IE tasks in CiteSeer

Compiled by Sujatha Das, Cornelia Caragea

Expertise Modeling for Matching Papers with Reviewers

David Mimno, Andrew McCallum
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA
{mimno,mccallum}@cs.umass.edu

ABSTRACT

An essential part of an expert-finding task, such as matching reviewers to submitted papers, is the ability to model the expertise of a person based on documents. We evaluate several measures of the association between a document to be reviewed and an author, represented by their previous papers. We compare language-model-based approaches with a novel topic model, Author-Persona-Topic (APT). In this model, each author can write under one or more "personas," which are represented as independent distributions over hidden topics. Examples of previous papers written by prospective reviewers are gathered from the *Rexa* database, which extracts and disambiguates author mentions from documents gathered from the web. We evaluate the models using a reviewer matching task based on human relevance judgments determining how well the expertise of proposed reviewers matches a submission. We find that the APT topic model outperforms the other models.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms

Experimentation

Keywords

Topic models, reviewer finding, expert retrieval

1. INTRODUCTION

Peer review is part of the foundation of the scientific method, but matching papers with reviewers can be a challenging process. The process is also a significant, time-consuming burden on the conference chair. There has been a recent trend towards bidding on submissions by reviewers, which consumes additional reviewer time, as well as raising questions about the confidentiality of the submissions process.

Matching papers with reviewers is a complicated task, with many sub-problems. Conference chairs must solve a large optimization problem involving constraints on the number of reviewers per paper and the number of papers per reviewer. One of the most important elements of the process, however, is modeling the expertise of a given reviewer with respect to the topical content of a given paper. This task is related to "expert finding," an area that has received increased interest in recent years in the context of the TREC Enterprise Track. In addition, for several years researchers in artificial intelligence have sought to automate, or at least streamline, the reviewer matching process.

In this paper, we evaluate several methods for measuring the affinity of a reviewer to a paper. These methods include language models with Dirichlet smoothing [Ponte and Croft, 1998, Zhai and Lafferty, 2001], the Author-Topic model [Rosen-Zvi et al., 2004], and a novel topic model, Author-Persona-Topic (APT).

We follow previous approaches in treating expert finding as an information retrieval task. The goal is to find relevant people rather than relevant documents, but we use the same basic tools. More specifically, we construct a model in which each potential reviewer has a distribution over words in the vocabulary, and then rank reviewers for a given paper based on the likelihood of the words in that paper under each reviewer's distribution. In this paper we evaluate several methods for constructing such models.

In order to discover expertise, it is necessary to consider how to represent expertise. Statistical topic models represent documents as mixtures of topical components, which are distributions over the words in the corpus. The APT model is motivated by the observation that authors frequently write about several distinct subject area combinations. It is rare that a person is an expert in all facets of a single topic. For example, even a topic as narrow as support vector machines is sufficiently rich and complex that almost no one would claim expertise in all facets of their use and theory.

People usually describe their expertise as the combination of several topics, and often have experience in several such intersections. The second author, for example, has expertise in Bayesian networks and language, and reinforcement learning and hidden states, but not in reinforcement learning and language, a combination used in dialog systems. Other examples of such topical intersections include game theory and Bayesian networks or information retrieval and algorithms. In the APT model, we not only learn the topical components, but also divide each author's papers into sev-

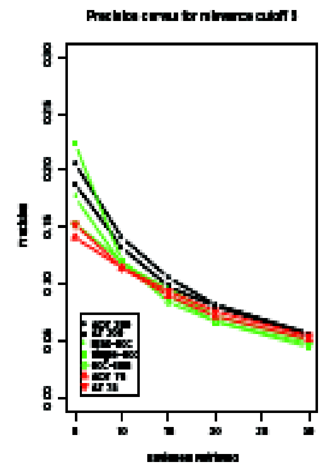


Figure 3: The precision of each model as more documents are retrieved for relevance cutoff 3. The same general patterns are present at this level of relevance as in the lower-cutoff evaluation. The topic models with 200 topics are the best overall, while the single-document author language model has the highest precision in the first five reviewers retrieved.

Table 2: Precision at relevance cutoff ≥ 2 after retrieving n reviewers.

Model	5	10	15	20	30
APT 200	0.4118	0.2971	0.2255	0.1824	0.1294
AT 200	0.3882	0.2765	0.2176	0.1794	0.1265
max-doc	0.3471	0.2500	0.1980	0.1588	0.1147
single-doc	0.4471	0.2735	0.1980	0.1529	0.1059
doc-sum	0.3412	0.2500	0.1882	0.1529	0.1118
APT 75	0.3059	0.2588	0.1961	0.1618	0.1176
AT 75	0.3529	0.2588	0.2020	0.1632	0.1275

Table 3: Precision at relevance cutoff 3 after retrieving n reviewers.

Model	5	10	15	20	30
APT 200	0.2059	0.1412	0.1059	0.0824	0.0569
AT 200	0.1882	0.1324	0.0980	0.0809	0.0549
max-doc	0.1765	0.1176	0.0961	0.0721	0.0510
single-doc	0.2235	0.1206	0.0902	0.0676	0.0451
doc-sum	0.1529	0.1206	0.0843	0.0676	0.0480
APT 75	0.1412	0.1147	0.0902	0.0721	0.0520
AT 75	0.1529	0.1147	0.0941	0.0765	0.0549

substantial performance boost. There are many areas for future work, such as taking advantage of citations and co-authorship data and building language models based on the partition of an author's papers provided by the APT model.

Ultimately, measuring the expertise of a person given a paper is only a part of a system for matching reviewers to papers. It is also necessary to ensure that reviewers receive a reasonable number of papers to review, and that every paper gets a certain minimum number of reviewers. As probabilistic models, the methods described in this paper could fit easily into a larger likelihood function that takes into account the number of reviewers per paper and the number of papers per reviewer. Finding a good matching for the conference as a whole would then be a matter of sampling matchings with high probability from that model.

Matching papers with reviewers is a highly constrained optimization problem. In addition to constraints on the number of papers per reviewer and the number of reviewers per paper, conflicts of interest are common. Indeed, in our experiments, of the top five reviewers retrieved for each paper, 5.0% of those retrieved by the APT model with 200 topics and 4.2% of those retrieved by the single document language model were in fact listed as authors on the paper in question. It is likely that if we removed all prospective reviewers with conflicts of interest, the number of available highly relevant reviewers would be much smaller. This phenomenon suggests that the additional accuracy of the topic modeling approaches at the 10 reviewer level and beyond could be valuable for real world reviewer matching applications.

6. ACKNOWLEDGMENTS

We thank the nine anonymous relevance judges from previous NIPS program committees. Desislava Petkova contributed substantially to the evaluation and the discussion of language models.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by NSF grant # CNS-0551597, in part by NSF Nano # DMI-0531171, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR*, 2006.

C. Basu, H. Hirsh, W. Cohen, and C. Nevill-Manning. Recommending papers by mining the web. In *IJCAI*, 1999.

S. Benferhat and J. Lang. Conference paper assignment. *Int. J. Intell. Syst*, 16:1183, 2001.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR*, 2004.

S. T. Dumais and J. Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *SIGIR*, page 233, 1992.

S. Hettich and M. J. Pazzani. Mining for proposal reviewers: lessons learned at the National Science Foundation. In *KDD*, 2006.

Example of a correctly extracted entry

CiteSeer$^{x}_{\beta}$

Search

Include Citations Advanced Search

Expertise modeling for matching papers with reviewers (2007)

by David Mimno , Andrew Mccallum

Venue: In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

Citations: 18 - 2 self

Save to List
Add to Collection
Correct Errors
Monitor Changes

Cached

Download Links

[www.cs.umass.edu]

Other Repositories/Bibliography

DBLP

Summary | Active Bibliography | Co-citation | Clustered Documents | Version History

Abstract

An essential part of an expert-finding task, such as matching reviewers to submitted papers, is the ability to model the expertise of a person based on documents. We evaluate several measures of the association between an author in an existing collection of research papers and a previously unseen document. We compare two language model based approaches with a novel topic model, Author-Persona-Topic (APT). In this model, each author can write under one or more "personas," which are represented as independent distributions over hidden topics. Examples of previous papers written by prospective reviewers are gathered from the Rexa database, which extracts and disambiguates author mentions from documents gathered from the web. We evaluate the models using a reviewer matching task based on human relevance judgments determining how well the expertise of proposed reviewers matches a submission. We find that the APT topic model outperforms the other models. 1.

Citations

684 A language modeling approach to information retrieval - Ponte, Croft - 1998

121 Retrieval evaluation with incomplete information - Buckley, Voorhees - 2004

109 Topic and role discovery in social networks - McCallum, Corrada-Emmanuel, et al. - 2005

47 Automating the assignment of submitted manuscripts to reviewers - Dumais, Nielsen - 1992

35 Hierarchical language models for expert finding in enterprise corpora - Petkova, Croft - 2006

11 Mining for proposal reviewers: Lessons learned at the national science foundation - HETTICH, J

7 An algorithm to determine peer-reviewers - Rodriguez, Bollen - 2008

BibTeX


```
@INPROCEEDINGS{Mimno07expertisemodeling,
    author = {David Mimno and Andrew
Mccallum},
    title = {Expertise modeling for matching
papers with reviewers},
    booktitle = {In Proceedings of the 13th
ACM SIGKDD International Conference on
Knowledge Discovery and Data Mining},
    year = {2007},
    pages = {500--509}
}
```


Bookmark

CiteSeerX
10.1.1.93.9359

Challenges in CiteSeer

- Since the performance of models trained on CiteSeerX data highly depends on the quality of the data, the CiteSeerX applications require accurate metadata extraction techniques.

- The metadata extraction in CiteSeerx is done using automated techniques.
- Although fairly accurate, metadata extraction in CiteSeerX still results in noisy metadata.

Example of a noisy entry

CiteSeer$^{x}_{\beta}$

Search

Include Citations   Advanced Search

An Architecture for (1993)

by Hong Li , Theodore Williams , Descriptiveness Vs. Prescriptiveness

Venue: IP Address Allocation with CIDR', RFC 1518

Citations: 34 - 0 self

Save to List
Add to Collection
Correct Errors
Monitor Changes

Cached

Download Links

[www.eil.utoronto.ca]

Summary | Active Bibliography | Co-citation | Clustered Documents | Version History

Abstract

� Goranson's comment during ICEIMT'92:

Citations

No citations identified.

BibTeX

```
@INPROCEEDINGS{Li93anarchitecture,
    author = {Hong Li and Theodore
Williams and Descriptiveness Vs.
Prescriptiveness},
    title = {An Architecture for},
    booktitle = {IP Address
Allocation with CIDR', RFC 1518},
    year = {1993}
}
```

Years of Citing Articles

Record Linkage where possible

- Uses information from (external sources) DBLP to automatically filter noise

DBLP provides manually curated metadata and bibliographic information on major computer science journals and proceedings.


```xml
<inproceedings mdate="2007-08-23" key="conf/kdd/MimnoM07">
<author>David M. Mimno</author>
<author>Andrew McCallum</author>
<title>Expertise modeling for matching papers with reviewers.</title>
<pages>500-509</pages>
<year>2007</year>
<crossref>conf/kdd/2007</crossref>
<booktitle>KDD</booktitle>
<ee>http://doi.acm.org/10.1145/1281192.1281247</ee>
<url>db/conf/kdd/kdd2007.html#MimnoM07</url>
</inproceedings>
```

Extracting Researcher Metadata

Researcher metadata available on homepages used in Author Disambiguation, Expert Profiling and Academic Network Extraction

Example

Professor

Andrew McCallum

Computer Science Department
University of Massachusetts Amherst
mccallum@cs.umass.edu
+1 413 545-1323 (vox)
+1 413 545-1789 (fax)

- https://people.cs.umass.edu/~mccallum/

- Sequential tagging for extracting researcher metadata

Treat the homepage content as a list of tokens and annotate each token using one of
{AFFL, EMAIL, FAX, PHN, POS, UNIV, O}

I	am	an	assistant	professor	at	CS	Dept,	Stanford.
O	O	O	POS	POS	O	AFFL	AFFL	UNIV

Feature types and templates for *Sequential Tagging*

1. Canonical term features: lowercase and all punctuation removed
2. Dictionary features: presence in a field-specific dictionary (e.g. phone, department, university, professor)
3. Surface form features: pattern of the token (e.g. onecap, onelower, allcaps, capitalized, digits)

Let F, G represent the above feature-types, i the position of the token.

Unigram features	F_i, $i = \{-2, \ldots 2\}$
Bigram features	$F_{-1}F_0$ and F_0F_1
Skip features	$F_{-1}F_1$
Conjunction features	$F_{-i}G_0$ and F_0G_i, $i = \{-1, 0, 1\}$

Example

Correct:. . .	POS	O	UNIV	O	PHN	PHN	PHN	O
Content:. . .	student	at	Stanford,	phone:	814	321	8184	fax
Positions. . .	3	4	5	6	7	8	9	10

Sample Features

tokenpos=5: stanford, isCapitalized, noDict, phnDict, at, at_stanford

tokenpos=6: phone, lowercase, phnDict, numeric, phnDict_numeric, isCapitalized_phone, isCapitalized_numeric

Error analysis

- Discriminatory cues appears more often with the “O” tag
- Multiple position cues “I am a grad student working with Professor X in....”
- University information appears without cue words (examples: Cornell, IITB, UIUC)
- Too many affiliation “patterns” and some patterns only covered by few examples in the training dataset

  “Department of Computer Science and Engineering
  “Computer Science Department”
  “Electrical Engineering and Computer Science Department”
  “CS Department”

Intuition and questions

Some fields are easy to extract (e.g. phone numbers)

How can we improve annotation performance without adding more labeled examples? Can we harness the easy fields to identify the difficult ones? Use labeled features and a two-step process...

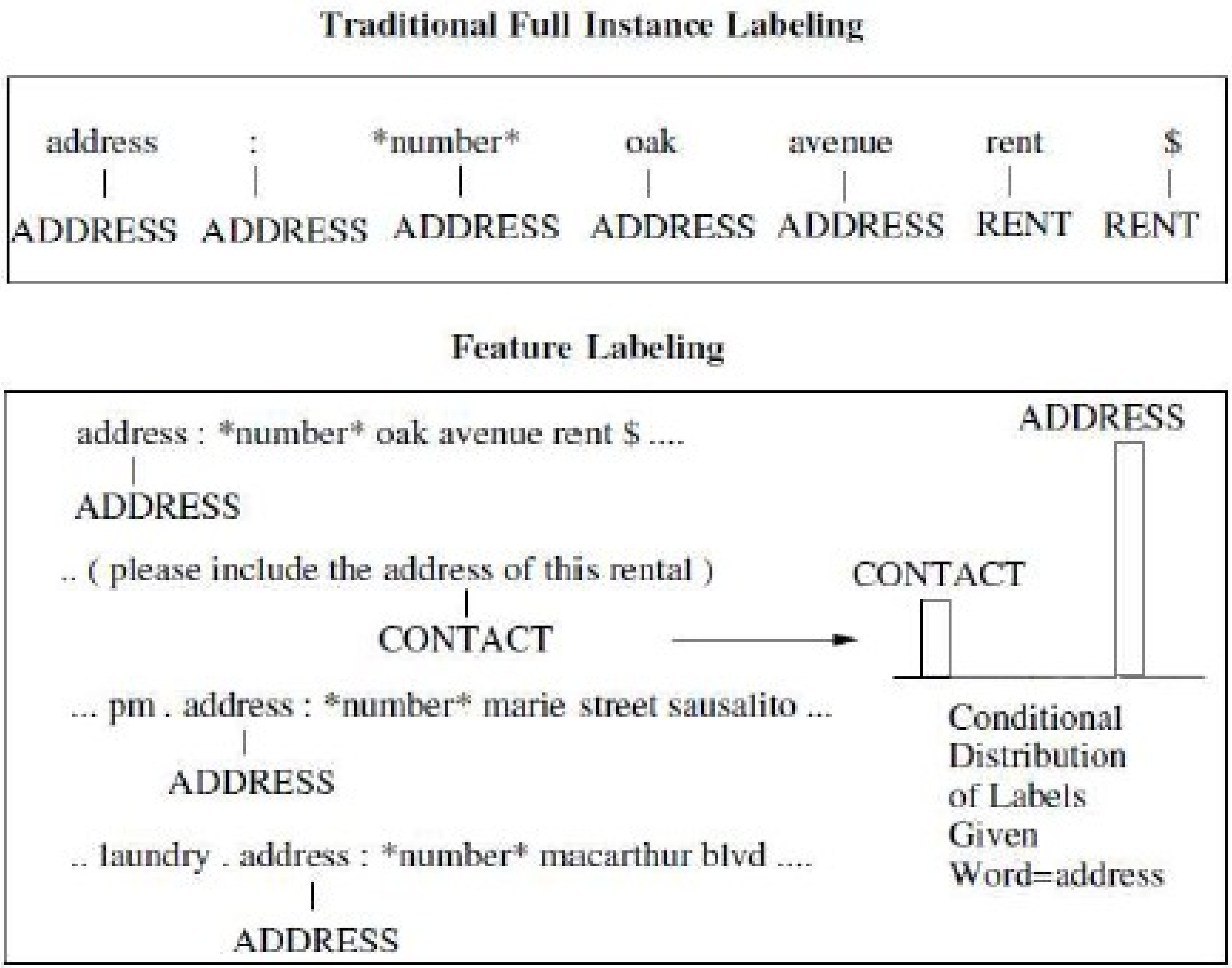
1. Use the basic set of features to train a tagger (Stage - 1).
2. Use predicted tags from the stage-1 tagger as additional features to train a second tagger.

Feature labeling for "weak" supervision

- A new form of supervision studied by Mann, Druck and McCallum [SIGIR 2008, ACL 2008]
- Discriminative classifiers/taggers trained on labeled features than labeled instances
- Labeled Feature = (feature, label-distribution)
- Examples
  - ('department': AFFL=0.9, O=0.1)
    "CS **department**" vs. "courses in our **department**..."
  - ('brown': NOUN=0.7, ADJ=0.3)
    "John **Brown** is running for president" vs. "Little **brown** jug"
  - ('research': POS=0.5, AFFL=0.3, O=0.2)
    "I am a **research** assistant at " vs. "Center for **research** in Plant Genetics" vs "My **research** interests..."

Why feature labeling?

Figure 1 from [ACL 2008]



How to find labeled features?

- Automated techniques based on LDA (Latent Dirichlet Allocation) and InfoGain for classification
- On average, given a feature, domain experts find it easy to pick a majority class [SIGIR 2008]
- For homepage annotation, easy to specify majority label for certain words

Words related to the **AFFL** tag
academy, association, center, centre, school, college, department, dept, dipartimento, division, foundation, group, institut, society institute,lab, laboratories, laboratory

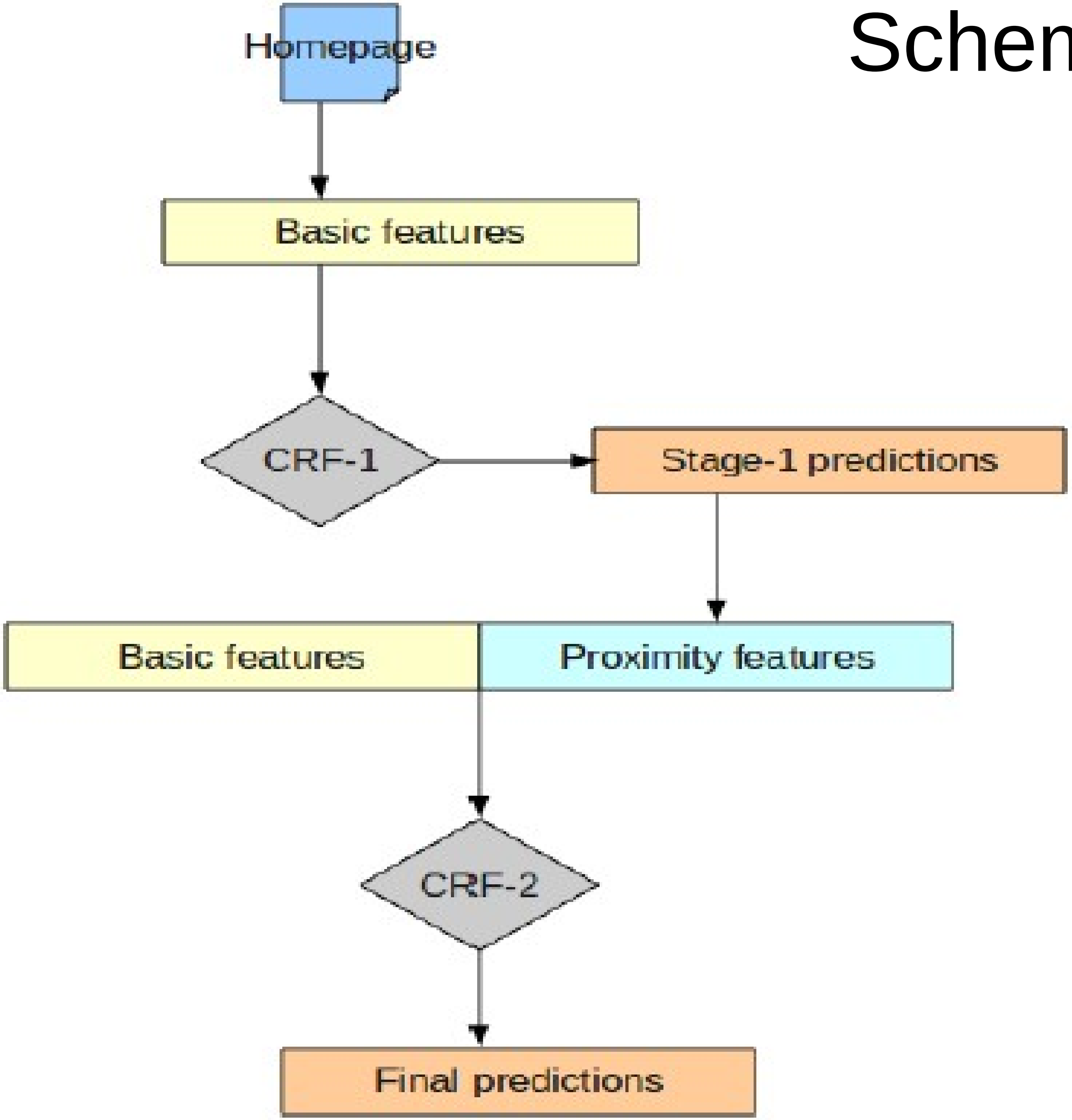
Words related to the **UNIV** tag
universiteit, universitat, university, univ

Big Idea

- Capture the conventions observed in metadata placement.
  Examples:
  - phone and fax information is typically close by
  - position is typically followed by affiliation
- Some fields are easy to annotate than the others

- Stage-1: *First find easy fields like phone/fax/email*
- *Capture layout information using predicted labels from stage-1 as proximity features*
- Stage-2: *Retrain by enforcing proximity as constraints via PR*

Schematic


Homepage
Basic features
CRF-1
Stage-1 predictions
Basic features
Proximity features
CRF-2
Final predictions

Tagging Performance

Setting	**AFFL**	**Agg**	**Agg/O**
`Stage-1`			
Basic	**0.4096**	**0.7479**	**0.7084**
+ dictionary PR	0.5179	0.7742	0.7390
`Stage-2 using Stage-1 preds`			
Basic	0.4697	0.7596	0.7219
+dictionary PR	0.5423	0.7843	0.7507
+window PR	0.5202	0.7689	0.7326
+all constraints PR	0.5359	0.7865	0.7532

Conclusions

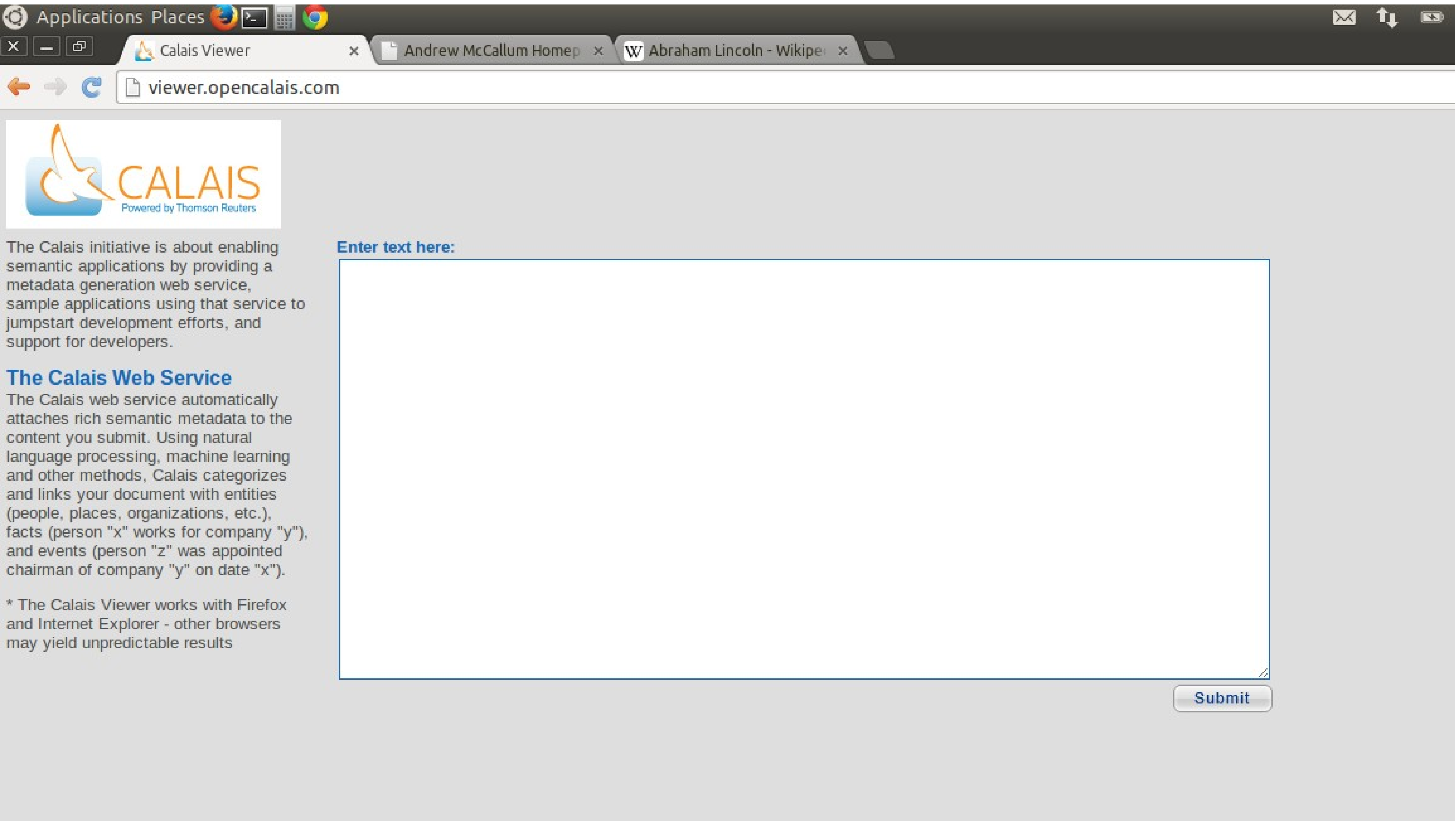
- Several IE tasks in CiteSeer
- Combination of models for extracting metadata
  - NLP knowledge
  - Regex patterns
  - CRFs/SVMs
- The extracted author-document, document-document networks and top-level applications like trend detection, author disambiguation crucially depend on accurate metadata

References

- Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, Zhenyue Zhang, and Edward A. Fox. 2003. Automatic document metadata extraction using support vector machines. JCDL '03
- Sujatha Das Gollapalli, Yanjun Qi, Prasenjit Mitra, C. Lee Giles: Extracting Researcher Metadata with Labeled Features. SDM 2014
- Cornelia Caragea, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, C. Lee Giles: CiteSeer x : A Scholarly Big Dataset. ECIR 2014

OpenCalais Demo

- http://viewer.opencalais.com/



people.cs.umass.edu/~mccallum/

Professor

Andrew McCallum

Computer Science Department
University of Massachusetts Amherst
mccallum@cs.umass.edu
+1 413 545-1323 (vox)
+1 413 545-1789 (fax)

Research

The main goal of my research is to dramatically increase our ability to mine actionable knowledge from unstructured text. I am especially interested in information extraction from the Web, understanding the connections between people and between organizations, expert finding, social network analysis, and mining the scientific literature & community. Toward this end my group develops and employs various methods in statistical machine learning, natural language processing, information retrieval and data mining---tending toward probabilistic approaches and graphical models. For more information see our current projects and publications.

News

- We are building an "open reviewing" system for ICLR 2013 and other venues. If you are interested in alternative approaches to peer review, please talk with me!
- FACTORIE is a toolkit for deployable probabilistic modeling, implemented as a software library in Scala. It provides its users with a succinct language for creating relational factor graphs, estimating parameters and performing inference.
- I was the General Chair of ICML 2012, with Program Chairs Joelle Pineau and John Langford.
- Generalized Expectation is an accurate way to train models by labeling features.

CALAIS
Powered by Thomson Reuters

Show RDF | Entry Page

Topics:

Education	100%
Health Medical Pharma	97%
Technology Internet	97%

Social Tags:

Technology Internet	★★★
Health Medical Pharma	★★★
Education	★★★
Year of birth missing	★★★
University of Massachusetts Amherst	★★★

Professor

Andrew McCallum

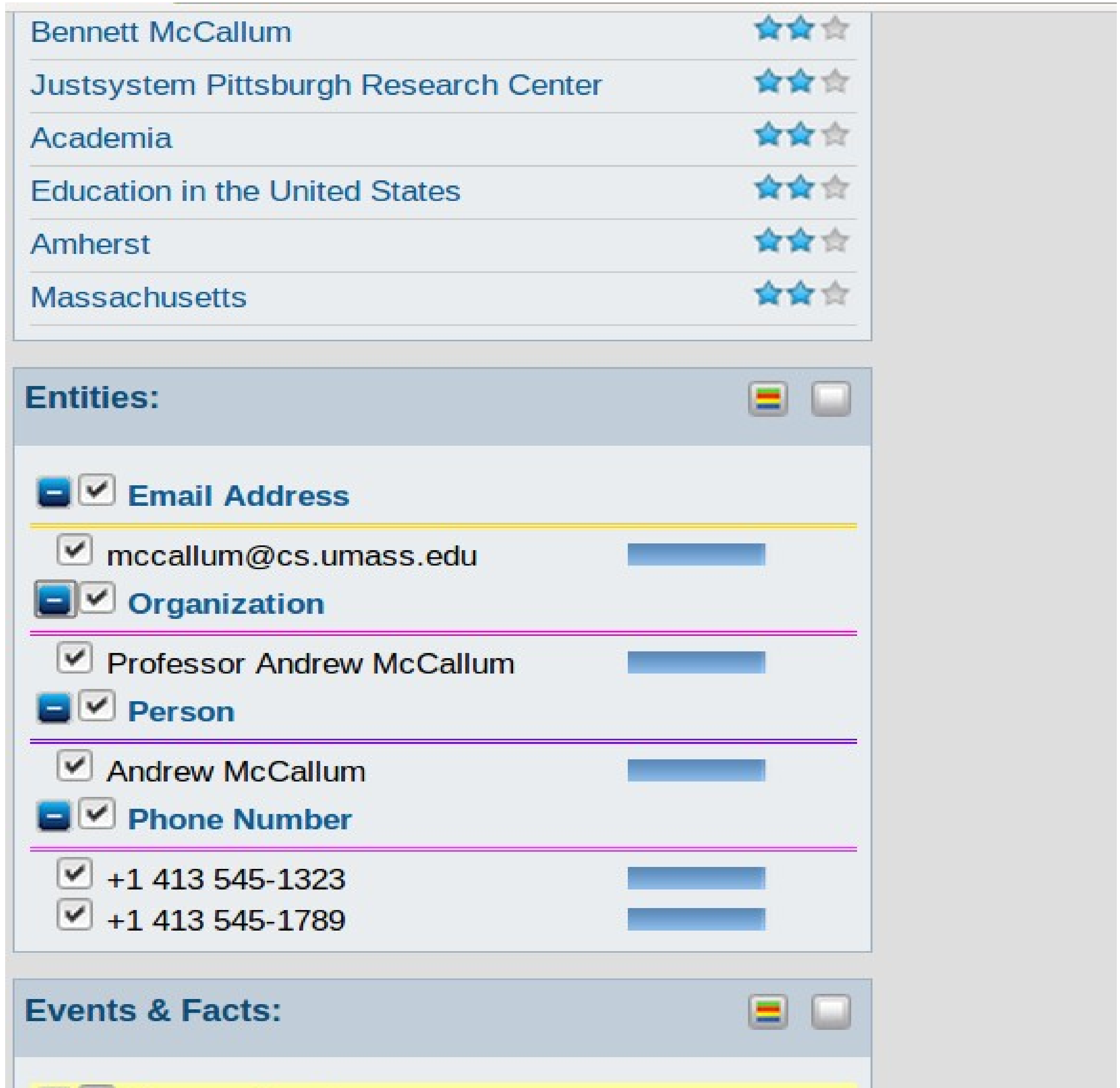
Computer Science Department

University of Massachusetts Amherst

mccallum@cs.umass.edu

+1 413 545-1323 (vox)

+1 413 545-1789 (fax)


Bennett McCallum
Justsystem Pittsburgh Research Center
Academia
Education in the United States
Amherst
Massachusetts
Entities:
Email Address
mccallum@cs.umass.edu
Organization
Professor Andrew McCallum
Person
Andrew McCallum
Phone Number
+1 413 545-1323
+1 413 545-1789
Events & Facts:

en.wikipedia.org/wiki/Abraham_Lincoln

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Article Talk Read View source View history Search

Abraham Lincoln

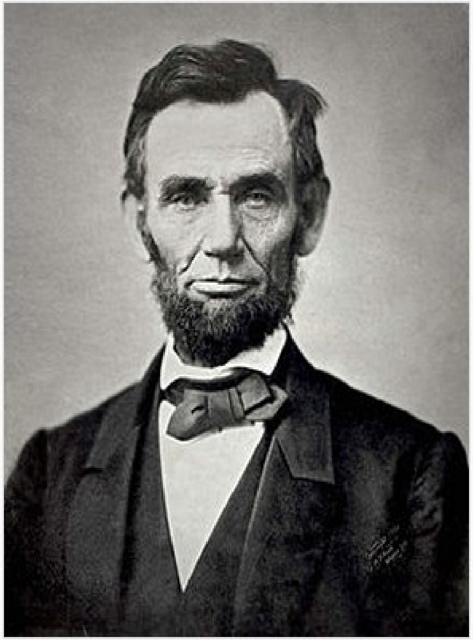
From Wikipedia, the free encyclopedia

This article is about the American president. For other uses, see Abraham Lincoln (disambiguation).

Abraham Lincoln /ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 – April 15, 1865) was the 16th president of the United States, serving from March 1861 until his assassination in April 1865. Lincoln led the United States through its Civil War—its bloodiest war and its greatest moral, constitutional and political crisis.[1][2] In so doing he preserved the Union, abolished slavery, strengthened the federal government, and modernized the economy. Reared in a poor family on the western frontier, Lincoln was a self-educated lawyer in Illinois, a Whig Party leader, state legislator during the 1830s, and a one-term member of the Congress during the 1840s. He promoted rapid modernization of the economy through banks, canals, railroads and tariffs to encourage the building of factories; he opposed the war with Mexico in 1846. After a series of highly publicized debates in 1858, during which Lincoln spoke out against the expansion of slavery, he lost the U.S. Senate race to his archrival, Democrat Stephen A. Douglas. Lincoln, a moderate from a swing state, secured the Republican Party presidential nomination in 1860. With very little support in the slave states, Lincoln swept the North and was elected president in 1860. His election prompted seven southern slave states to form the Confederacy before he took the office. No compromise or reconciliation was found regarding slavery.

When the North enthusiastically rallied behind the national flag after the Confederate attack on Fort Sumter on April 12, 1861, Lincoln concentrated on the military and political dimensions of the war effort. His goal was to reunite the nation. He suspended habeas corpus, arresting and temporarily detaining thousands of suspected secessionists in the border states without trial. Lincoln averted British intervention by defusing the Trent Affair in late 1861. His numerous complex moves toward ending slavery centered on the Emancipation Proclamation in 1863, using the Army to protect escaped slaves, encouraging the border states to outlaw slavery, and helping

Abraham Lincoln



An 1863 daguerreotype of Lincoln, at the age of 54.

16th President of the United States

In office

Abraham Lincoln Listeni/ˈeɪbrəhæm ˈlɪŋkən/ (February 12, 1809 –

April 15, 1865) was the 16th **president of the United States**, serving from March 1861 until **his** assassination in April 1865. **Lincoln** led the **United States** through its Civil Wa

its bloodiest war and its greatest moral, constitutional and political crisis.[1]

[2] In so doing **he** preserved the Union, abolished slavery, strengthened the **federal government**, and modernized the economy. Reared in a poor family on the western frontier,

f-educated lawyer in **Illinois**, a **Whig Party leader, state legislator** during the 1830s, and a one-

term member of the **Congress** during the 1840s. **He** promoted rapid modernization of the economy

e war with **Mexico** in 1846. After a series of highly publicized debates in 1858, during which **Lincol**

ocrat **Stephen A. Douglas**. **Lincoln**, a moderate from a swing state, secured the **Republican Part**

North and was elected **president** in 1860. **His** election prompted seven southern slave states to fo

ng slavery.

Entities:

City

- Lincoln,Illinois,United States

Country

- Mexico
- United States

Organization

- Congress
- federal government
- Republican Party
- United States Senate
- Whig Party

Person

- Abraham Lincoln
- Stephen A. Douglas

Political Event

- Senate race

Position

- leader , state legislator
- president
- president of the United States
- self-educated lawyer