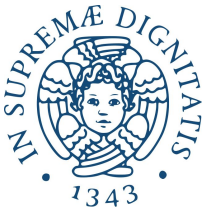
Document Aboutness via Sophisticated Syntactic and Semantic Features

Marco Ponza, Paolo Ferragina, and Francesco Piccinno

Dipartimento di Informatica
University of Pisa



NLdb 2017
Liège, Belgium

Menu

0. Introduction
    - The Document Aboutness Task
    - Through a Semantic Representation
1. Entity Salience
    - Our Solution
    - Three-Stage Framework
2. Experiments
    - Competitors & Datasets
    - Results
3. Conclusion & Future Work

Introduction

The Document Aboutness Task

Succinct Representation of a

Document's Subject Matter (Bruza, AIR '96)

- N-Grams (Turney, IR '00)
- Sentences (Mihalcea, EMNLP '04)
- Terms of Dictionary (Paranjpe, CIKM '08)
- Keywords (Liu, EMNLP '14)

Goal

Classical Aboutness Representation

Limitations

(Hasan, ACL '14)

Introduction

The Document Aboutness Task

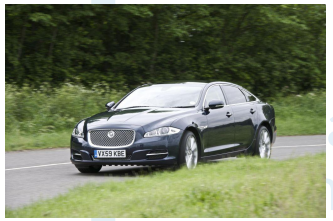
Input Document

Aboutness

Jaguar

Jaguar_(felin)

or



Jaguar_Cars

- Interpretation
- Overgeneration
- Infrequency
- Redundancy

Limitations

(Hasan, ACL '14)

Introduction

Through a Semantic Representation

- Interpretation
- Over-Generation
- Infrequency
- Redundancy

Limitations

(Hasan, ACL '14)

> *[...] errors could be addressed using background knowledge.*
>
> (Hasan, ACL '14)

Adding semantics into the aboutness representation

Introduction

Through a Semantic Representation

Semantic aboutness representation

1. Entity ∈ *Knowledge Base*

- ▷ Deployment of Entity Annotators

(Cucerzan, EMNLP '07) and many others

Maradona won against Mexico

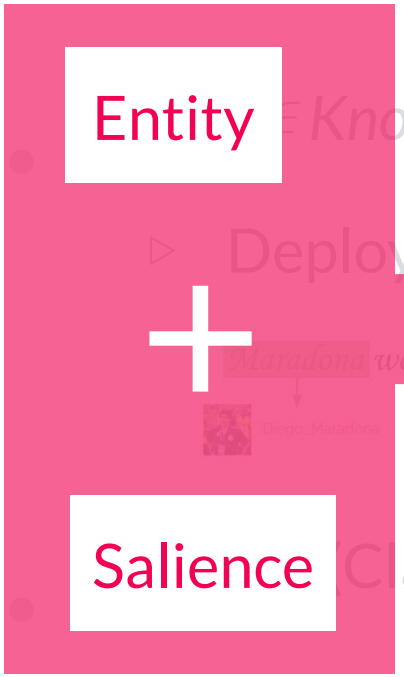
Diego_Maradona

Mexico_National_
Football_Team

Introduction

Through a Semantic Representation

Semantic Aboutness Representation



Entity Salience Extraction

(Dunietz, EACL '14), (Trani, DocEng '16)

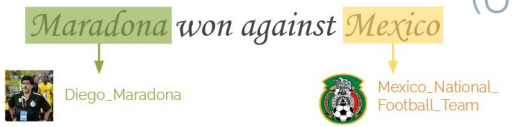
Introduction

Through a Semantic Representation

Semantic Aboutness Representation

1. Entity $\in$ *Knowledge Base*

- Deployment of Entity Annotators



(Usbeck, WWW '15)

(Dunietz, EACL '14), (Trani, DocEng '16)

2. Salience (Class or Score)

- Relevance of an Entity

Entity Salience

Our Solution

- Three-Stage Framework 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

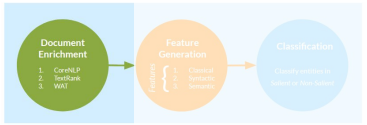
Classification

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

Salient Entities

Entity Salience

Three-Stage Framework



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

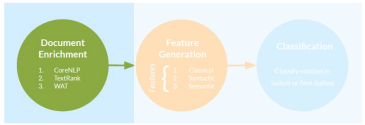
Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

1. Document Enrichment

- CoreNLP (Manning, ACL '14)

Entity Salience

Three-Stage Framework



1. Document Enrichment

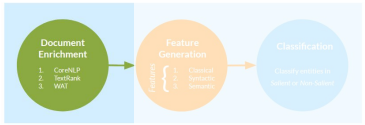
- ▷ CoreNLP (Manning, ACL '14)

Module

Sentence Splitting

Entity Salience

Three-Stage Framework



1. Document Enrichment

- ▷ CoreNLP (Manning, ACL '14)

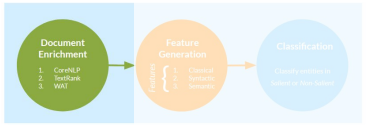
Module

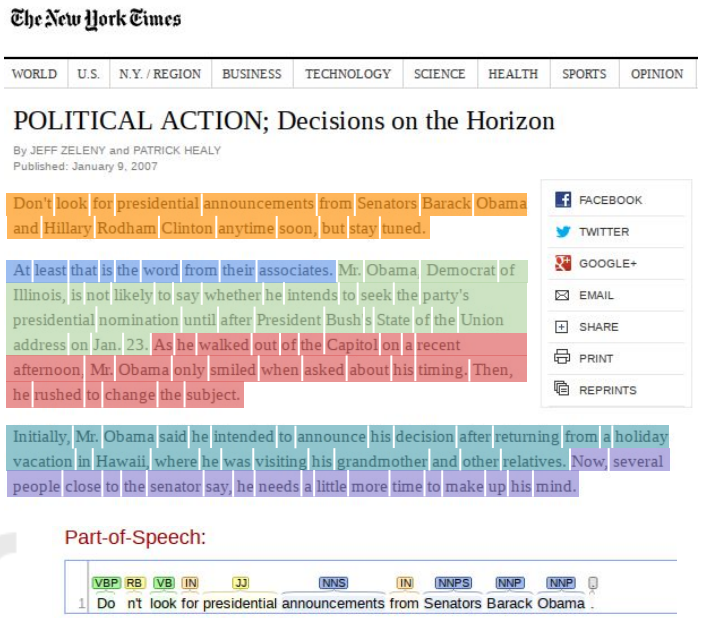
Sentence Splitting

Tokenization

Entity Salience

Three-Stage Framework





Images via
http://corenlp.run

1. Document Enrichment

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

Module

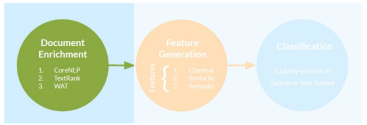
Sentence Splitting

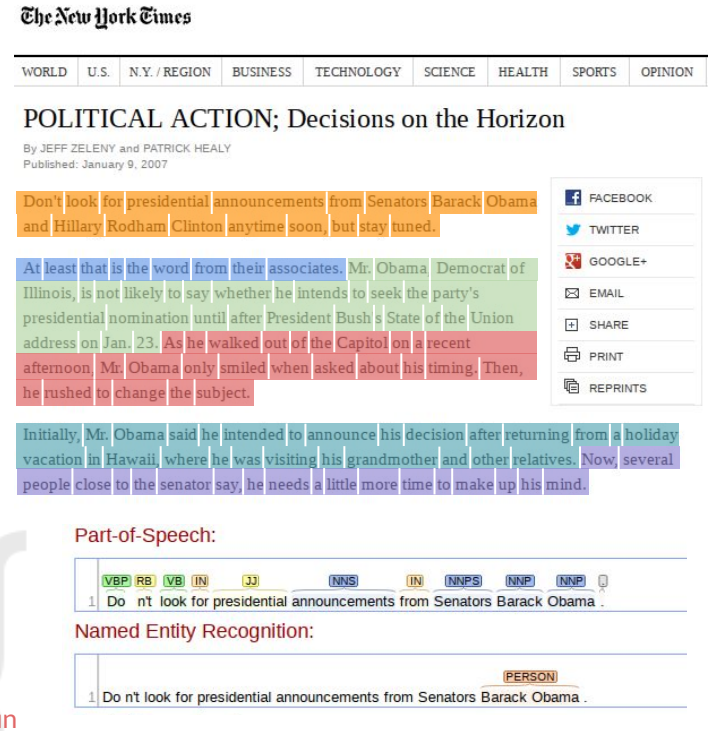
Tokenization

POS-Tagging

Entity Salience

Three-Stage Framework





Images via
http://corenlp.run

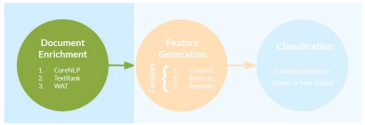
1. Document Enrichment

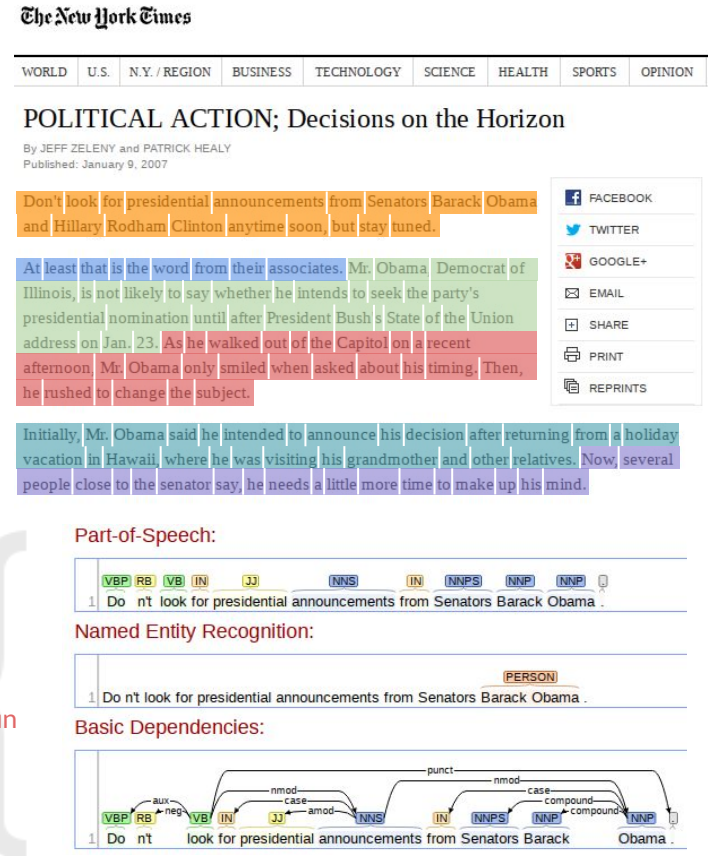
- ▷ CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition

Entity Salience

Three-Stage Framework





Images via http://corenlp.run

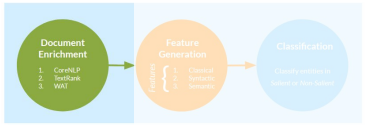
1. Document Enrichment

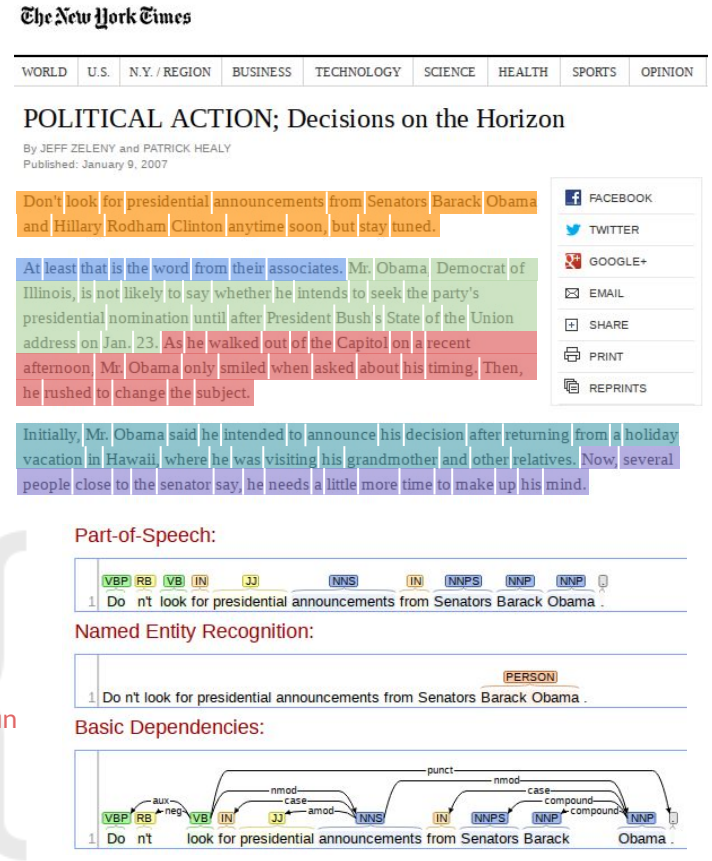
- ▷ CoreNLP (Manning, ACL '14)

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

Entity Salience

Three-Stage Framework





Images via http://corenlp.run

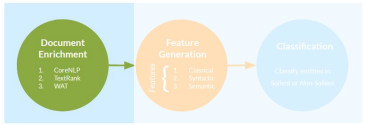
1. Document Enrichment

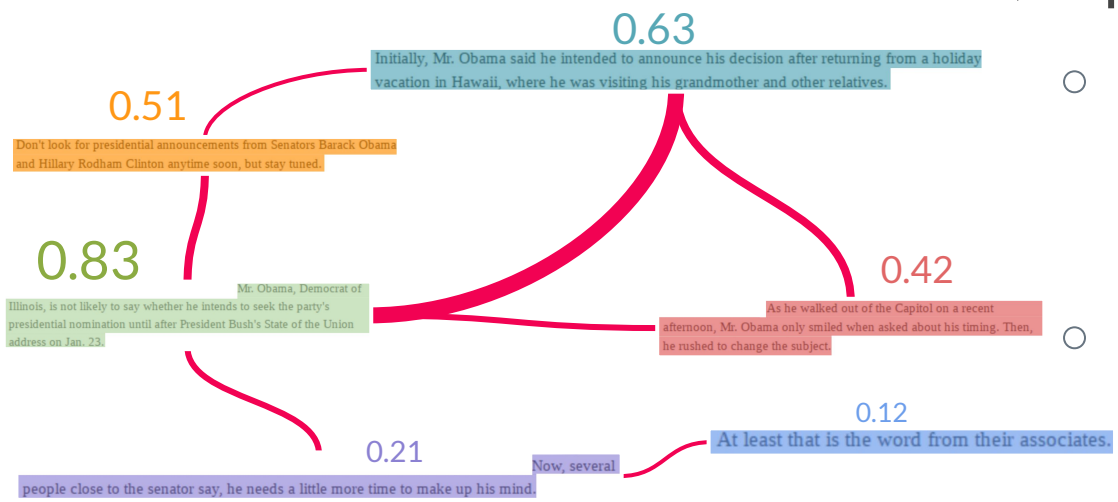
- ▷ CoreNLP (Manning, ACL '14)

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

Entity Salience

Three-Stage Framework





1. Document Enrichment

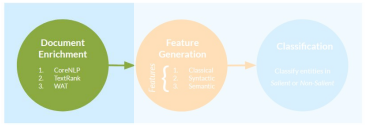
- ▷ CoreNLP (Manning, ACL '14)

Sentences

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

Entity Salience

Three-Stage Framework



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

1. Document Enrichment

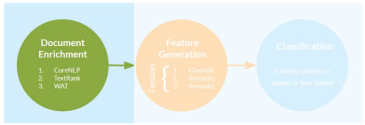
- ▷ CoreNLP (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



1. Document Enrichment

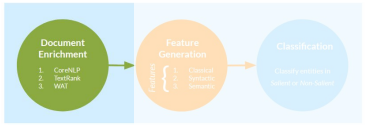
- **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



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

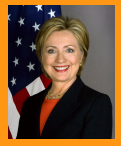
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

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

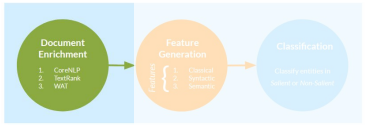
- **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



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.

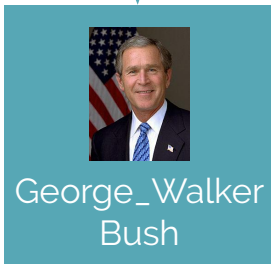
At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only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

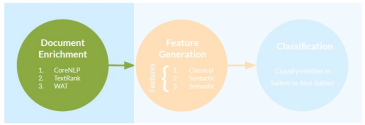
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



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

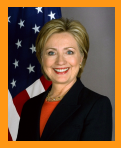
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

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

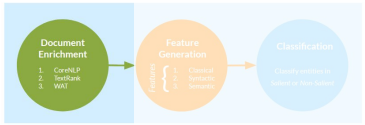
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



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

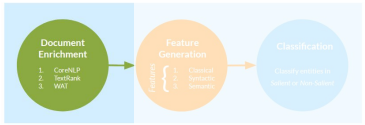
- ▷ **CoreNLP** (Manning, ACL '14)

Named Entities + Proper/Common Nouns

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

Entity Salience

Three-Stage Framework



The New York Times

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

POLITICAL ACTION; Decisions on the Horizon

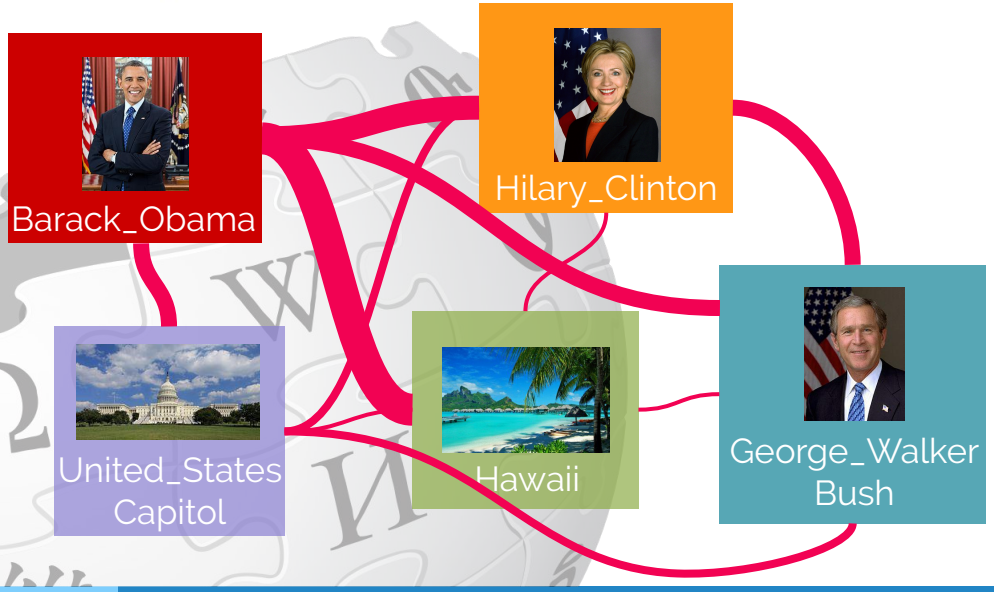
By JEFF ZELENY and PATRICK HEALY
Published: January 9, 2007

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)

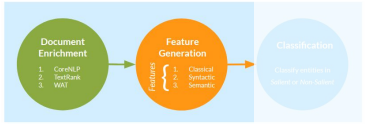
Named Entities + Proper/Common Nouns

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

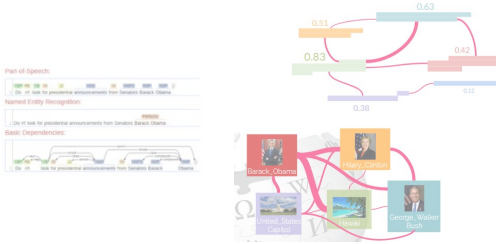

Barack Obama
From Wikipedia, the free encyclopedia
Barack Obama
44th President of the United States
In office
January 20, 2009 – January 20, 2017
Vice President Joe Biden
Preceded by George W. Bush
Succeeded by Donald Trump
United States Senator from Illinois
In office
January 3, 2005 – November 16, 2008
Preceded by Peter Fitzgerald
Succeeded by Roland Burris
Member of the Illinois Senate from the 13th district
President of the United States
Bill Clinton
Illinois
Hillary Clinton
From Wikipedia, the free encyclopedia
Hillary Clinton
67th United States Secretary of State
In office
January 21, 2009 – February 1, 2013
President Barack Obama
Deputy James Steinberg
William Joseph Burns
Preceded by Condoleezza Rice
Succeeded by John Kerry
United States Senator from New York
In office
January 3, 2001 – January 21, 2009
Preceded by Daniel Patrick Moynihan
Succeeded by Kirsten Gillibrand
First Lady of the United States
Entity Salience
Three-Stage Framework
Document Enrichment
Named Entities
CoreNLP
SWAT (Piccinno, SIGIR '14)
Annotates them with
Wikipedia Entities
Releatedness Graph
Nodes = Entities
Weights = Wikipedia
Jaccard In-Links
Barack_Obama
Hillary_Clinton
George_Walker_Bush

Entity Salience

Three-Stage Framework



1. Document Enrichment



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

2. Feature Generation

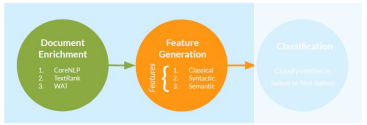
Standard Entity Features	
$ef(e,d), idf(e), ef\text{-}idf(e,d)$	Entity frequency (number of times WAT annotates e in d), inverse document frequency for e and their product.
$position\text{-}stats_{\{s,t\}}(e,d)$	Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of sentence- (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

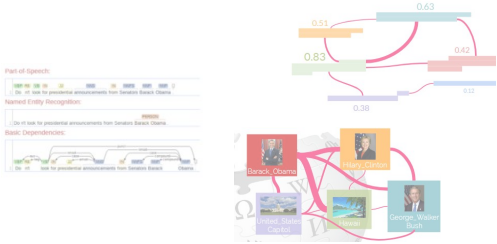
- Standard Entity Features
  - Frequency
  - Positions
  - …
- CMU-Google Features
  - POS-Tags, Coreference Freq.
  - PageRank on a graph whose weights are based on co-occ.
  - …
- Syntactic Features
  - Statistics on Sentence Ranks
  - Frequency/Positions of Dependency Relations
  - …
- Semantic Features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

Entity Salience

Three-Stage Framework

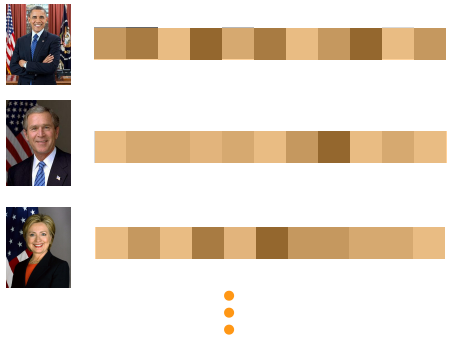


1. Document Enrichment



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

Experiments

Competitors & Datasets

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

  Limitations

  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

Experiments
Results

Dataset

New York Times

System	Micro		
	P	R	F1
CMU-Google (Dunietz, EACL '14)	60.5	**63.5**	61.5
CMU-Google-ours	58.8	62.6	60.7
SWAT	**62.2**	63.0	**62.6**

Experiments
Results

Dataset

New York Times

System	Micro		
	P	R	F1
CMU-Google (Dunietz, EACL '14)	60.5	**63.5**	61.5
CMU-Google-ours	58.8	62.6	60.7
SWAT	**62.2**	63.0	**62.6**

- Up to +1.9 % on Micro-F1

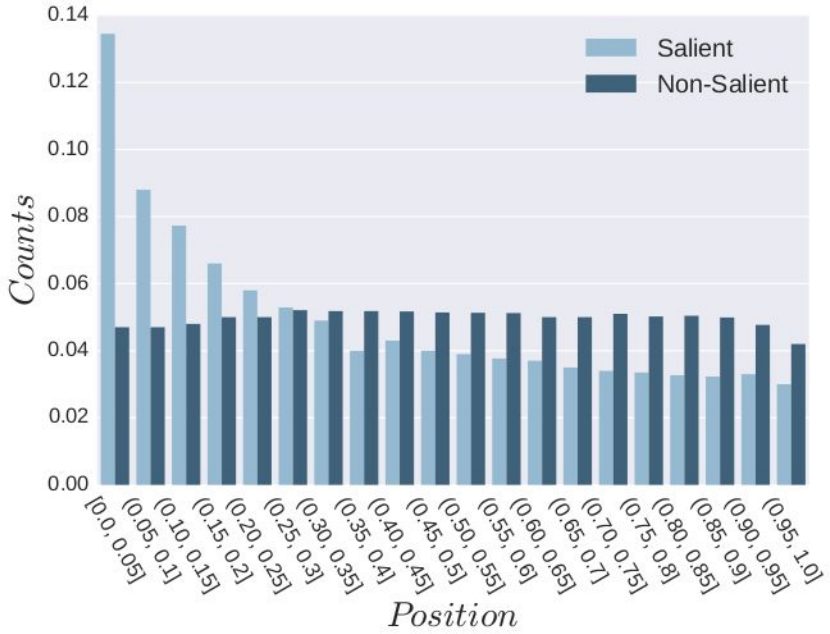
Experiments
Results

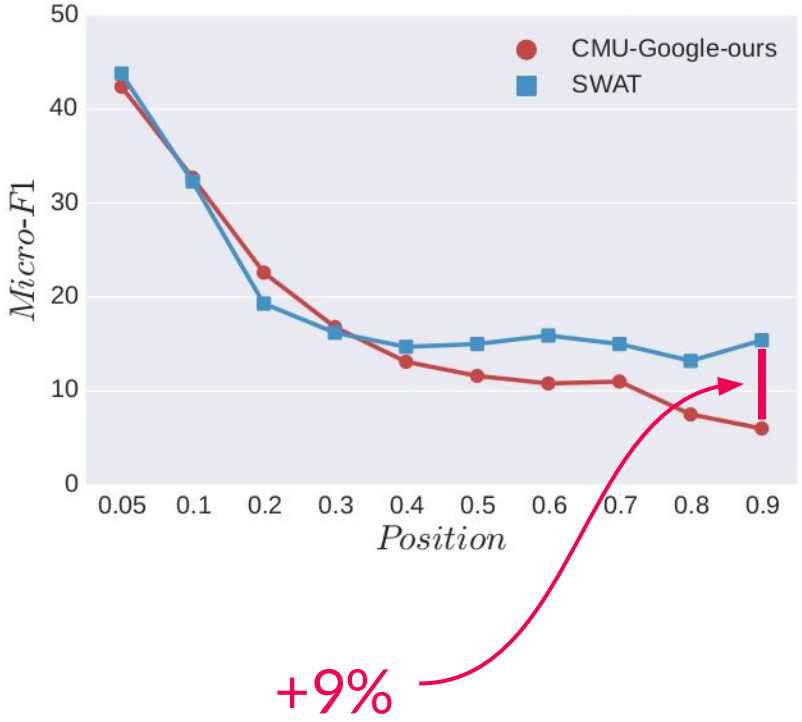
Dataset

New York Times

Independence from position of salient entities

Goal





Experiments
Results

Dataset

Wikinews

System	Macro		
	P	R	F1
CMU-Google-ours	35.7	48.6	38.0
SEL (Trani, DocEng '16)	42.0	51.0	43.0
SWAT	**47.3**	**55.2**	**47.8**

Experiments
Results

Dataset

Wikinews

System	Macro		
	P	R	F1
CMU-Google-ours	35.7	48.6	38.0
SEL (Trani, DocEng '16)	42.0	51.0	43.0
SWAT	**47.3**	**55.2**	**47.8**

- Up to +9.8% on Macro-F1

Conclusion & Future Work

Conclusion & Future Work

- We proved that the deployment of semantic knowledge eliminates the limitations described by (Hasan, ACL '14)

- Interpretation
- Over-Generation
- Infrequency
- Redundancy

Limitations
(Hasan, ACL '14)

- We aim to
  - Improve SWAT
    - Deployment of WAT 2.0 & Deep Learning
  - Ranking of Entities
  - Other datasets (not only news)
  - Deploy our system for other applications
    - Entity Annotation of Query
    - Document Similarity
    - …
    - *Please, suggest your favourite one!*

Thanks!

Any questions?