Algorithms and Applications for Web-Scale Knowledge Graphs

Marco Ponza

Supervisor

Prof. Paolo Ferragina

Menu

- Results of the first two years of PhD

1. Document Aboutness *full paper at*

2. Entity Relatedness *full paper at*

- What next? Last year of PhD

1.
Document Aboutness

Based on the full paper

Document Aboutness via Sophisticated Syntactic and Semantic Features

Marco Ponza, Paolo Ferragina and Francesco Piccinno

Introduction

The Document Aboutness Task

Succinct Representation of a Document's Subject Matter (Bruza, AIR '96)

The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

Still, Mr. Obama has been busy telephoning crucial Democrats in Iowa, New Hampshire and other states. There is, of course, only one reason for him to be making such inquiries.

Last week on Capitol Hill, Mr. Obama bumped into Ethel Kennedy, who has been a big admirer. When asked about him, she said, "He can't run soon enough."

Mrs. Clinton, meanwhile, plans to announce her decision in the next several weeks, her advisers say. According to several Democrats who have spoken to her, as well as advisers, Mrs. Clinton has given every indication that she is running, short of saying so, and no signals that she is not.

She is making phone calls to Democratic officials, labor leaders and supporters in early nominating states. And she continues to talk to possible consultants and donors, yet she has not made any travel plans to kick off a campaign. JEFF ZELENY and PATRICK HEALY

Aboutness

Entity	Relevance
Barack_Obama	0.85
Hillary_Clinton	0.8
Hawaii	0.3
George_W._Bush	0.2
...	...

Introduction

The Document Aboutness Task

Succinct Representation of a Document's Subject Matter (Bruza, AIR '96)

Semantic Aboutness Representation

1. Entity ∈ *Knowledge Graph*

   - ▷ Deployment of Entity Linkers (Cucerzan, EMNLP '07)

   ...and many others!

Maradona won against Mexico

Diego_Maradona

Mexico_National_Football_Team

Introduction

The Document Aboutness Task

Succinct Representation of a Document's Subject Matter (Bruza, AIR '96)

Semantic Aboutness Representation

1. Entity $\in$ *Knowledge Graph*

   - Deployment of Entity Linkers (Cucerzan, EMNLP '07)

     ...and many others!

2. Salience (Class or Score)

   - Relevance of an Entity

State-of-the-Art

Competitors & Datasets

- CMU-Google System (Dunietz, EACL '14)

Re-implemented

- Proprietary modules substituted with open-source tools
  - WAT + scikit learn

- SEL (Trani, DocEng '16)

Limitations

- No comparison with CMU-Google System
- Benchmark on small dataset
- Not publicly available

Entity Salience

Our Solution

- Three-Stage System for Entity Salience Extraction
- In-Depth Feature Engineering:
  - Syntactic:
    - Sentence Ranking
    - Dependency Trees
  - Semantic:
    - Entity Annotations
    - Relatedness Graph
- Improves current solutions
  - Up to +9.8%
- The first publicly available API

Entity Salience

General Structure


The New York Times

What? Condoms Can Prevent AIDS? No Way!

Input Document

The Salient Wikipedia Annotator of Text

SWAT

Document Enrichment

1. CoreNLP
2. TextRank
3. WAT

Feature Generation

Features

1. Basics
2. Syntactic
3. Semantic

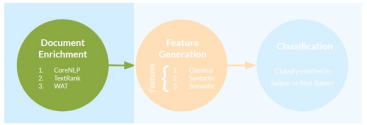
Classification

Classify entities in *Salient* or *Non-Salient*

Salient Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

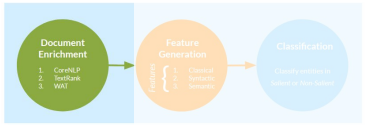
Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

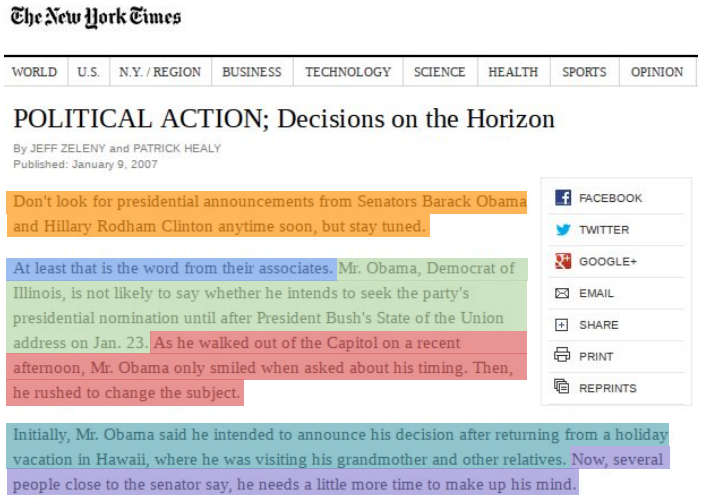
1. Document Enrichment

- CoreNLP (Manning, ACL '14)

Entity Salience

Three-Stage System





1. Document Enrichment

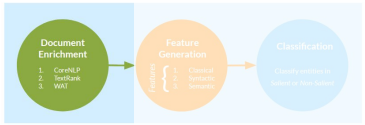
- CoreNLP (Manning, ACL '14)

Module

Sentence Splitting

Entity Salience

Three-Stage System



1. Document Enrichment

- ▷ CoreNLP (Manning, ACL '14)

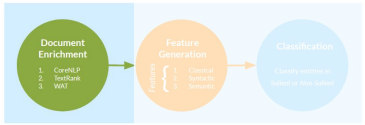
Module

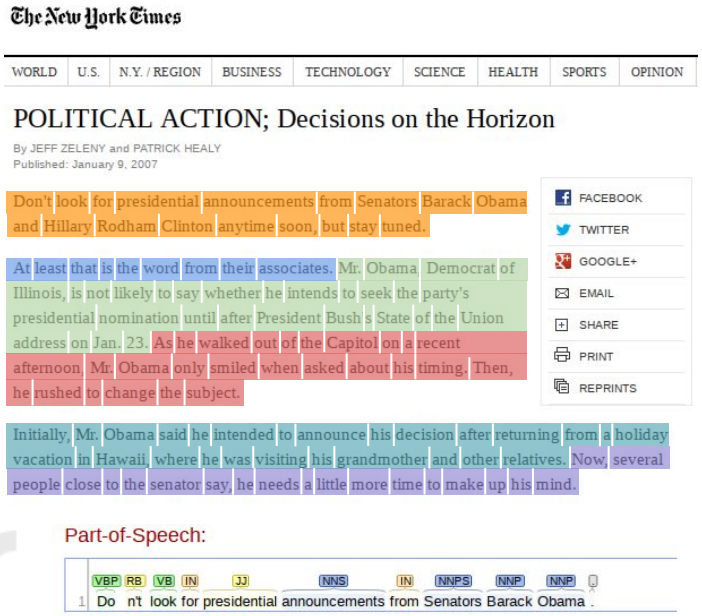
Sentence Splitting

Tokenization

Entity Salience

Three-Stage System





Images via
http://corenlp.run

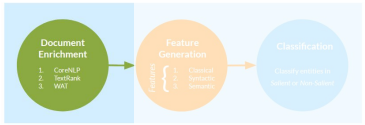
1. Document Enrichment

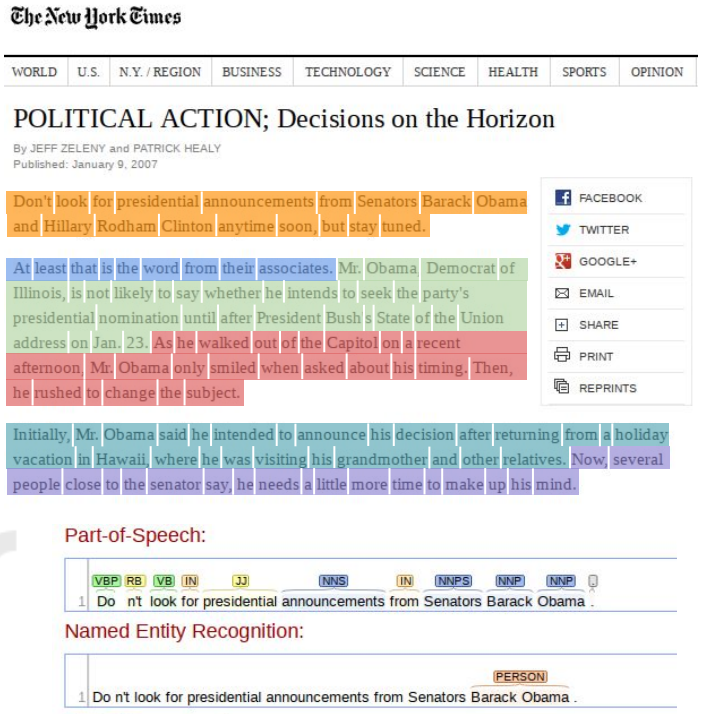
- **CoreNLP** (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging

Entity Salience

Three-Stage System





Images via
http://corenlp.run

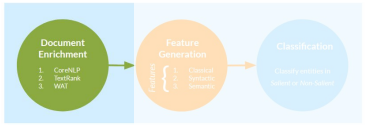
1. Document Enrichment

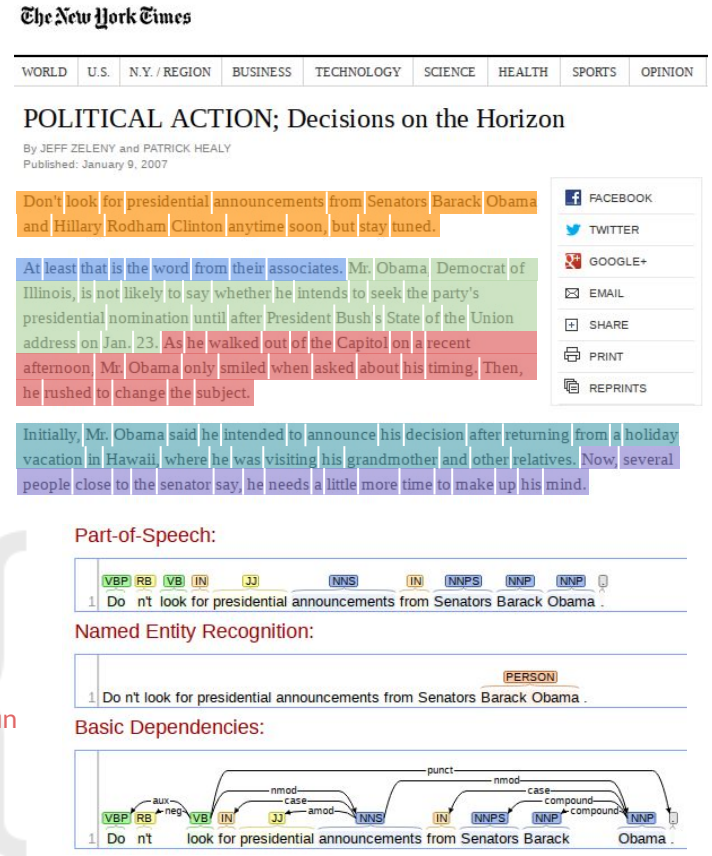
- ▷ CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition

Entity Salience

Three-Stage System





Images via http://corenlp.run

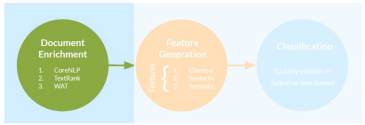
1. Document Enrichment

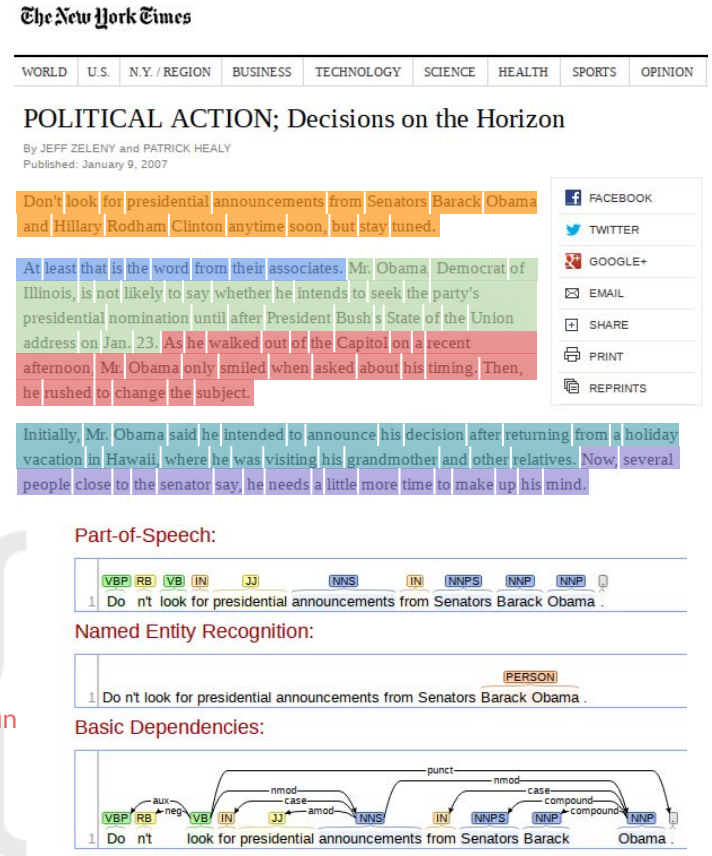
- ▷ CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition
Dependency Parsing

Entity Salience

Three-Stage System





Images via http://corenlp.run

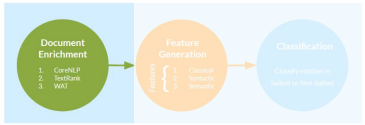
1. Document Enrichment

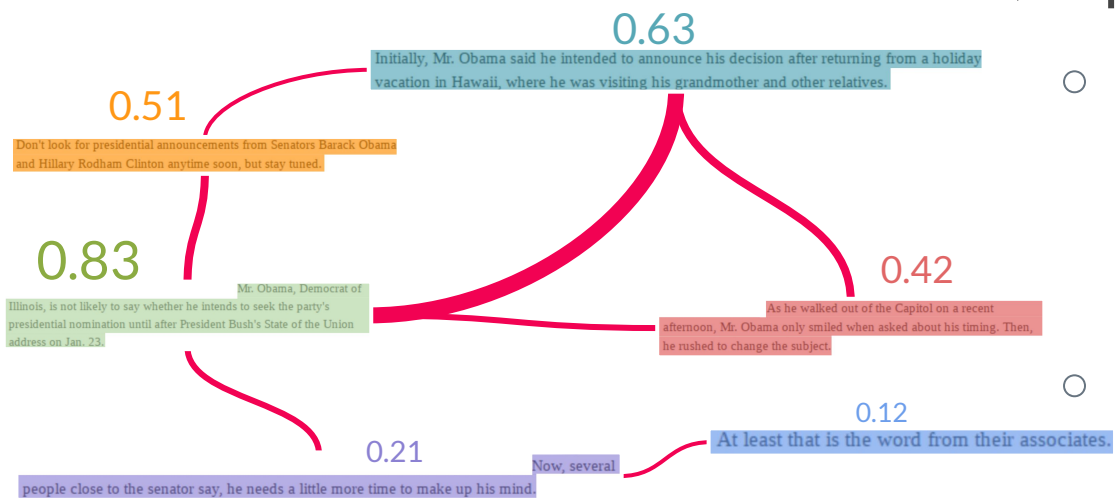
- ▷ CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition
Dependency Parsing
Coreference

Entity Salience

Three-Stage System





1. Document Enrichment

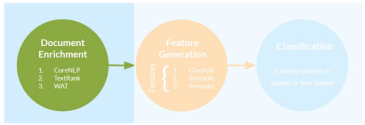
- ▷ **CoreNLP** (Manning, ACL '14)

Sentences

- ▷ **TextRank** (Mihalcea, EMNLP '04)
  - ○ Graph-Based Summarizer
    - ■ Nodes = Sentences
    - ■ Weights = Normalized Token Overlap
  - ○ Sentence Ranking via PageRank

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

1. Document Enrichment

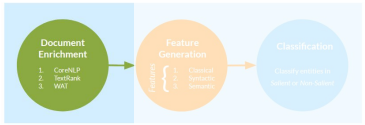
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ **WAT** (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



1. Document Enrichment

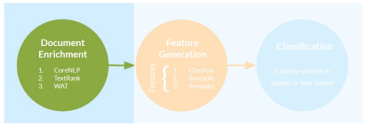
- **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- **WAT** (Piccinno, SIGIR '14)
  - Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS


Barack_Obama


Hilary_Clinton

1. Document Enrichment

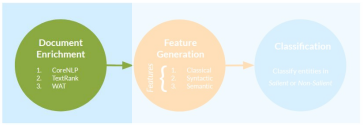
- ▷ CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ WAT (Piccinno, SIGIR '14)
  - ○ Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

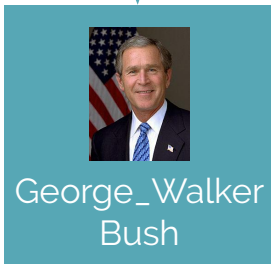
At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS


Barack_Obama


Hilary_Clinton



1. Document Enrichment

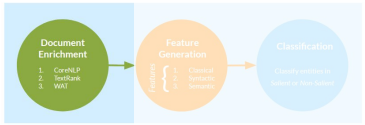
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ **WAT** (Piccinno, SIGIR '14)
  - ○ Annotates them with **Wikipedia Entities**

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

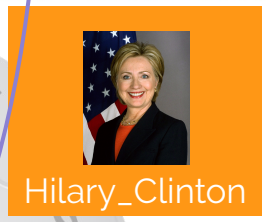
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS


Barack_Obama


Hilary_Clinton


United_States Capitol


George_Walker Bush

1. Document Enrichment

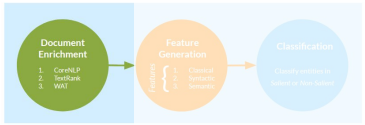
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- ▷ **WAT** (Piccinno, SIGIR '14)
  - ○ Annotates them with **Wikipedia Entities**

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.


Barack_Obama


Hilary_Clinton


United_States Capitol


Hawaii


George_Walker Bush

1. Document Enrichment

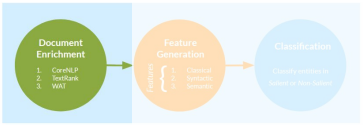
- **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

- **WAT** (Piccinno, SIGIR '14)
  - Annotates them with Wikipedia Entities

Entity Salience

Three-Stage System



The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

POLITICAL ACTION; Decisions on the Horizon

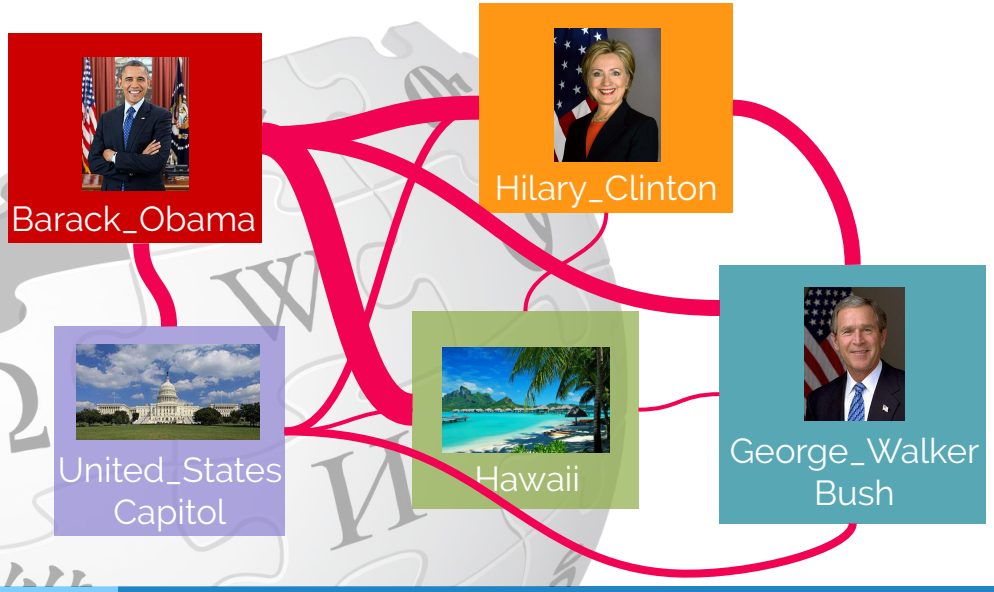
By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.



1. Document Enrichment

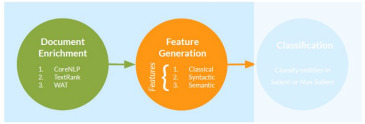
- **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

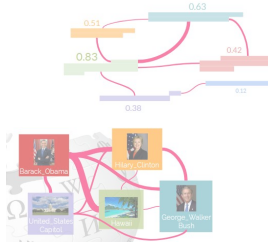
- **WAT** (Piccinno, SIGIR '14)
  - Annotates them with Wikipedia Entities
  - Releatedness Graph
    - Nodes = Entities
    - Weights = Wikipedia Jaccard In-Links

Entity Salience

Three-Stage System



1. Document Enrichment



Tokens, POS Tags, Dependency Relations, Coreference Chains, Sentence Ranks, Wikipedia Entities and their Relatedness

2. Feature Generation

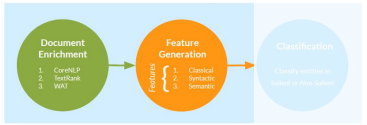
Standard Entity Features	
$ef(e,d), idf(e), ef\text{-}idf(e,d)$	Entity frequency (number of times WAT annotates e in d), inverse document frequency for e and their product.
$position\text{-}stats_{\{s,t\}}(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of sentence- (resp. token-) positions of e in d.
$mention\text{-}title(e,d)$	Presence of a *mention* of e in the title of d.
$entity\text{-}title(e,d)$	Presence of e in the title of d.
$is\text{-}upper(e,d)$	True if one of the mentions of e appear in d in uppercase, false otherwise.
Cmu-Google's Features	
$1st\text{-}loc(e,d)$	Index of the sentence in which the first mention of e appears in d, modeled as a binary indicator.
$head\text{-}count(e,d)$	Frequency of head word of entity e in the document d, modeled as a binary indicator.
$mentions(e,d)$	Sum between entity frequency and co-referenced frequency of e in d, modeled as a binary indicator.
$headline(e,d)$	POS tag of each word of e that appears in at least one mention and also in the headline of d, normalized via feature hashing.
$head\text{-}lex(e,d)$	Lower-cased head word of the first mention of e in d, normalized via feature hashing.
$google\text{-}centrality(e,d)$	PageRank score of e on the entity graph generated from d, where weights are the co-occurrence probability of two entities, computed on the training set.
Novel Features Introduced	
$spread_{\{s,t\}}(e,d)$	Difference between the maximum and minimum sentence- (resp. token-) positions of e in d.
$bucketed\text{-}freq_{\{s,t\}}(e,d)$	Vector of bucketed frequencies through sentence- (resp. token-) positions of e in d.
$textrank\text{-}stats(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of TEXTRANK scores of sentences where e appears in d.
$dep\text{-}freq(e,d)$	Frequency of e in d when it appears as dependent of the dependency relation dep.
$dep\text{-}bucketed\text{-}freq_{\{s,t\}}(e,d)$	Vector of bucketed frequencies through sentence- (resp. token-) positions of e in d where only the mentions where e appears as dependent of a dependency relation dep are considered.
$dep\text{-}position\text{-}stats_{\{s,t\}}(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of sentence- (resp. token-) positions of e in d, where only the mentions where e appears as dependent of a dependency relation dep are considered.
$dep\text{-}textrank\text{-}stats(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of TEXTRANK scores where only the sentences where e appears as dependent of a dependency relation dep are taken into account.
$rel\text{-}stats(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the relatedness scores between e and all other entities annotated in d.
$rel\text{-}bucketed\text{-}stats_{\{s,t\}}(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the relatedness scores between e and all other entities present in d, bucketed over document positions (both at sentence- and token-level).
$comm\text{-}stats(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the *commonness* values of e in d (see WAT).
$\rho\text{-}stats(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the ρ-score values of e in d (see WAT).
$rel\text{-}centrality(e,d)$	Degree, PageRank, Betweenness, Katz, HITS, Closeness and Harmonic [2] scores of e computed on the entity graph of d. Details in text.

2. Feature Generation

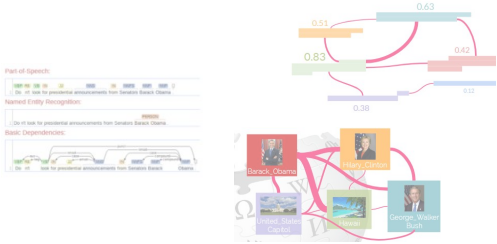
- Standard Entity Features
  - Frequency
  - Positions
  - …
- CMU-Google Features
  - POS-Tags, Coreference Freq.
  - PageRank on a graph whose weights are based on co-occ.
  - …
- Syntactic Features
  - Statistics on Sentence Ranks
  - Frequency/Positions of Dependency Relations
  - …
- Semantic Features
  - Statistics on annotations (coherence, commonness)
  - Graph Centralities on Relatedness Graph
  - Relatedness over Positions
  - …

Entity Salience

Three-Stage System

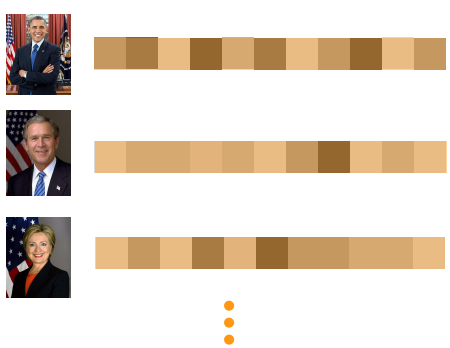


1. Document Enrichment



Tokens, POS Tags, Dependency Relations, Coreference Chains, Sentence Ranks, Wikipedia Entities and their Relatedness

2. Feature Generation



Entity Feature Vectors

3. Salience Classification

dmlc XGBoost

(Chen, SIGKDD '16)

Salient Entities

Experiments

Results

System	New York Times			Wikinews		
	Micro			Macro		
	P	R	F1	P	R	F1
CMU-Google (Dunietz, EACL '14)	60.5	**63.5**	61.5	n.a.	n.a.	n.a.
CMU-Google-ours	58.8	62.6	60.7	35.7	48.6	38.0
SEL (Trani, DocEng '16)	n.a.	n.a.	n.a.	41.0	51.0	43.0
SWAT	**62.2**	63.0	**62.6**	**47.3**	**55.2**	**47.8**

Experiments
Results

System	New York Times Micro			Wikinews Macro		
	P	R	F1	P	R	F1
CMU-Google (Dunietz, EACL '14)	60.5	**63.5**	61.5	n.a.	n.a.	n.a.
CMU-Google-ours	58.8	62.6	60.7	35.7	48.6	38.0
SEL (Trani, DocEng '16)	n.a.	n.a.	n.a.	41.0	51.0	43.0
SWAT	**62.2**	63.0	**62.6**	**47.3**	**55.2**	**47.8**

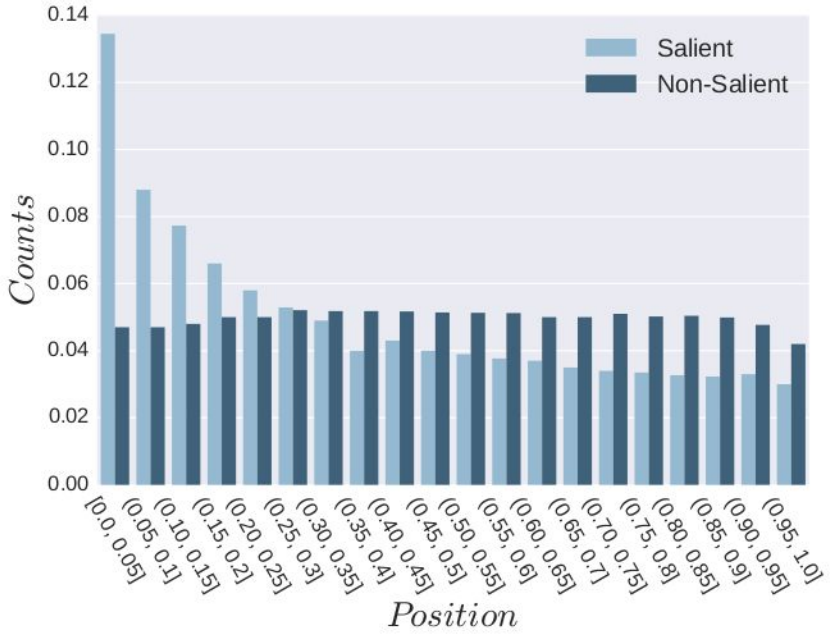
- Up to +1.9% on New York Times and +9.8% on Wikinews

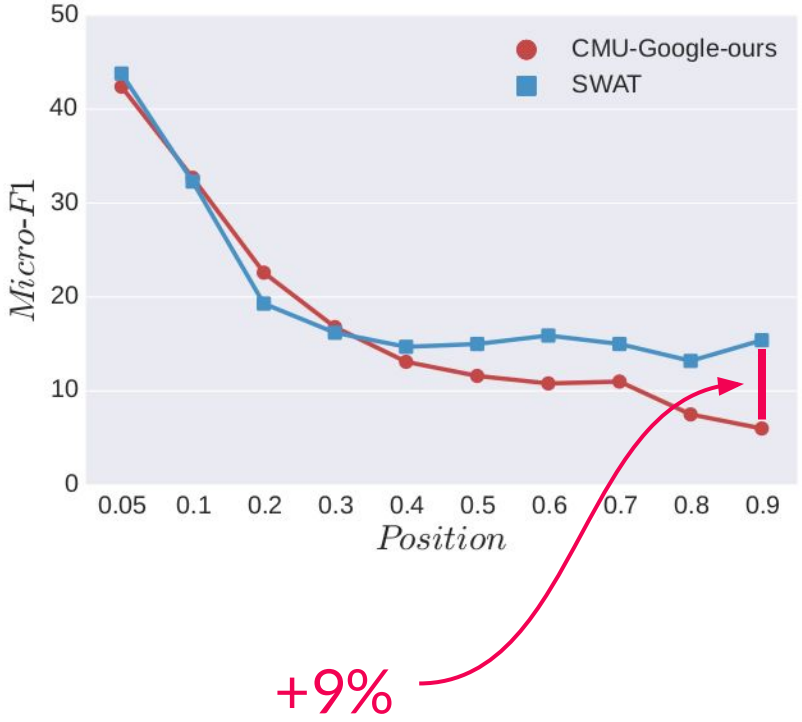
Experiments
Results

Dataset

New York Times

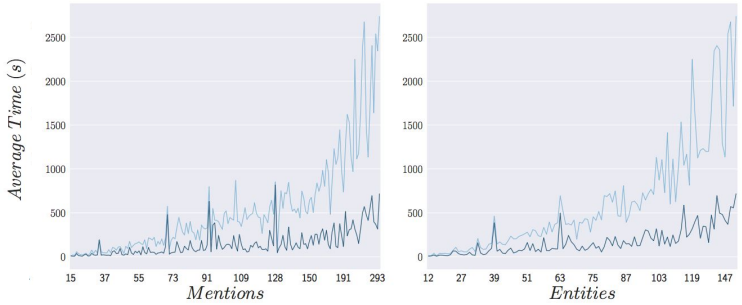
Independence from position of salient entities





Recent Improvements & Future Works

- We recently enhance our system by using several new features based on Entity Embeddings (Perozzi, KDD '14), (Ni, WSDM '16)
- Improvements up to
  - +14% (micro-F1) on New York Times (+5% with respect to 1.0)
  - +9.9% (macro-F1) on Wikinews (+0.1% with respect to SWAT 1.0)
- System now 5x faster



- Journal paper almost concluded
- Deploy SWAT on SoBigData infrastructure

2.
Entity Relatedness

Based on the full paper

A Two-Stage Framework for Computing Entity Relatedness in Wikipedia

Marco Ponza, Paolo Ferragina and Soumen Chakrabarti

Entity Relatedness

Motivation

Proliferation of the usage of Knowledge Graphs


Customers

- Retrieval of Information (Blanco, WSDM '15), (Cornolti, WWW '16)
- Entity Linking (Mihalcea, CIKM '07), (Meij, WSDM '12), (Ganea, WWW '16)
- Document Clustering , Classification and Similarity
  (Scaiella, WSDM '12), (Vitale, ECIR '12), (Ni, WSDM '16)

Need for computing *entity relatedness*

Compute how much two entities are related

Relatedness : Entities x Entities →Real

The Wikipedia Knowledge Graph

- Our Knowledge Graph (KG): WIKIPEDIA The Free Encyclopedia

The Wikipedia Knowledge Graph

- Our Knowledge Graph (KG): WIKIPEDIA The Free Encyclopedia
  - Entity?

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).
This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".[4] According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".[4] Marco Rosci notes that while there is much speculation regarding his life and personality, his view of the world was logical rather than mysterious, and that the empirical methods he employed were unorthodox for his time.[5]

Leonardo da Vinci

Portrait by Francesco Melzi.

- Entity = Wikipedia Page = Node of our KG

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).
This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as the prime exemplar of the "Universal Genius" or "Renaissance Man", an individual of "unquenchable curiosity" and "feverishly inventive imagination".[4] According to art historian Helen Gardner, the scope and depth of his interests were without precedent in recorded history, and "his mind and personality seem to us superhuman, while the man himself mysterious and remote".[4] Marco Rosci notes that while there is much speculation regarding his life and personality, his view of the world was logical rather than mysterious, and that the empirical methods he employed were unorthodox for his time.[5]

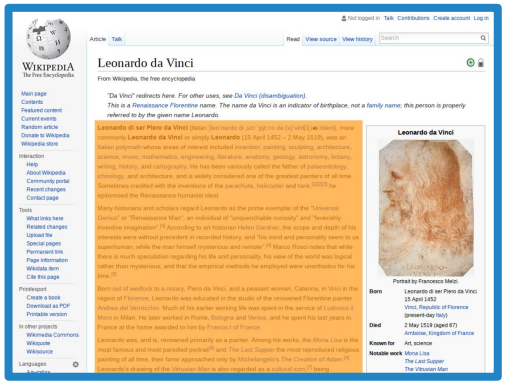
Leonardo da Vinci

Portrait by Francesco Melzi.

- Entity = Wikipedia Page = Node of our KG
- Label of an Entity = Textual Description of a Wikipedia Page

Terminology

- Our Knowledge Graph (KG): WIKIPEDIA The Free Encyclopedia
  - Entity = Wikipedia Page (a node of KG)
  - Label = Textual Description of the Wikipedia Page
  - Edges?



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

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Wikimedia Commons
Wikiquote
Wikisource

Languages

Leonardo da Vinci

From Wikipedia, the free encyclopedia

"Da Vinci" redirects here. For other uses, see Da Vinci (disambiguation).
This is a Renaissance Florentine name. The name da Vinci is an indicator of birthplace, not a family name; this person is properly referred to by the given name Leonardo.

Leonardo di ser Piero da Vinci (Italian: [leoˈnardo di ˌsɛr ˈpjɛːro da (v)ˈvintʃi] (listen)), more commonly **Leonardo da Vinci** or simply **Leonardo** (15 April 1452 – 2 May 1519), was an Italian polymath whose areas of interest included invention, painting, sculpting, architecture, science, music, mathematics, engineering, literature, anatomy, geology, astronomy, botany, writing, history, and cartography. He has been variously called the father of palaeontology, ichnology, and architecture, and is widely considered one of the greatest painters of all time. Sometimes credited with the inventions of the parachute, helicopter and tank,[1][2][3] he epitomised the Renaissance humanist ideal.

Many historians and scholars [illegible] "Genius" or "Renaissance Man [illegible] of his

[illegible] ed in Rome, Bologna and Venice, and he spent his last [illegible] d to him by Francis I of France.

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona* most famous and most parodied portrait[6] and *The Last Supper* the most reproduce painting of all time, their fame approached only by Michelangelo's *The Creation of A* Leonardo's drawing of the *Vitruvian Man* is also regarded as a cultural icon,[7] being

Leonardo da Vinci

The Vitruvian Man

Science

From Wikipedia, the free encyclopedia

This article is about the general term. For other uses, see Science (disambiguation).

Science[nb 1][2]:58[3] is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[nb 2]

Contemporary science is typically subdivided into the natural sciences, which study the material universe; the social sciences, which study people and societies; and the formal sciences, such as mathematics. The formal sciences are often excluded as they do not depend on empirical observations.[4] Disciplines which use science like engineering and medicine may also be considered to be applied sciences.[5]

During the Middle Ages in the Middle East, foundations for the scientific method were laid by Ibn al-Haytham in his *Book of Optics*.[6][7][8][9][10] From classical antiquity through the 19th century, science as a type of knowledge was more closely linked philosophy than it is now and, in fact, in the Western world, the term "natural philosophy" encompassed fields of study that today associated with science, such as astronomy, medicine, and physics.[11][nb 3] While the classification of the material wo the ancient Indians and Greeks into air, earth, fire and water was more philosophical, medieval Middle Eastern scientists us practical, experimental observation to classify materials.[12]

In the 17th and 18th centuries, scientists increasingly sought to formulate knowledge in terms of *laws of nature*. Over the co of the 19th century, the word "science" became increasingly associated with the scientific method itself, as a disciplined way to study the natural world. It was in the 19th century that scientific disciplines such as biology, chemistry, and physics reached their modern shapes. The same time period also included the origin of the terms "scientist" and "scientific community," the founding of scientific institutions, and increasing significance of the interactions with society and other aspects of culture.[13][14]

Invention

From Wikipedia, the free encyclopedia

"Inventor" and "Invented" redirect here. For other uses, see Invention (disambiguation).
For more details on inventions throughout history, see Timeline of historic inventions.
For the CAD design software, see Autodesk Inventor.

An **invention** is a unique or novel device, method, composition or process. The invention p engineering and product development process. It may be an improvement upon a machine an object or a result. An invention that achieves a completely unique function or result may novel and not obvious to others skilled in the same field. An inventor may be taking a big st

Some inventions can be patented. A patent legally protects the intellectual property rights c claimed invention is actually an invention. The rules and requirements for patenting an inve process of obtaining a patent is often expensive.

Another meaning of invention is **cultural invention**, which is an innovative set of useful so passed on to others.[1] The Institute for Social Inventions collected many such ideas in mag important component of artistic and design creativity. Inventions often extend the boundari capability.

Astronomy

From Wikipedia, the free encyclopedia

This article is about the scientific study of celestial objects. For other uses, see Astronomy

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets nebulae) and processes (such as supernovae explosions, gamma ray bursts, and cosmic micr physics, chemistry, and evolution of such objects and processes, and more generally all phenc atmosphere of Earth. A related but distinct subject, physical cosmology, is concerned with stud

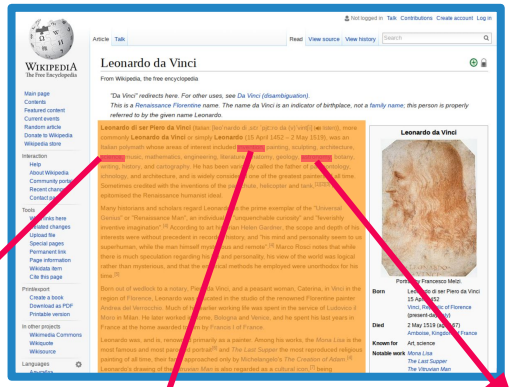
Astronomy is the oldest of the natural sciences. The early civilizations in recorded history, such Egyptians, Nubians, Iranians, Chinese, and Maya performed methodical observations of the ni included disciplines as diverse as astrometry, celestial navigation, observational astronomy an professional astronomy is nowadays often considered to be synonymous with astrophysics.[2]

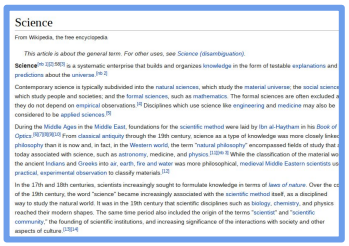
During the 20th century, the field of professional astronomy split into observational and theoret is focused on acquiring data from observations of astronomical objects, which is then analyzed Theoretical astronomy is oriented toward the development of computer or analytical models to phenomena. The two fields complement each other, with theoretical astronomy seeking to exp observations being used to confirm theoretical results.

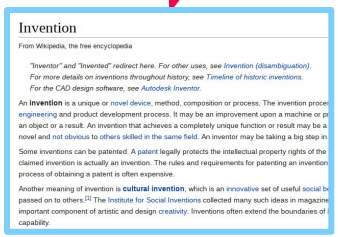
Astronomy is one of the few sciences where amateurs can still play an active role, especially in transient phenomena. Amateur astronomers have made and contributed to many important as new comets.

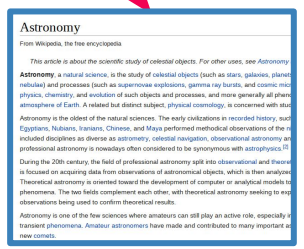
The Wikipedia Knowledge Graph

- Our Knowledge Graph (KG): WIKIPEDIA The Free Encyclopedia
  - Entity = Wikipedia Page (a node of KG)
  - Label = Textual Description of the Wikipedia Page
  - Edge = Wikipedia Hyperlinks









Known Relatedness Methods

A large number of methods proposed in literature...

- Personalized Web Search (Haveliwala, WWW '02)
- Link Prediction (Liben-Nowell, JAIST '07)
- Word and Document Similarity (Gabrilovich, IJCAI '07)
- Document Annotation (Piccinno, SIGIR '14)
- Machine Translation (Rothe, ACL '14)
- Document Classification (Perozzi, KDD '14), (Tan, WWW '15)

...that have been applied or are similar to our problem

We have experimented them
on the Entity Relatedness task

Our Two-Stage Framework

- Built on the top of existing relatedness algorithms
- Improves current approaches
  - More accurate relatedness scores
  - Fast at query time

- The two stages of our framework:

1. A small and weighted subgraph is dynamically grown around the two *query entities*

2. Computing the relatedness between the two *query entities* according with the generated subgraph

- Motivations
  - Wikipedia edges are noisy (introduced for citation, explanation, ...)
  - Subgraph nodes are strongly related to the query entities (they are good bridges)
  - Subgraph edges are less noisy (confined to few meaningful bridge nodes)

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*

Tiger

Cat

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*

Tiger

Cat

How can we populate the subgraph?

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*


Tiger


Siberian_tiger


Leopard


Jaguar


European_cat


Cat_anatomy


Felidae

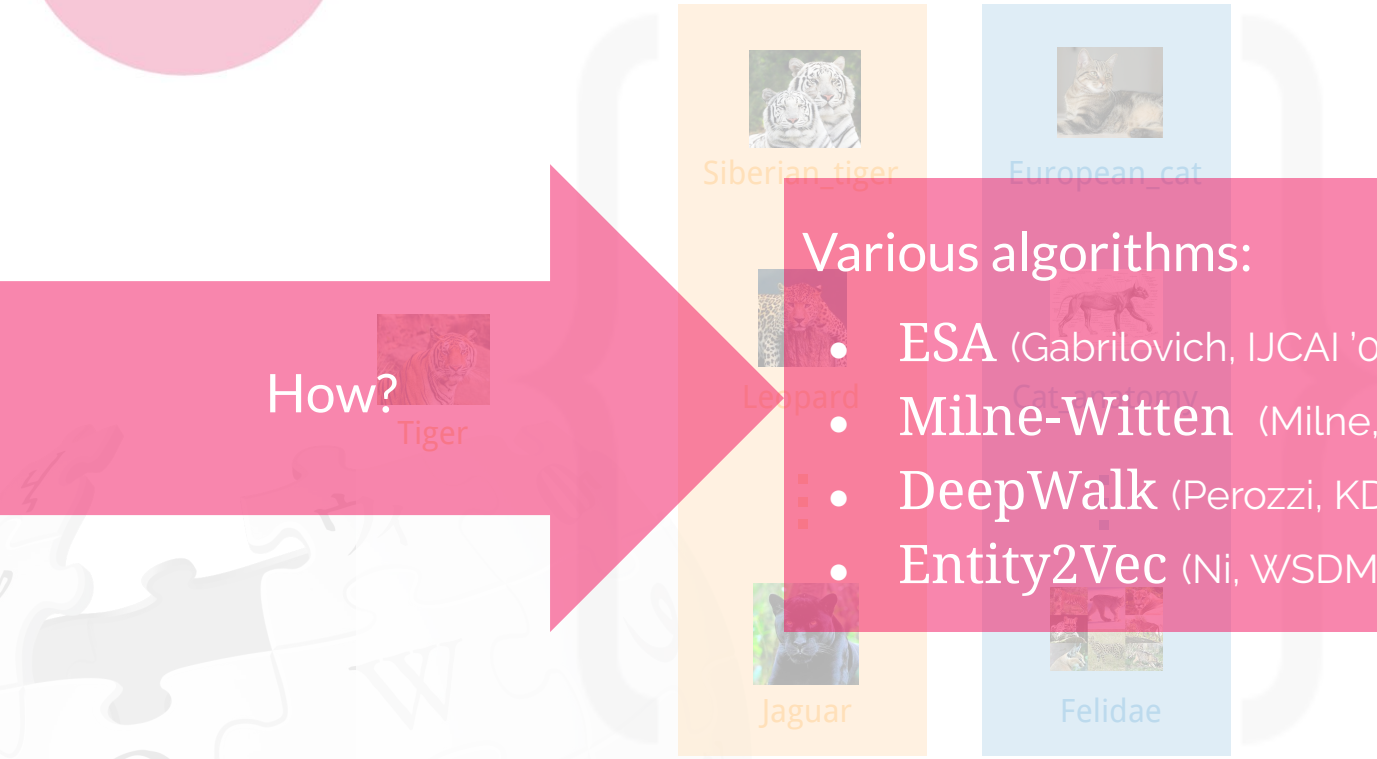

Cat

Populating the subgraph. Choosing the *top-k* nodes *most related* to the query entities

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*



How?

Various algorithms:

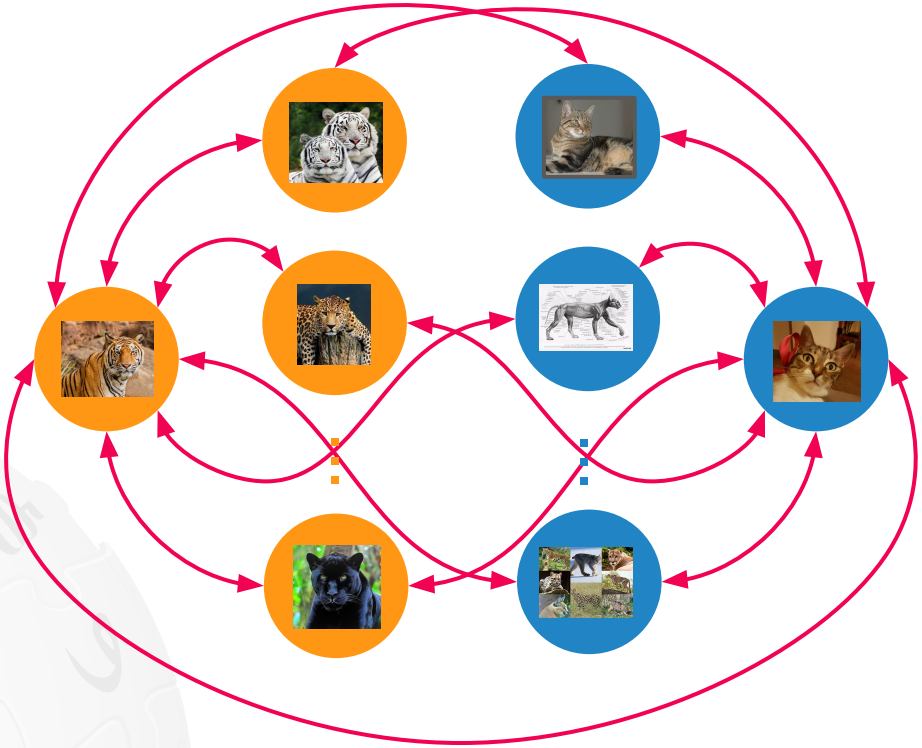
- ESA (Gabrilovich, IJCAI '07)
- Milne-Witten (Milne, AAAI '08)
- DeepWalk (Perozzi, KDD '14)
- Entity2Vec (Ni, WSDM '16)

Populating the subgraph. Choosing the *top-k* nodes *most related* to the query entities

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*



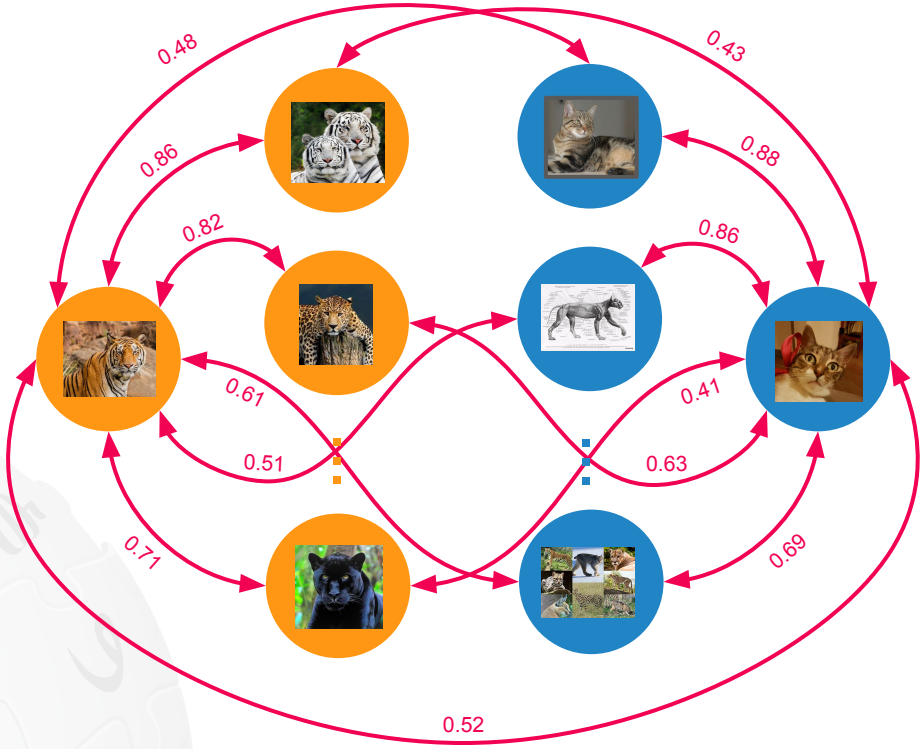
Creating the edges. Each query entity is linked to

- *the other query entity*
- *its top-k related entities*
- *the other top-k related entities*

Our Two-Stage Framework

1

A small and weighted subgraph is dynamically grown around the two *query entities*



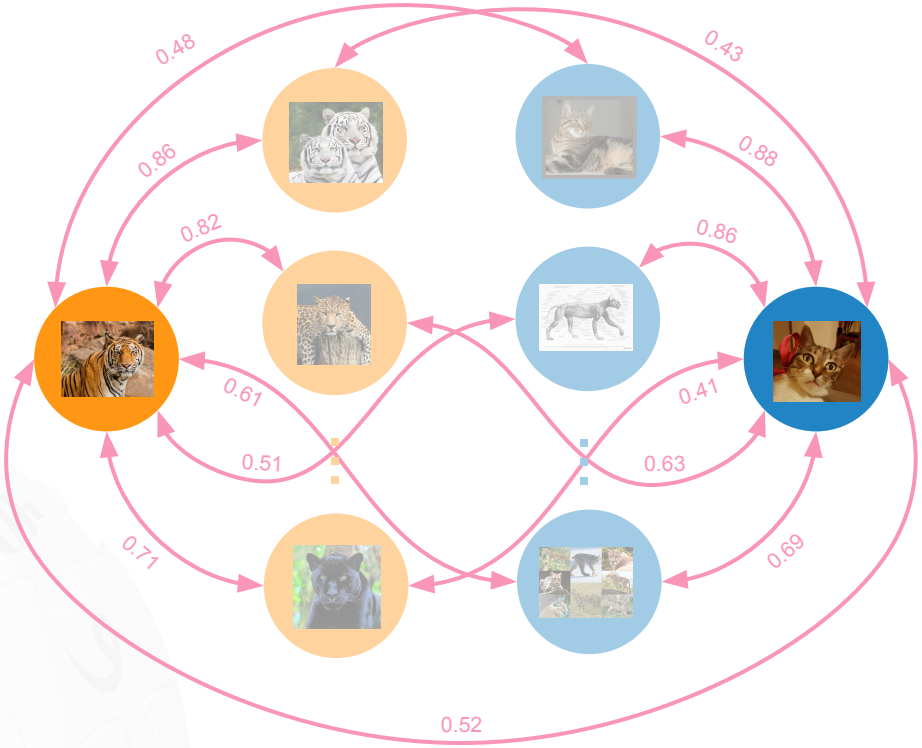
Weighting the edges. How?

- Milne-Witten (Milne, AAAI '08)
- DeepWalk (Perozzi, KDD '14)
- Entity2Vec (Ni, WSDM '16)

Our Two-Stage Framework

2

Computing the relatedness between the two *query entities* according with the generated subgraph



Computing Relatedness

CoSimRank (Rothe, ACL '14)

relatedenss ( , ) = *0.65*

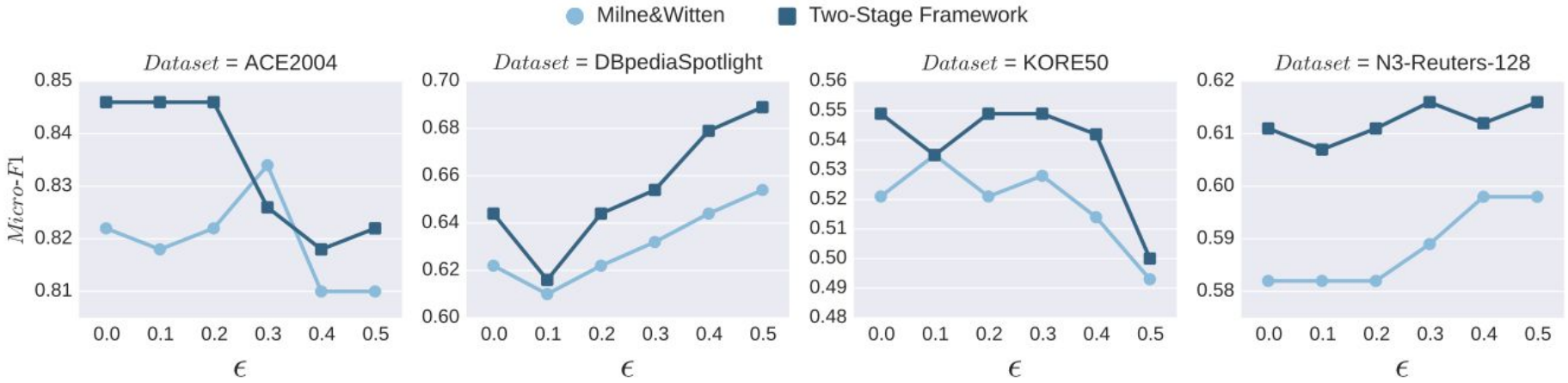
Experiments
Intrinsic Evaluation

Method	WikiSim			WiRe			AVG
	Pearson	Spearman	Harmonic	Pearson	Spearman	Harmonic	
ESA	0.61	0.72	0.67	0.60	0.63	0.62	0.645
Milne-Witten	0.62	0.65	0.63	0.77	0.69	0.72	0.675
DeepWalk	0.71	0.70	0.71	0.74	0.68	0.71	0.710
Entity2Vec	0.68	0.70	0.69	0.74	0.70	0.72	0.705
Two-Stage Framework	**0.74**	**0.75**	**0.74**	**0.83**	**0.75**	**0.79**	**0.765**

- Pearson measures predicted-vs-correct scores
- Spearman focuses on the ranking order among entity pairs
- Two-Stage Framework instantiated with
  - Milne-Witten as Top-k Retrieval
  - Weights are the average between Milne-Witten and DeepWalk
- More experiments in the paper (first known comparison among *more than 15 methods!*)

Experiments
Extrinsic Evaluation

- Domain of *Entity Linking*
  - Linking short but meaningful sequence of words with proper Wikipedia Entities
- Entity Linker used for experiments: TAG·ME
  - We replaced the relatedness method used in TagMe (e.g. Milne-Witten) with our Two- Stage Framework



- Our relatedness measure not only improves TagMe, but also makes it more insensitive to choices of the ε-parameter in TagMe

Experiments
Optimizations & Efficiency

- Preprocessing of Milne-Witten on the out-neighbors for each entity
- Compressing the Wikipedia Graph with Webgraph (Boldi, WWW '04) using the Elias-Fano codes
- Compressing DeepWalk embeddings with FEL (Blanco, WSDM '15)

	Uncompressed	Compressed
Space	5 GB	445 MB
Average Time	0.5 ms	3 ms

- Our framework fits in few hundred of MB and the computation of the relatedness is still fast at query time!

Future Work

Several *open issues* are there.

- Extending our framework to other KGs:
  - YAGO
  - WikiData
  - ...

- How can we further speedup our framework?
  - LSH (Gionis, VLDB '99)
  - Sketches (Akiba, KDD '16)
  - ...

- Impact of our framework to other domains?
  - Query understanding (Cornolti, WWW '16)
  - Document similarity (Ni, WSDM '16)
  - ...

3.
What next?

What next?
Last year of PhD

- Novel application is Expert Finding (with P. Cifariello) where we are using the Relatedness results on several tasks:
  - Query expansion
  - Ranking the expertise of people
  - State-of-the-art is a paper in WWW 2016
- Journal version of the work on Entity Relatedness
- Currently visiting at Max-Planck Institute
  - just started working on Open Information Extraction
    - extract *highly compact* but still *readable facts* (subject, relation, object)
    - computing their *relevance*

...and last but not least...

...write the PhD Thesis!

Thanks!

Any questions?