourie, Power, Ammar</td>
<td>JCDL 2018</td>
</tr>
<tr>
<td>Neural Relation Extraction using a Combination of Full Supervision and Distant Supervision</td>
<td>Beltagy, Lo, Ammar</td>
<td>arXiv preprint 2018</td>
</tr>
<tr>
<td>Citation Count Analysis for Papers with Preprints</td>
<td>Feldman, Lo, Ammar</td>
<td>arXiv preprint 2018</td>
</tr>
</tbody>
</table>
Bringing it All to Production
semanticscholar.org: Scale and Reach

- Comprehensive literature graph to explore papers, authors, topics.
- View and search across 40MM+ Computer Science & BioMedical academic papers.
- Expansion to all of science in 2019.
- Topic extraction covering 350k+ topics, 230MM+ mentions, and 6.7MM+ relationships.
- Used by over 2.2MM unique users per month.
- Global adoption.
Normalizing Paper Records

Data Acquisition
- Crawler
- PubMed
- Springer
- Etc...

Sourced Paper
- Title: N/A
- Authors: N/A
- PDF: a2ho-4.pdf etc...

Sourced Paper
- Title: Analyzing...
- Authors: A, B, C
- PDF: N/A etc...

Sourced Paper
- Title: Findings of...
- Authors: A, B, C
- PDF: findings.pdf etc...

Amazon S3
Pipeline

Junk Filter

De-duplicate & Cluster

Create Citation Graph

Join Extracted Data

Create Search Index

Distributed Paper Resolution

Title Block: analyzescientificdocument

Title: Analyzing Scientific Documents
Author: John Doe

Title: Analyzing scientific documents
Author: J. Doe

Title: Analyze Scientific Documents
Author: Jane Smith

Title: Analyzing Scientific Documents
Author: John Doe

Title: Analyze Scientific Documents
Author: Jane Smith

Title Block: deeplearningfornlp

Title: Deep Learning for NLP
Author: Jane Smith

Title: Deep Learning for NLP
Author: Jane Smith
Distributed Citation/Reference Matching

Title: Structured Content Extraction
Author: John Doe

References:
1. Analyzing Scientific Documents, Doe et al., SCIDOCA 2016
2. Deep Learning with NLP, Smith, NAACL 2017
Other Resources
Open Research Corpus

- Full Corpus of Papers:
  - Some data restricted due to some licensing/copyright conditions.
  - Record-per-line JSON archive, 36GB.
  - [http://labs.semanticscholar.org/corpus/](http://labs.semanticscholar.org/corpus/)

- API:
  - Lightweight single paper or author access.
  - JSON over REST API.
  - Link resolver for easy linking to Semantic Scholar.
  - [http://api.semanticscholar.org/](http://api.semanticscholar.org/)
Use of Open Corpus as Training Set

An independent researcher created a system for embedding papers as a vector to support similarity measurements and recommendation generation.

https://github.com/Santosh-Gupta/Research2Vec
ArXiv Added a Bibliography via API

ArXiv used our API to fetch and re-display citation and reference lists on their pages.
Open Source Repositories

**Cite-o-matic**: Find relevant citations for drafts of papers.  
https://github.com/allenai/citeomatic

**Science Parse v1**: CRF-based extraction of metadata from PDFs.  
https://github.com/allenai/science-parse

**Science Parse v2**: LSTM-based extraction of metadata from PDFs.  
https://github.com/allenai/spv2

**Deep Figures**: Extract figures and tables from papers.  
https://github.com/allenai/deepfigures-open
What would you like to see?
Thank you!

milesc@allenai.org

http://allenai.org

https://semanticscholar.org
Appendix
1st dataset of peer reviews

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi Mishra, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, Roy Schwartz. *NAACL 2018*

<table>
<thead>
<tr>
<th></th>
<th>ICLR</th>
<th>cs.cl</th>
<th>cs.lg</th>
<th>cs.ai</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>57.6</td>
<td>68.9</td>
<td>67.9</td>
<td>92.1</td>
</tr>
<tr>
<td>Ours (Δ)</td>
<td>65.3</td>
<td>75.7</td>
<td>70.7</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>+7.7</td>
<td>+6.8</td>
<td>+2.8</td>
<td>+0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presentation format</th>
<th>Oral</th>
<th>Poster</th>
<th>Δ</th>
<th>stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recommendation</td>
<td>3.83</td>
<td>2.92</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>Substance</td>
<td>3.91</td>
<td>3.29</td>
<td>0.62</td>
<td>0.84</td>
</tr>
<tr>
<td>Clarity</td>
<td>4.19</td>
<td>3.72</td>
<td>0.47</td>
<td>0.90</td>
</tr>
<tr>
<td>Meaningful comparison</td>
<td>3.60</td>
<td>3.36</td>
<td>0.24</td>
<td>0.82</td>
</tr>
<tr>
<td>Impact</td>
<td>3.27</td>
<td>3.09</td>
<td>0.18</td>
<td>0.54</td>
</tr>
<tr>
<td>Originality</td>
<td>3.91</td>
<td>3.88</td>
<td>0.02</td>
<td>0.87</td>
</tr>
<tr>
<td>Soundness/Correctness</td>
<td>3.93</td>
<td>4.18</td>
<td>-0.25</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Prepublishing before paper gets accepted is associated with 65% more citations!

(a) Histogram of the number of paper citations in the year following a conference.

(b) Histogram of the difference between the arXiv submission date and the conference deadline.

Sergey Feldman, Kyle Lo, Waleed Ammar. arXiv 2018
Gender bias in clinical trials

- arthritis
- asthma
- cancer
- cardiovascular diseases
- depression
- diabetes
- healthy
- hiv
- obesity
- pain
Cite-o-matic

Table 2: F1@20 and MRR results for two baselines and three variants of our method. BM25 results are based on our implementation of this baseline, while ClusCite results are based on the results reported in Ren et al. (2014). “NNSelect” ranks candidates using cosine similarity between the query and candidate documents in the embedding space (phase 1). “NNSelect + NNRank” uses the discriminative reranking model to rerank candidates (phase 2), without encoding any of the metadata features. “+ metadata” encodes the metadata features (i.e., keyphrases, venues and authors), achieving the best results on all datasets. Mean and standard deviations are reported based on five trials.

Chandra Bhagavatula, Sergey Feldman, Russell Power, Waleed Ammar. NAACL 2018
Figure 1: Part of the literature graph.

Ontology alignment

Figure 2: Siamese network architecture for computing entity embeddings for each source and target entity in a candidate entity pair.
Context-sensitive word representations

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_1 \pm \text{std}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chiu and Nichols (2016)</td>
<td>90.91 ± 0.20</td>
</tr>
<tr>
<td>Lample et al. (2016)</td>
<td>90.94</td>
</tr>
<tr>
<td>Ma and Hovy (2016)</td>
<td>91.37</td>
</tr>
<tr>
<td>Our baseline without LM</td>
<td>90.87 ± 0.13</td>
</tr>
<tr>
<td>TagLM</td>
<td>91.93 ± 0.19</td>
</tr>
</tbody>
</table>

Table 1: Test set $F_1$ comparison on CoNLL 2003 NER task, using only CoNLL 2003 data and unlabeled text.

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_1 \pm \text{std}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang et al. (2017)</td>
<td>94.66</td>
</tr>
<tr>
<td>Hashimoto et al. (2016)</td>
<td>95.02</td>
</tr>
<tr>
<td>Søgaard and Goldberg (2016)</td>
<td>95.28</td>
</tr>
<tr>
<td>Our baseline without LM</td>
<td>95.00 ± 0.08</td>
</tr>
<tr>
<td>TagLM</td>
<td>96.37 ± 0.05</td>
</tr>
</tbody>
</table>

Table 2: Test set $F_1$ comparison on CoNLL 2000 Chunking task using only CoNLL 2000 data and unlabeled text.

---

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power. **ACL 2017**

Waleed Ammar, Matthew E. Peters, Chandra Bhagavatula, Russell Power. **SemEval 2017**
Figure extraction

Next steps

1. Extract Meaningful Structures

2018

Figures
Tables
Entities
Relations
Results

2019

C-peptide [contraindicated with] Diabetes Mellitus

The addition of MbPA reaches a test perplexity of 29.2 which is, to the authors' knowledge, state-of-the-art at time of writing.
Next steps

1. Extract Meaningful Structures

2018

2019

Address more paper FAQs (e.g., PICO elements).

<table>
<thead>
<tr>
<th>Framework Item</th>
<th>Think about:</th>
<th>Example:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>What are the patient's demographics such as age, gender and ethnicity? Or what is the or problem type?</td>
<td>Work-related neck muscle pain</td>
</tr>
<tr>
<td>Problem (or Population)</td>
<td>What type of intervention is being considered? For example is this a medication of some type, or exercise, or rest?</td>
<td>Strength training of the painful muscle</td>
</tr>
<tr>
<td>Intervention</td>
<td>Is there a comparison treatment to be considered? The comparison may be with another medication, another form of treatment such as exercise, or no treatment at all.</td>
<td>Rest</td>
</tr>
<tr>
<td>Comparison or Control</td>
<td>What would be the desired effect you would like to see? What effects are not wanted? Are there any side effects involved with this form of testing or treatment?</td>
<td>Pain relief</td>
</tr>
</tbody>
</table>

The addition of MbPA reaches a test perplexity of 29.2 which is, to the authors' knowledge, state-of-the-art at time of writing.
Next steps

1. Extract Meaningful Structures

2018

- Figures
- Tables
- Entities
- Relations
- Results

2019

Model salience of extracted structures at the document level.

Topics
- Sentiment analysis
- Question answering
- Bidirectional transformation
- Textual entailment
- Text corpus
- Language model
- Word embedding
- Downstream (software development)
Next steps

1. Extract Meaningful Structures

2018

User’s implicit/explicit feedback.

2019

Topics
- Sentiment analysis
- Question answering
- Bidirectional transformation
- Textual entailment

- Text corpus
- Language model
- Word embedding
- Downstream (software development)
Next steps

1. Extract Meaningful Structures

2018

Tables

Figures

Entities

Relations

Results

2019

User’s implicit/explicit feedback.

Extracted Numerical Results

1. We found $T = 500$, $y = 0.25$ worked best, achieving a test perplexity of 34.3 on this dataset (Table 1).

2. The cache reduces the test perplexity by 1.6 for the LSTM and 4.4 for LSTM + Hebbian Softmax. The addition of MdpA reaches a test perplexity of 29.2 which is, to the authors’ knowledge, state-of-the-art at time of writing.

3. We observe a 9.8-point drop in perplexity, from 53.5 to 43.7, illustrated in Table 4.2.2.
Next steps

1. Extract Meaningful Structures

2018

Figures

Tables

Entities

Relations

Results

2019

User’s implicit/explicit feedback.

Hello Waleed Ammar,

Semantic Scholar is working on building a better search engine for academic papers. Your assistance as an author will greatly help us accelerate our development of machine learning models that can better process papers like yours. Please take a moment to answer the question below. We appreciate your input!

In your paper: "Towards a Universal Analyzer of Natural Languages", does the following sentence express a numerical result?

“For example, we use the block dropout to teach the parser to ignore the predicted POS tag embeddings all the time at first by initializing μ = 1.0 (i.e., always dropout, setting l = 0), and dynamically update μ to match the error rate of the POS tag”

[Radio button choices: YES, NO]
Next steps

1. Extract Meaningful Structures

2018

Adapt existing models to more domains.

2019

Ganin and Lempitsky. ICML’15
Next steps

2. Establish Connections

2018

2019
Next steps

2. Establish Connections

2018

More work on ontology matching and deduplication.

2019
Next steps

2. Establish Connections

2018

2019

More work on author disambiguation.
Next steps

2. Establish Connections

2018

2019

Implement v1.0 of Knowledge Graph Querying

Example query: find researchers with first-author publications in ACL or NIPS since 2017, worked on knowledge base completion or question answering since 2010, and are affiliated with a university in the US.
Next steps

3. Macro Analysis

2018

- pre-publishing vs. citation rates
- understanding peer reviews
- quantify demographic bias in clinical trials

2019
Next steps

3. Macro Analysis

2018

- pre-publishing vs. citation rates
- understanding peer reviews
- quantify demographic bias in clinical trials

2019

Continue analyzing biases in clinical trials.
Next steps

3. Macro Analysis

2018

- pre-publishing vs. citation rates
- understanding peer reviews
- quantify demographic bias in clinical trials

2019

Study influence of government vs. industry funding.
Next steps

3. Macro Analysis

2018

pre-publishing vs. citation rates
understanding peer reviews
quantify demographic bias in clinical trials

2019

Study evolution and influence of research communities.

AKBC 2019
1st Conference on Automated Knowledge Base Construction (AKBC) 2019
in Amherst, Massachusetts, May 20-22nd, 2019.