Algorithms and Applications for Web-Scale Knowledge Graphs

Marco Ponza

Supervisor

Prof. Paolo Ferragina

Menu

1. Entity Annotation
    - The Modeling of Knowledge
    - Terminology
    - The Annotation Pipeline
    - Applications
    - A New Text Representation

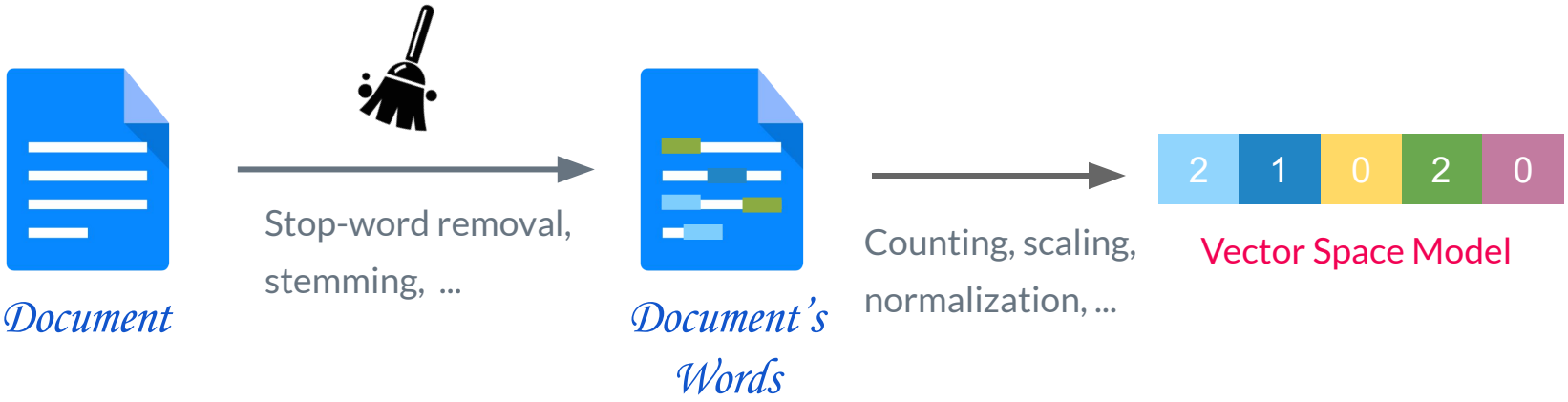
2. Work done in the first year
    - Entity Relatedness
    - Document Aboutness

3. Future Work

1.
Entity Annotation

The Modeling of Knowledge

- Classical approaches
  - Document Knowledge = Words
  - Bag-of-words (aka BoW)
  - Vector Space Model (aka VSM) (Salton, 1971)



The Modeling of Knowledge

- Well-known issues (Jurafsky, '00)
  - Ambiguity (Polysemy and Synonymy)

Jaguar

or

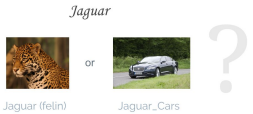
Jaguar (felin)

Jaguar_Cars

?

The Modeling of Knowledge

- Well-known issues (Jurafsky, '00)
  - Ambiguity (Polysemy and Synonymy)
  - Semantic Connections

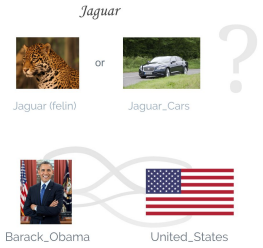


Barack_Obama

United_States

The Modeling of Knowledge

- Well-known issues (Jurafsky, '00)
  - Ambiguity (Polysemy and Synonymy)
  - Semantic Connections



- Algorithmic solutions
  - Latent Approaches (e.g. LSI/LSA, Word2Vec)
    - Unintelligible for humans (Gabrilovich IJCAI '07)
  - *"Knowledge is Power"* Hypothesis (Lenat, '91; Gabrilovich SIGIR '16)
    - Semantic and unambiguous concepts
    - Depend on the design of **Entity Annotators**

Entity Annotation

Terminology

- Wikipedia Knowledge Graph
- Node?

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
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

In other projects
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 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]

Born out of wedlock to a notary, Piero da Vinci, and a peasant woman, Caterina, in Vinci in the region of Florence, Leonardo was educated in the studio of the renowned Florentine painter Andrea del Verrocchio. Much of his earlier working life was spent in the service of Ludovico il Moro in Milan. He later worked in Rome, Bologna and Venice, and he spent his last years in France at the home awarded to him by Francis I of France.

Leonardo was, and is, renowned primarily as a painter. Among his works, the *Mona Lisa* is the most famous and most parodied portrait[6] and *The Last Supper* the most reproduced religious painting of all time, their fame approached only by Michelangelo's *The Creation of Adam*.[4]

Leonardo da Vinci

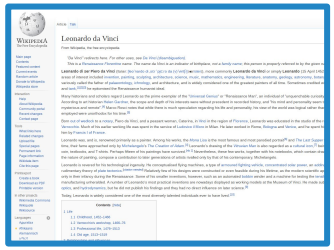
Portrait by Francesco Melzi.

Born	Leonardo di ser Piero da Vinci 15 April 1452 Vinci, Republic of Florence (present-day Italy)
Died	2 May 1519 (aged 67) Amboise, Kingdom of France
Known for	Art, science
Notable work	*Mona Lisa* *The Last Supper*

Entity Annotation

Terminology

- Wikipedia Knowledge Graph



- Node: Wikipedia Page (Entity)

- Link?

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]

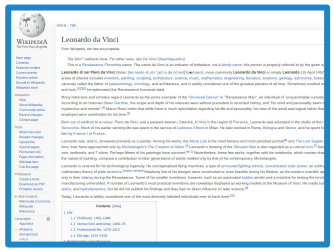
During the 20th century, the field of professional astronomy split into observational and theoret is focused on acquiring data from observations of astronomical objects, which is then analyzed Theoretical astronomy is oriented toward the development of computer or analytical models to phenomena. The 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.

Entity Annotation

Terminology

- Wikipedia Knowledge Graph



- Node: Wikipedia Page (Entity)

- Link: Wikipedia Hyperlink

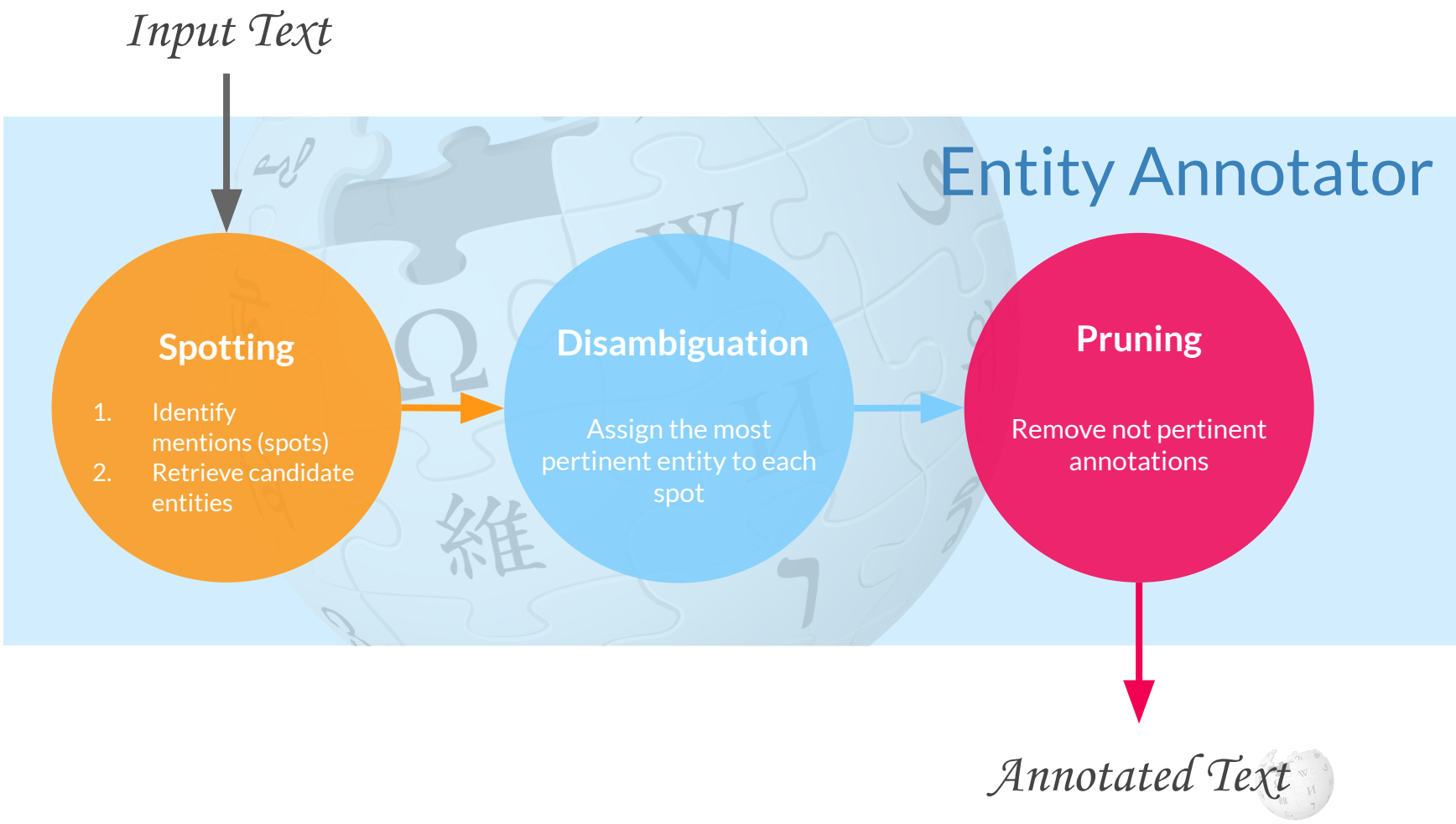
Enrich a text $\mathcal{T}$ with proper annotations

Annotation = (*mention*, entity)

Goal

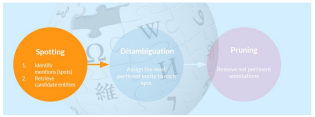
Entity Annotation

The Annotation Pipeline



Entity Annotation

The Annotation Pipeline



Spotting

Yesterday *Maradona* *won against* *Mexico.*

Mention Detection

Candidate Generation


Yesterday_(Time)

Yesterday_(Beatles_song)


Yesterday_(Guns_N_Roses_song)

...


Diego_Maradona


Diego_Sinagra

Maradona_by_Kusturica

...


Mexico


Mexico,_New_York


Mexico_national_football_team

...

Entity Annotation

The Annotation Pipeline

Spotting

1. Mention Detection
   - Named Entity Recognition (aka NER)
   - N-gram generation

2. Candidate Generation
   - Gazetteer: { mention →entities }
     - How?

Florence: Leonardo's artistic and social background

Lorenzo Ghiberti's *Gates of Paradise* (1425–52) were a source of communal pride. Many artists assisted in their creation

Florence at the time of Leonardo's youth was the centre of Christian Humanist thought and culture.[22] Leonardo commenced his apprenticeship with Verrocchio in 1466, the year that Verrocchio's master, the great sculptor Donatello, died. The painter Uccello, whose early experiments with perspective were to influence the development of landscape painting, was a very old man. The painters Piero della Francesca and Filippo Lippi, sculptor Luca della Robbia, and architect and writer Leon Battista Alberti were in their sixties. The successful artists of the next generation were Leonardo's teacher Verrocchio, Antonio del Pollaiuolo and the portrait sculptor Mino da Fiesole, whose lifelike busts give the most reliable likenesses of Lorenzo Medici's father Piero and uncle Giovanni.[53][54][55][56]

Leonardo's youth was spent in a Florence that was ornamented by the works of these artists and by Donatello's contemporaries, Masaccio, whose figurative frescoes were imbued with realism and emotion, and Ghiberti, whose *Gates of Paradise*, gleaming with gold leaf, displayed the art of combining complex figure compositions with detailed architectural backgrounds. Piero della Francesca had made a detailed study of perspective,[57] and was the first painter to make a scientific study of light. These studies and Alberti's treatise *De Pictura*[58] were to have a profound effect on younger artists and in particular on Leonardo's own observations and artworks.[53][55][56]

Massaccio's "Expulsion from the Garden of Eden" depicting the naked and distraught Adam and Eve created a powerfully expressive image of the human form, cast into three dimensions by the use of light and shade, which was to be developed in the works of Leonardo in a way that was to be influential in the course of painting. The humanist influence of Donatello's "David" can be seen in Leonardo's late paintings, particularly *John the Baptist*.[53][54]

A prevalent tradition in Florence was the small altarpiece of the Virgin and Child. Many of these were created in tempera or glazed terracotta by the workshops of Filippo Lippi, Verrocchio and the prolific della Robbia family.[53] Leonardo's early Madonnas such as *The Madonna with a carnation* and the

Article Talk

Fresco

From Wikipedia, the free encyclopedia

For other uses, see Fresco (disambiguation).

Fresco (plural ***frescos*** or ***frescoes***) is a technique of mural painting executed plaster, the painting becomes an integral part of the wall. The word fresco (Ital painting techniques, which are applied to dried plaster, to supplement painting

Contents [hide]

1 Technology
2 Other types of wall painting
3 History
3.1 Egypt and Ancient Near East

Article Talk

Lorenzo Ghiberti

From Wikipedia, the free encyclopedia

Lorenzo Ghiberti (Italian: [loˈrɛntso giˈbɛrti]; 1378 – 1 December 1455), born Florence Baptistery, called by Michelangelo the *Gates of Paradise*. Trained a important writing on art, as well as what may be the earliest surviving autobio

Contents [hide]

1 Life
1.1 Early life
1.2 Florence Baptistery doors

Article Talk

Florence Baptistery

From Wikipedia, the free encyclopedia

The **Florence Baptistery** (Italian: *Battistero di San Giovanni*), also known as t baptistery stands in both the Piazza del Duomo and the Piazza San Giovanni,

The Baptistery is one of the oldest buildings in the city, constructed between 1 the Pisan Romanesque or Lombard styles, its influence was decisive for the s Brunelleschi, and the other architects created Renaissance architecture. In the the late antique architectural tradition in Italy, as in the cases of the Basilica of

The Baptistry is renowned for its three sets of artistically important bronze doo Ghiberti.[1] The east doors were dubbed by Michelangelo the *Gates of Paradis*

The Italian poet Dante and many other notable Renaissance figures, including Florentines were baptized here.

Entity Annotation

The Annotation Pipeline

Spotting

1. Mention Detection
   - Named Entity Recognition (aka NER)
   - N-gram generation

2. Candidate Generation
   - Gazetteer: { mention →entities }
     - How? Wikipedia anchor texts!
     - *Ranking (+ Thresholding)*
       - Commonness (Ferragina, CIKM '10; Guo, CIKM '14)
       - Entity-context Similarity (Zwicklbauer, SIGIR '16)
       - ...

Entity Annotation

The Annotation Pipeline

Disambiguation

Yesterday *Maradona* *won against* *Mexico*.

Yesterday_(Time)	 Diego_Maradona	Mexico
 Yesterday_(Beatles_song)	Diego_Sinagra	Mexico_New_York
Yesterday_(Guns_N_Roses_song)	Maradona_by_Kusturica	 Mexico_national_football_team
...	...	...

Entity Annotation

The Annotation Pipeline

Disambiguation

Yesterday *Maradona* *won against* *Mexico.*

Yesterday_ (Beatles_song)

0.1

Diego_Maradona

0.8

Mexico_national_ football_team

0.7

- Spots have been disambiguated
  - Ambiguous lexical elements (words) are now labeled with unambiguous concepts
- Finally, coherence scores are assigned

Entity Annotation

The Annotation Pipeline

Disambiguation

LS NED (Cucerzan, ACL '07)

Wikifier (Mihalcea, CIKM '07)

TAGME (Scaiella, CIKM '10)

DBpedia Spotlight (Mendes, SemSys '11)

...

fy (Moro, ACL '14)

WAT (Piccinno, SIGIR '14)

AIDA (Nguyen, LDOW '14)

PBoH (Ganea, WWW '16)

DoSeR (Zwicklbauer, SIGIR '16)

Entity Annotation

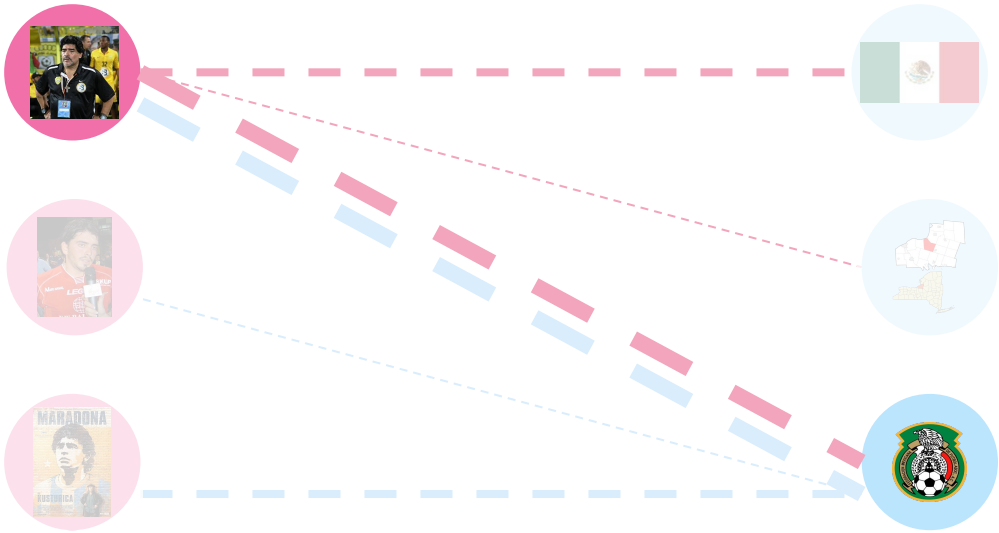
The Annotation Pipeline

Disambiguation

Algorithm: TAG-ME WAT (Scaiella, CIKM '10; Piccinno, SIGIR '14)

[...] Maradona won against Mexico.

- Voting Scheme
- M&W / Jaccard Relatedness



Entity Annotation

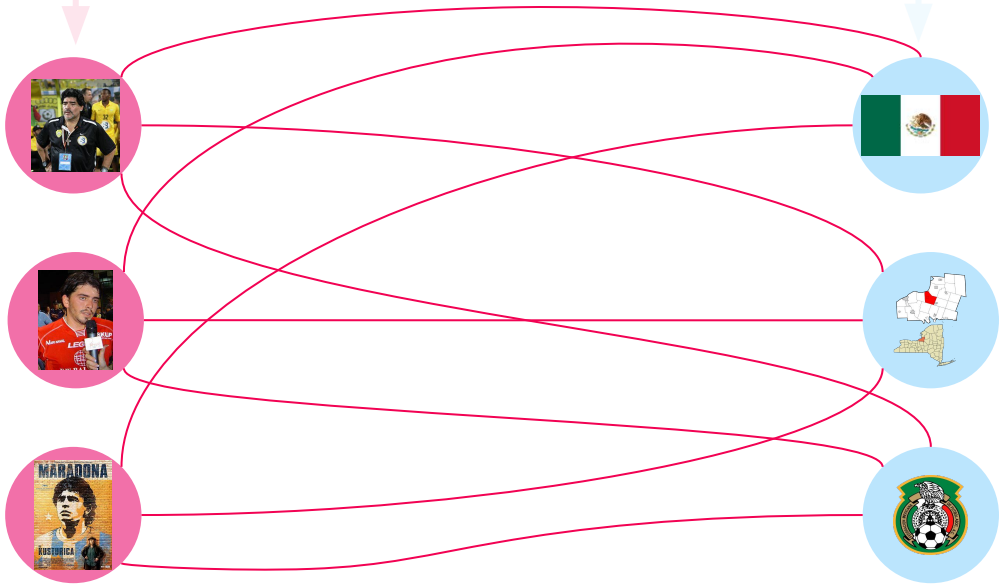
The Annotation Pipeline

Disambiguation

Algorithm: DoSeR (Zwicklbauer, SIGIR '16)

[...] Maradona won against Mexico.

- Graph of candidates



Entity Annotation

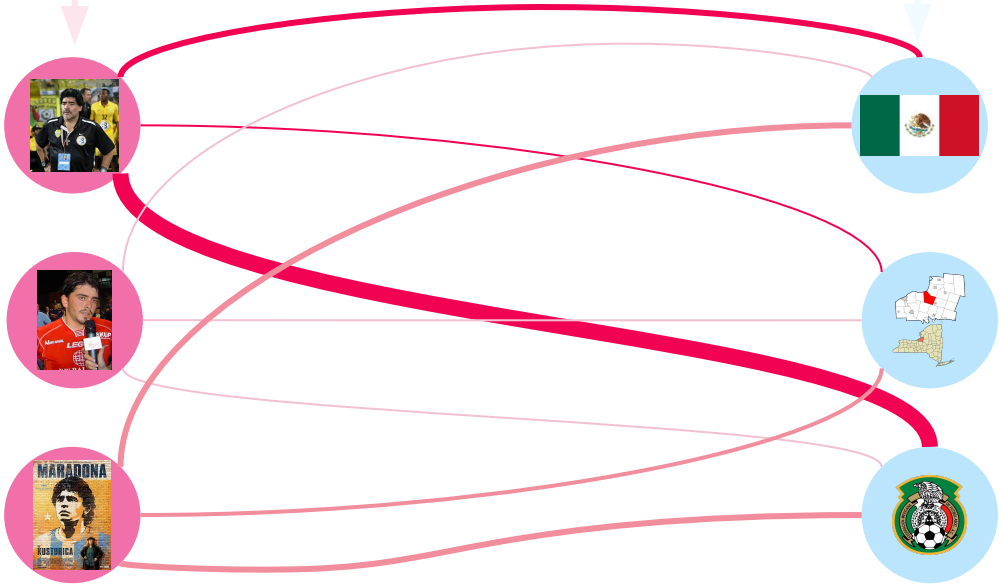
The Annotation Pipeline

Disambiguation

Algorithm: DoSeR (Zwicklbauer, SIGIR '16)

[...] Maradona won against Mexico.

- Graph of candidates
- Entity2Vec Relatedness



Entity Annotation

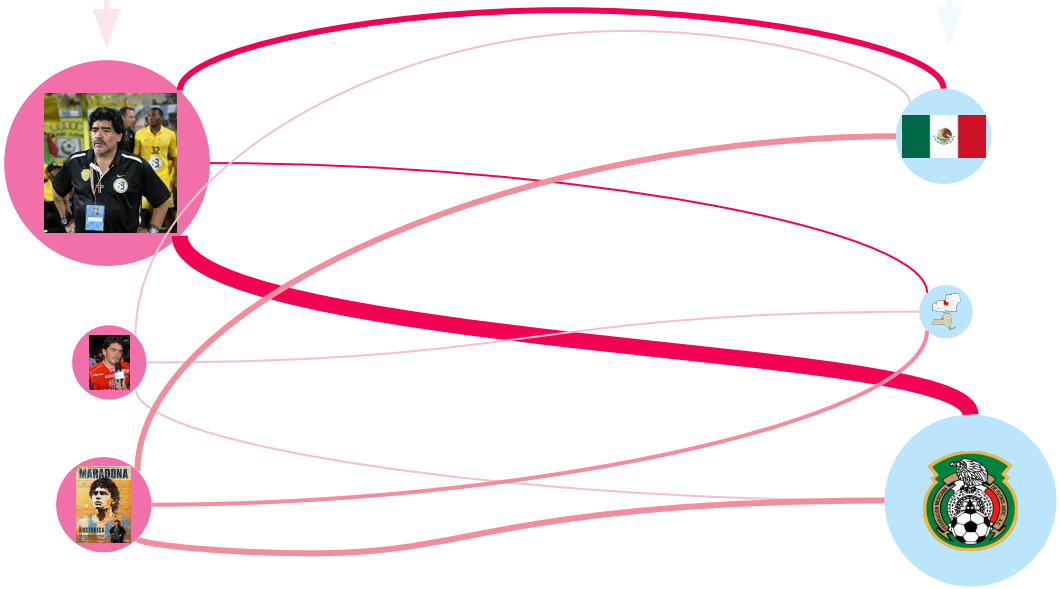
The Annotation Pipeline

Disambiguation

Algorithm: DoSeR (Zwicklbauer, SIGIR '16)

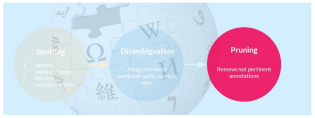
[...] Maradona won against Mexico.

- Graph of candidates
- Entity2Vec Relatedness
- PageRank



Entity Annotation

The Annotation Pipeline



Pruning

Yesterday *Maradona* *won against* *Mexico*.

Yesterday_ (Beatles_song)

0.1

Diego_Maradona

0.8

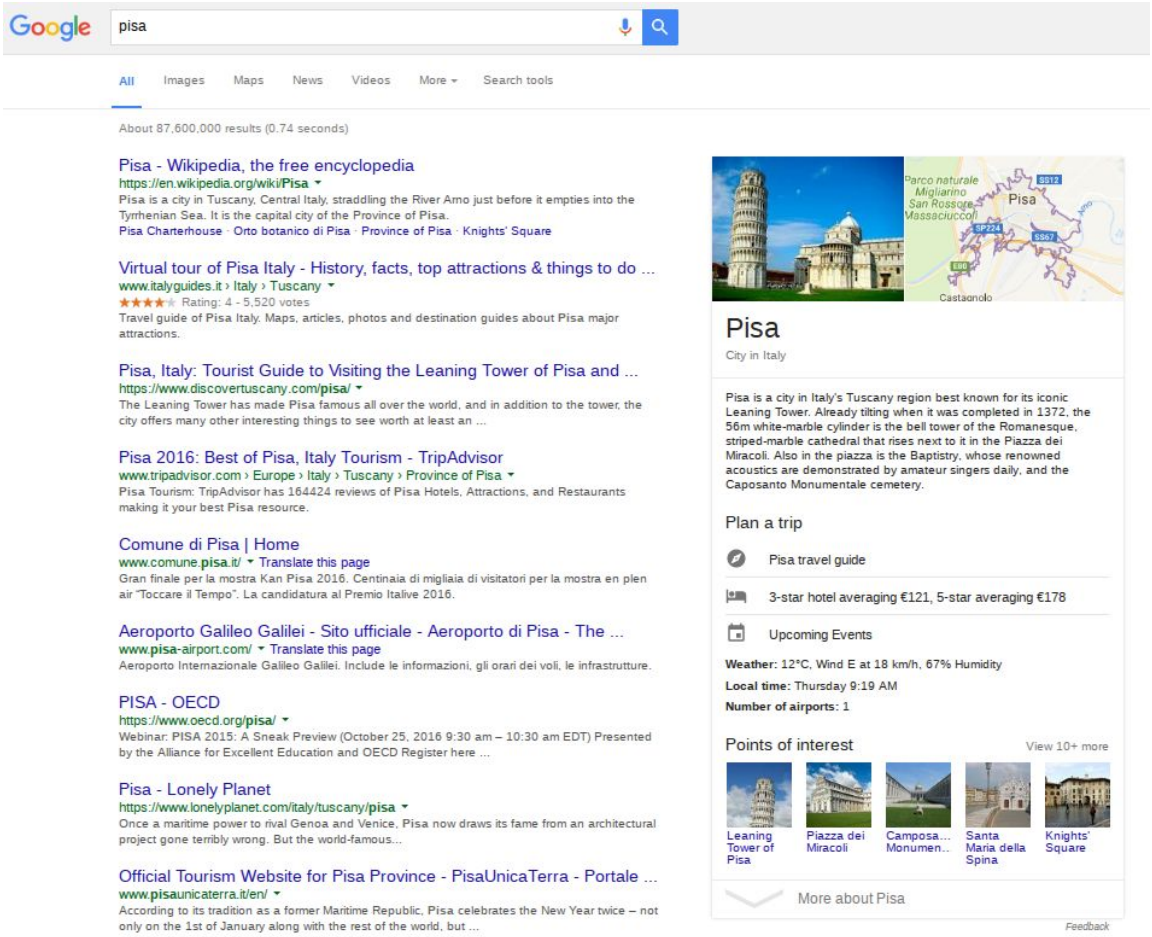
Mexico_national_ football_team

0.7

- Remove not pertinent annotations
- Clear text from erroneous annotations
- Coherence thresholding

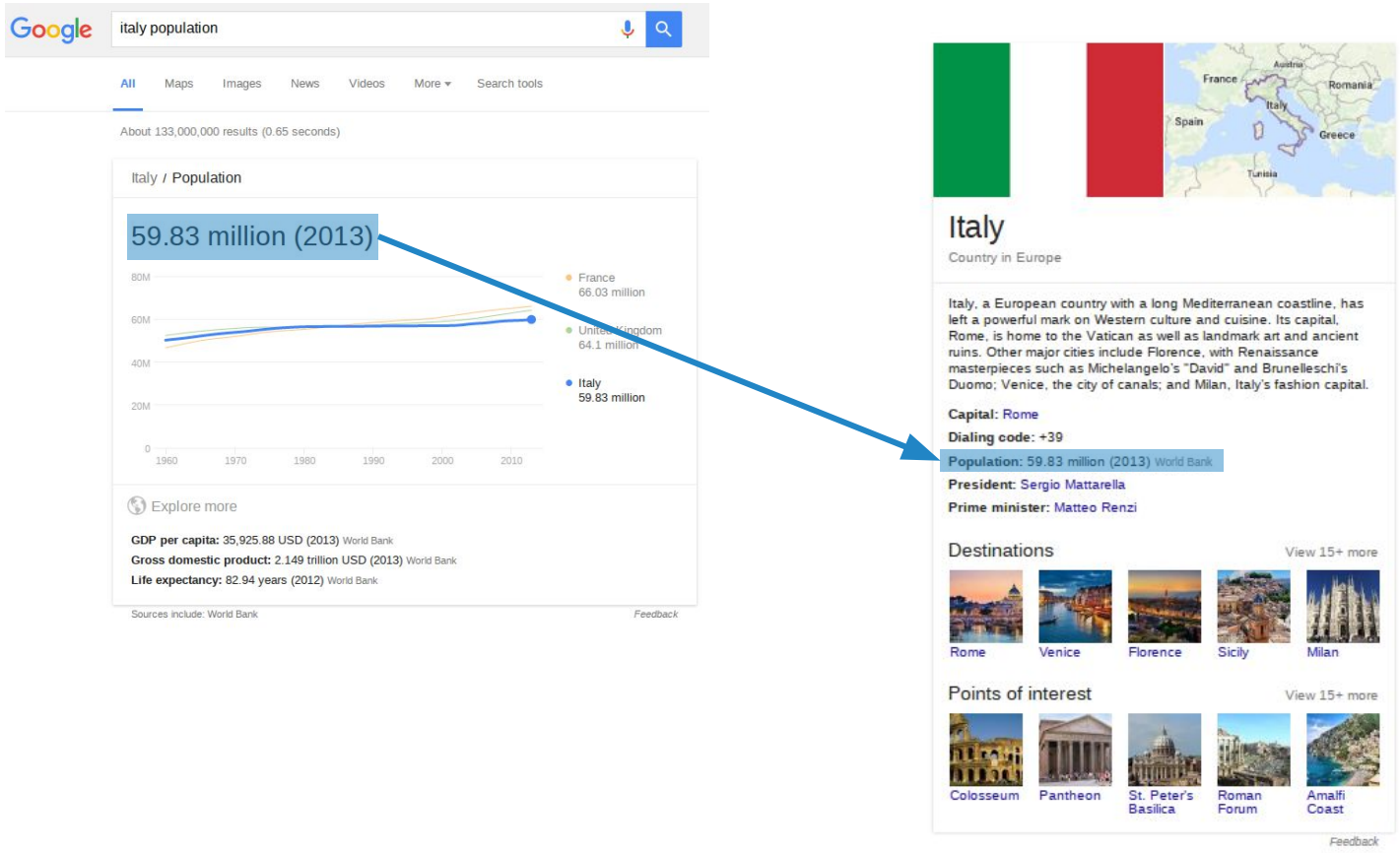
Applications

Web Search Results (Gabrilovich, SIGIR '16)



Applications

Web Search Results (Gabrilovich, SIGIR '16)



Applications

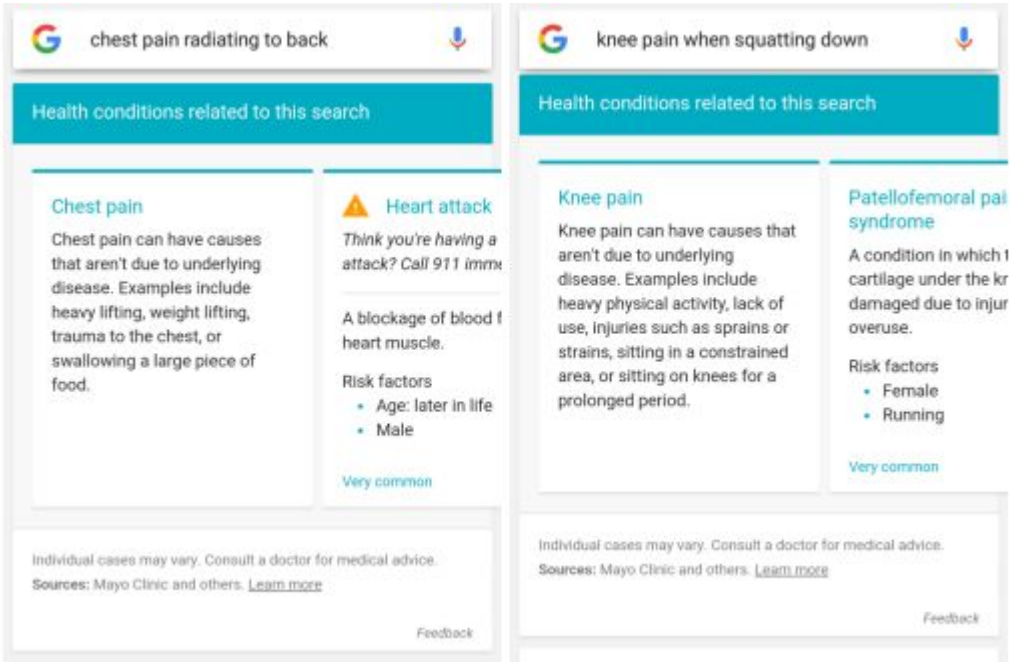
Question Answering (Gabrilovich, SIGIR '16)

Google

amazon

Applications

Implicit Questions (Gabrilovich, SIGIR '16)

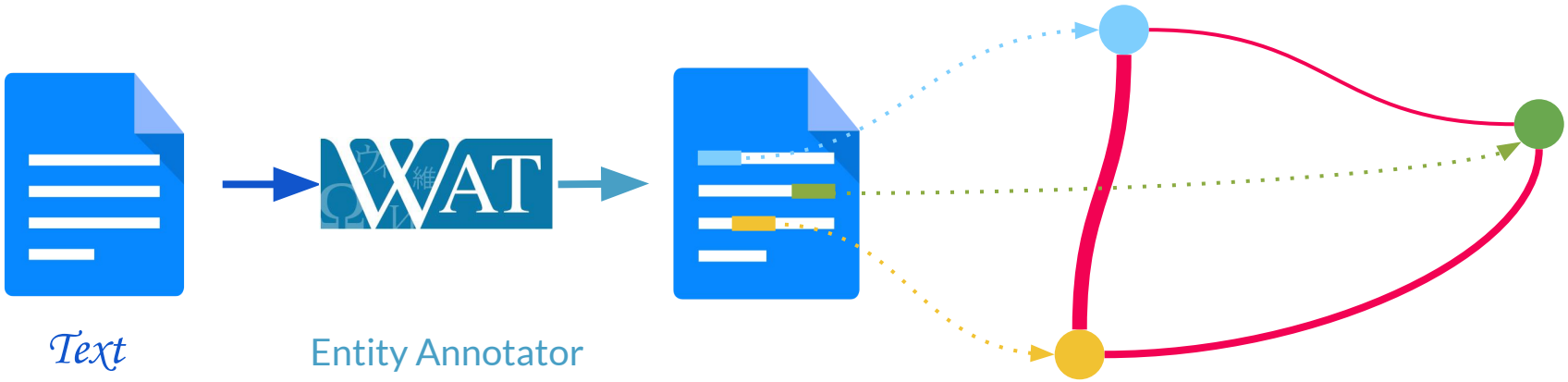


Condition → *What does it mean?*

Symptoms →*What do they indicate?*

A New Text Representation

- Originally introduced by (Scaiella, WSDM '12)
  - Widely deployed (Dunietz, EACL '14; Schuhmacher, WSDM '14; Ni, WSDM '15), ...
- *Text* = Graph of Entities
- What about...



A New Text Representation

- Originally introduced by (Scaiella, WSDM '12)
  - Widely deployed (Dunietz, EACL '14; Schuhmacher, WSDM '14; Ni, WSDM '15), ...
- *Text* = Graph of Entities

- What about...
  - ...edge weights?
  - ...node weights?

} Work done in the first year

2.
Work done in the first year

Entity Relatedness & Document Aboutness

Entity Relatedness

Entity Relatedness

Compute how much two entities are related

Relatedness : Entities x Entities → Real

Goal

- How much related are...
  - ...Bank with Money?
  - ...Wood with Book?

- Semantic Reasoning:
  - Human: Background Knowledge
  - Machines: Knowledge Graph

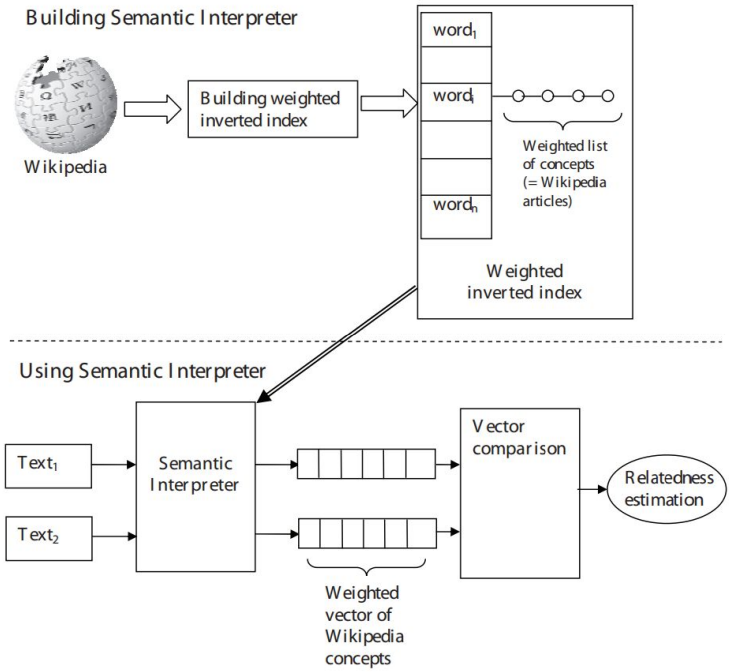
Entity Relatedness

(A brief list of) Algorithms and Applications

- Document/Word Similarity
  - WikiRelate (Strube, AAI '06)
  - Explicit Semantic Analysis (Gabrilovich, IJCAI '07)
    - WikiWalk (Yeh, ACL '09)
    - Temporal Semantic Analysis (Radinsky, WWW '11)
    - Concept Graph Representation (Ni, WSDM '16)
  - Milne & Witten (Witten, AAI '08)
  - Salient Semantic Analysis (Hassan, AAI '11)
- Machine Translation (Agirre, NAACL '09; Rothe, ACL '14)
- Document Classification (Perozzi, WWW '14; Tang, WWW '15)
- ...

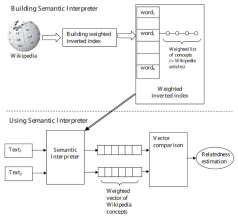
Entity Relatedness

- Two entities are related whether…
  - …they are described by related texts (Corpus-based)
    - Example: ESA (Gabrilovich, IJCAI '07)
      - Concepts grounded in human cognition
      - Opposite to *latent* concepts



Entity Relatedness

- Two entities are related whether…
  - …they are described by related texts (Corpus-based)
    - Example: ESA (Gabrilovich, IJCAI '07)
      - Concepts grounded in human cognition
      - Opposite to *latent* concepts



  - …they are referenced by related entities (Graph-based)
    - Example: CoSimRank (Rothe, ACL '14)

Entity Relatedness

CoSimRank (Rothe, ACL '14)

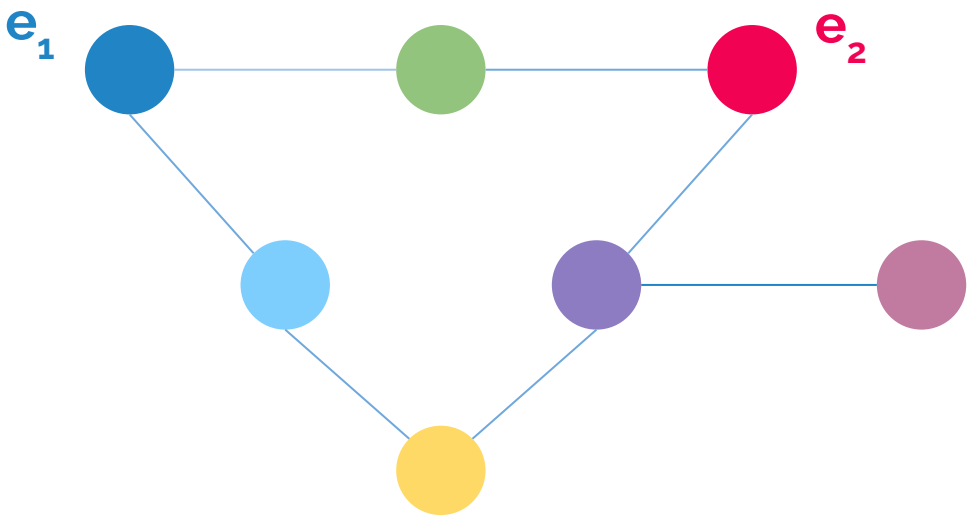
- Graph-based approach
- Relatedness algorithm for nodes in a graph
- Exploits Random Walks
- Algorithm (in brief)

1. *Sets damping vectors for* e_1 *and* e_2
2. *Runs an iteration of PageRank*
3. *Updates relatedness score*

$e_1, e_2 \in$ Entities

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^0(e_1)$

1.0
0.0
0.0
0.0
0.0
0.0
0.0

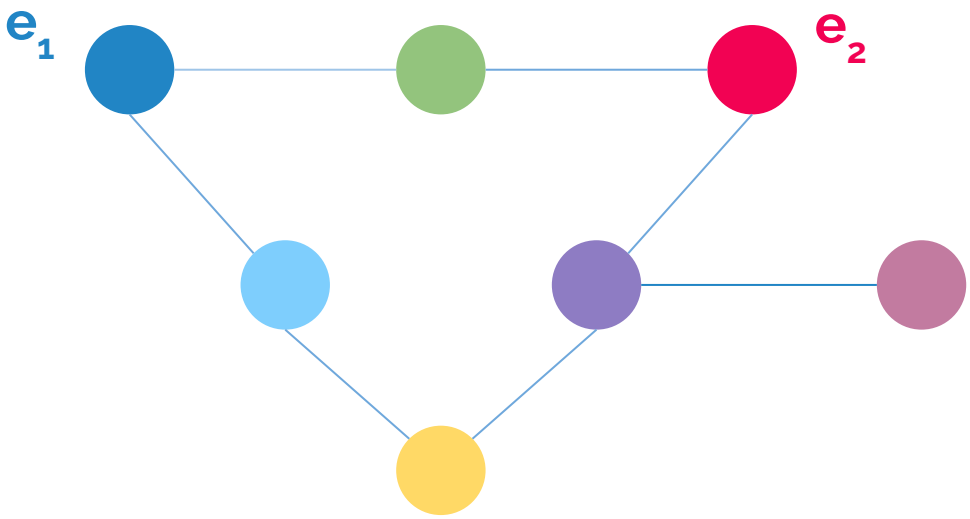
$p^0(e_2)$

0.0
0.0
1.0
0.0
0.0
0.0
0.0

$\text{Relatedness}^0(e_1, e_2) = 0.0$

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^1(e_1)$

0.2
0.4
0.0
0.4
0.0
0.0
0.0

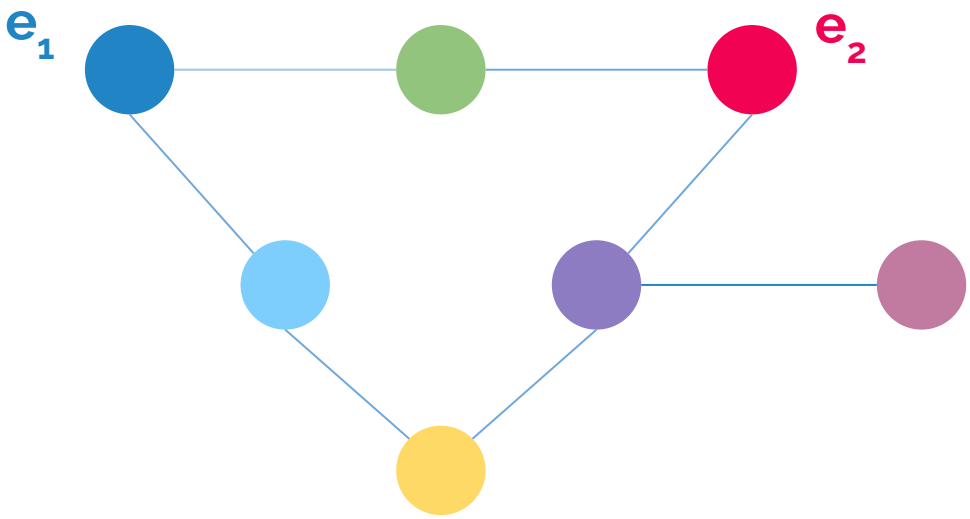
$p^1(e_2)$

0.0
0.4
0.2
0.0
0.4
0.0
0.0

$\text{Relatedness}^1(e_1, e_2) = 0.16$

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^2(e_1)$

0.52
0.08
0.16
0.08
0.0
0.0
0.16

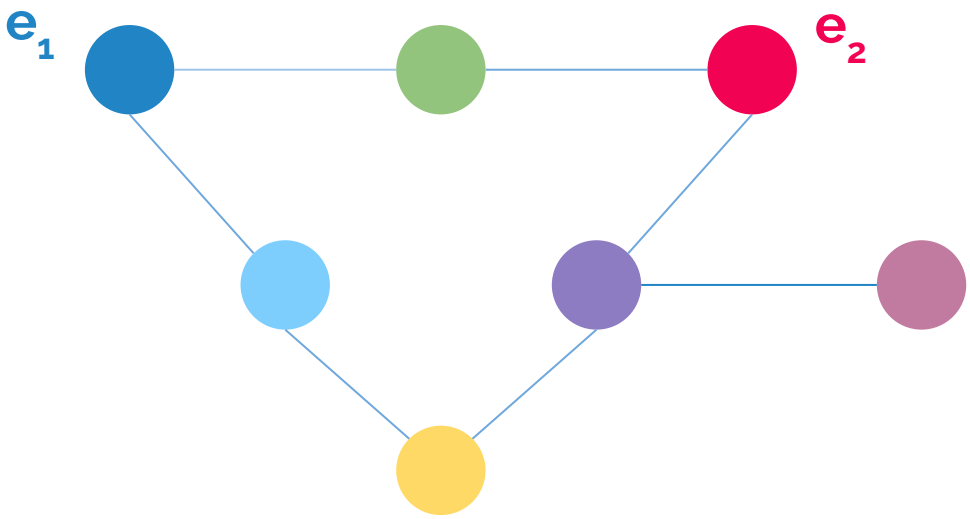
$p^2(e_2)$

0.16
0.08
0.46
0.0
0.05
0.21
0.10

$\text{Relatedness}^2(e_1, e_2) = 0.33$

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^3(e_1)$

0.26
0.27
0.03
0.27
0.08
0.00
0.03

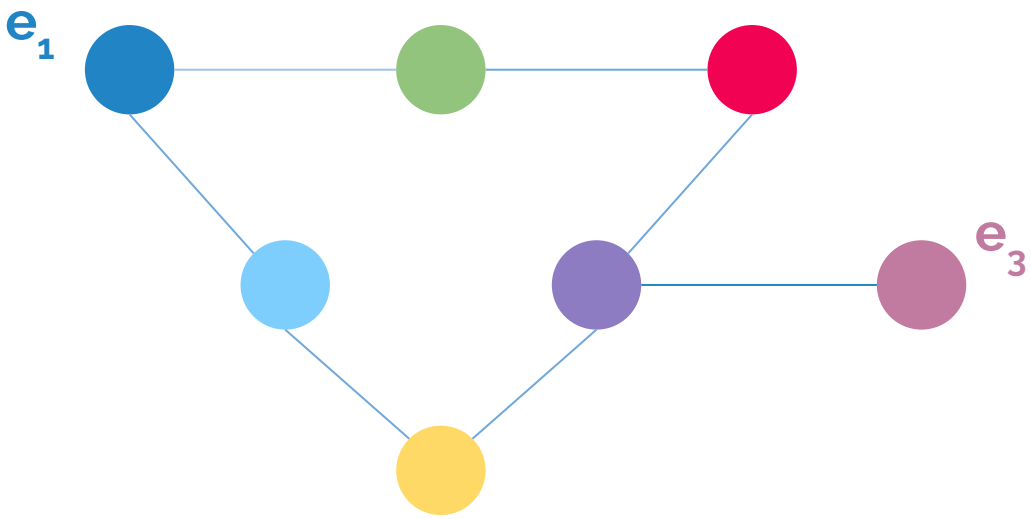
$p^3(e_2)$

0.03
0.25
0.25
0.10
0.20
0.04
0.02

$\text{Relatedness}^3(e_1, e_2) = 0.47$

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^0(e_1)$

1.0
0.0
0.0
0.0
0.0
0.0
0.0

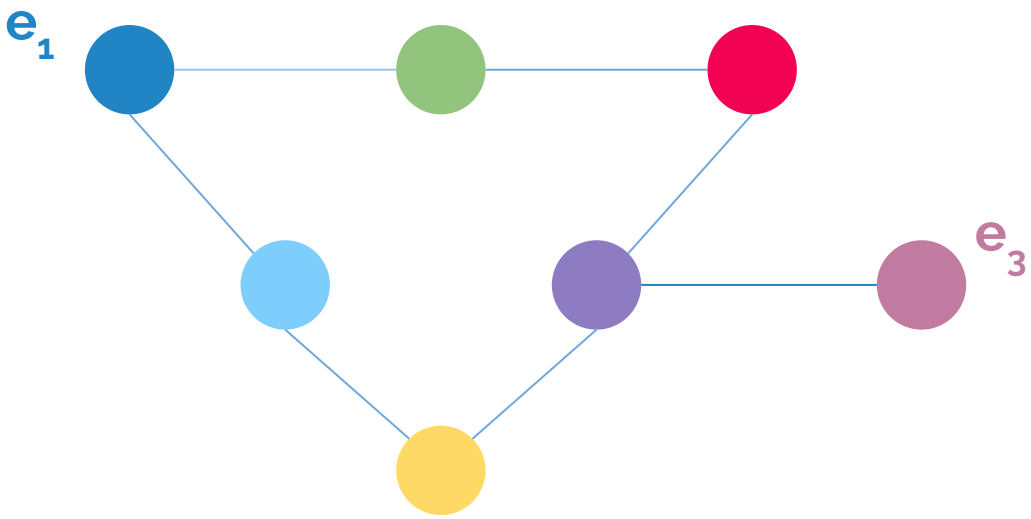
$p^0(e_3)$

0.0
0.0
0.0
0.0
0.0
1.0
0.0

$\text{Relatedness}^0(e_1, e_3) = 0.0$

Entity Relatedness

CoSimRank (Rothe, ACL '14)



$p^3(e_1)$

0.26
0.27
0.03
0.27
0.08
0.00
0.03

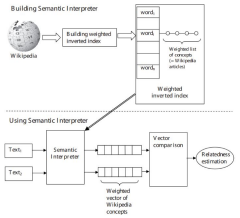
$p^3(e_2)$

0.0
0.04
0.02
0.04
0.16
0.24
0.02

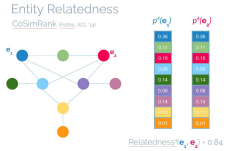
$\text{Relatedness}^3(e_1, e_3) = 0.13$

Entity Relatedness

- Two entities are related whether…
  - …they are described by related texts (Corpus-based)
    - Example: ESA (Gabrilovich, IJCAI '07)
      - Concepts grounded in human cognition
      - Opposite to *latent* concepts



  - …they are referenced by related entities (Graph-based)
    - Example: CoSimRank (Rothe, ACL '14)



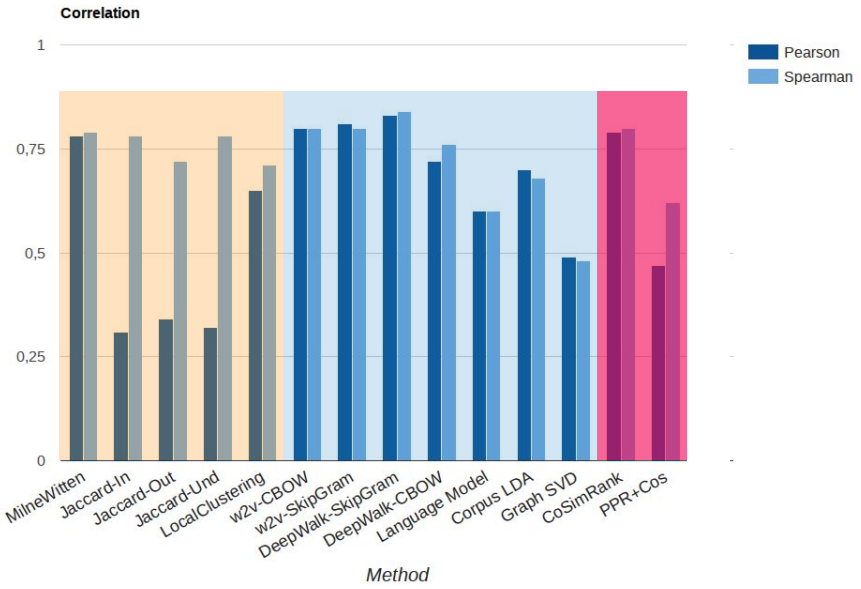
- *Need* of a fair and meaningful comparison

Entity Relatedness

Preliminary Results: The Relatedness Framework

- Design algorithms based on
  - Set Operations (Milne & Witten, Jaccard, ...)
  - Embeddings (Word2Vec, LDA, ...)
  - Random Walk (CoSimRank, PPR+Cos)

- Preliminary results
  - Analyse entity pairs
  - Deploy corpus-based algorithms
  - A new algorithm: LLP



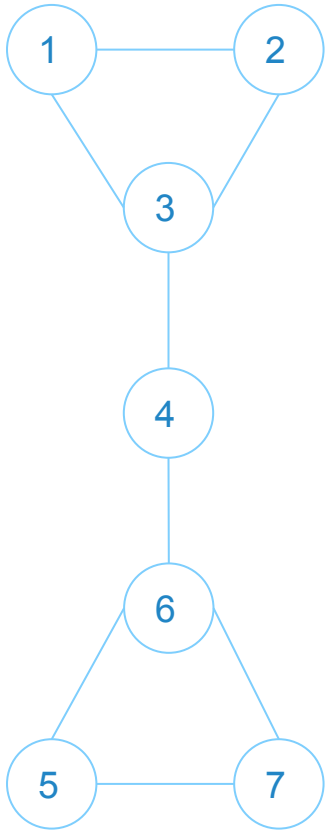
Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

- Standard Label Propagation (Newman, Phys. Rev. '04)
  - Clustering algorithm (node labeling)
  - Pro: Scales on very large graphs
  - Cons: Can generate few big clusters
- Layered Label Propagation
- Standard Label Propagation with a resolution parameter γ
  - Graph compression
  - Algorithm (in brief)
    1. *Randomly initialize each node with a* label (cluster)
    2. *Update label according to a specific* rule
       - Maximize nonlocal discount (Ronhovde, Phys. Rev. '10)

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



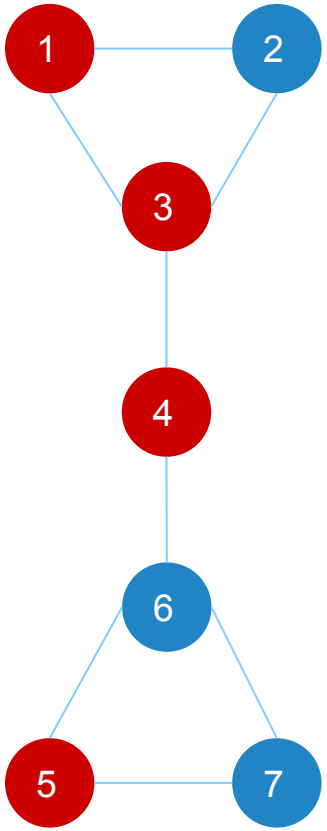
Round: 1

Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



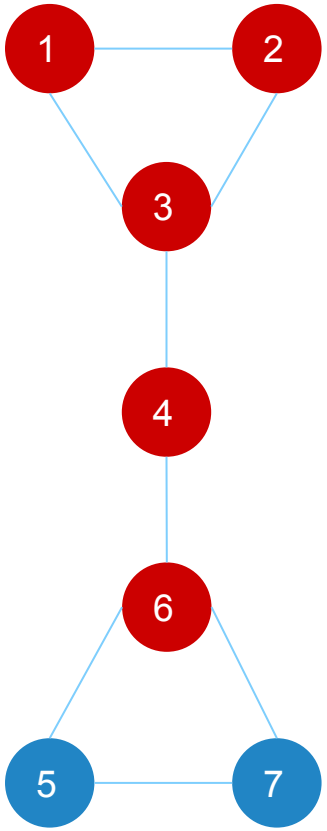
Round: 1

Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



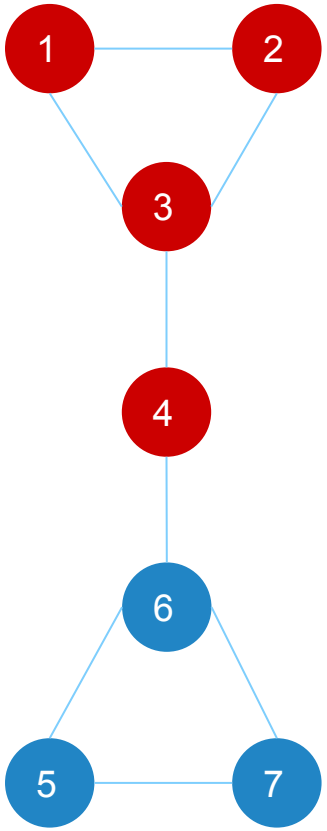
Round: 1

Step: Updating (1)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



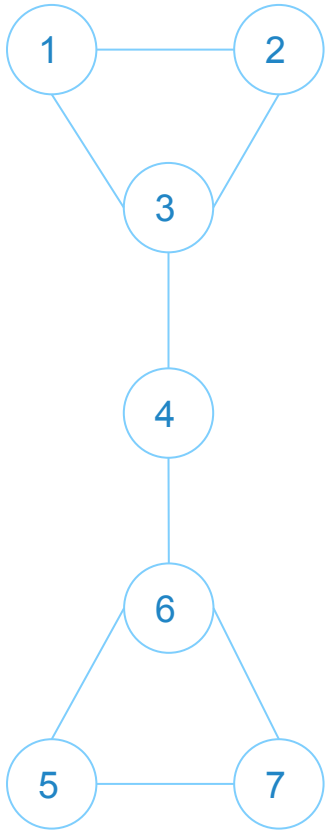
Round: 1

Step: Updating (2)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



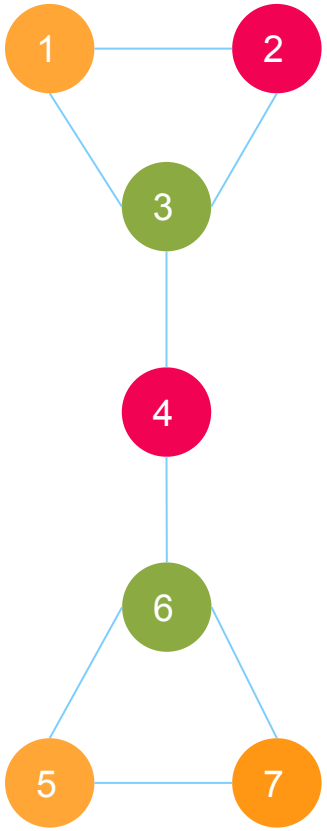
Round: 2

Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



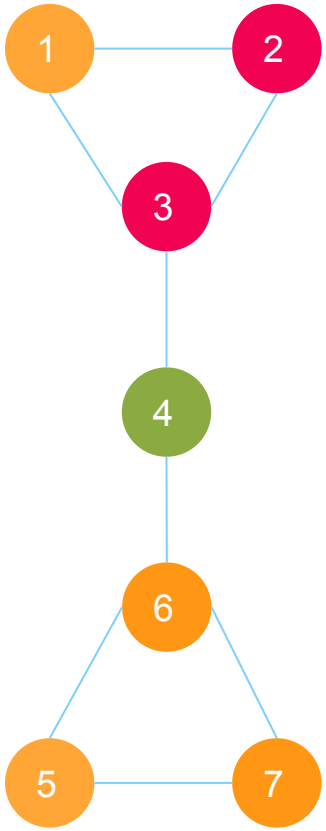
Round: 2

Step: Initialization

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



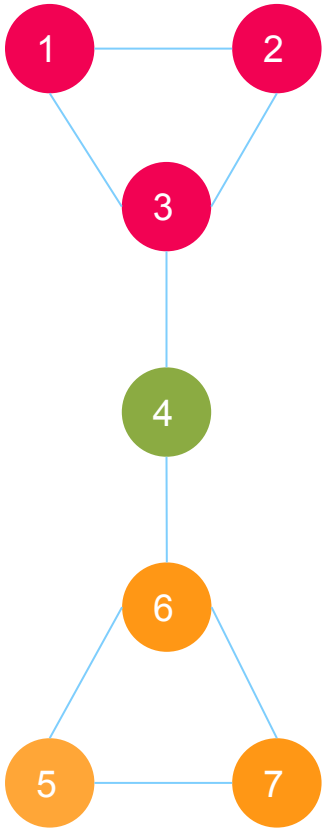
Round: 2

Step: Updating (1)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



Round: 2

Step: Updating (2)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

Round: 2

Step: Updating (2)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)

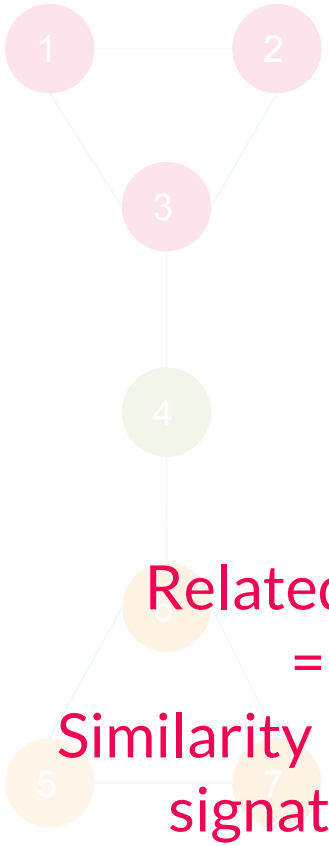
Round: 2

Step: Updating (2)

Node	Labels
1	
2	
3	
4	
5	
6	
7	

Entity Relatedness

A New Algorithm: Layered Label Propagation (Boldi, WWW '11)



Relatedness
=
Similarity between signatures

Round: 2
Step: Updating (2)

Node	Labels			
1				
2				
3				
4				
5				
6				
7				

Signatures

Document Aboutness

Document Aboutness

Aboutness = Succinct representation of a document's subject matter (Hutchins, 1977)

Goal

- Weight information (e.g. entities, words, ...) within a document

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	Weight
Barack_Obama	0.85
Hillary_Clinton	0.8
Hawaii	0.3
George_W._Bush	0.2
…	…

Document Aboutness

Aboutness = Succinct representation of a document's subject matter (Hutchins, 1977)

Goal

- Weight information (e.g. entities, words, ...) within a document
- Wide range of practical applications:
  a. Recommendation
  b. Categorization
  c. Exploratory search
  d. Web Ranking
  e. …

Document Aboutness

	Keyphrase Extraction	Entity Salience
Aboutness	Words Proper nouns Sentences ...	Entities 
Candidate Extraction	Dictionary POS tags ...	Entity Annotation
Subject Matter Identification	Ranking/Classification	
Issues	Interpretation Overgeneration Infrequency Redundancy	Dependency on KG ?

Keyphrase Extraction

"[...] errors could be addressed using background knowledge."

(Hasen, ACL '14)

"[...] features more directly linked to Wikipedia [...] can provide more focused background information."

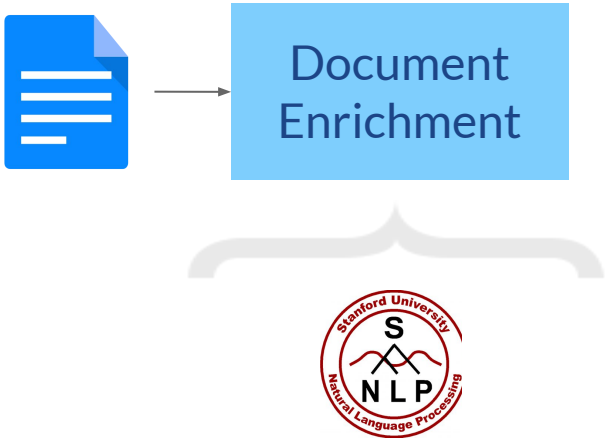
(Dunietz, EACL '14)

Entity Salience

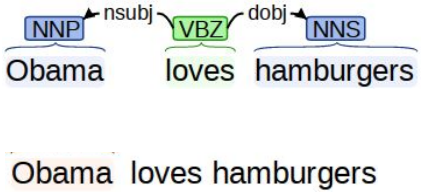
Document Aboutness

Our Proposal

- Entity Salience Approach



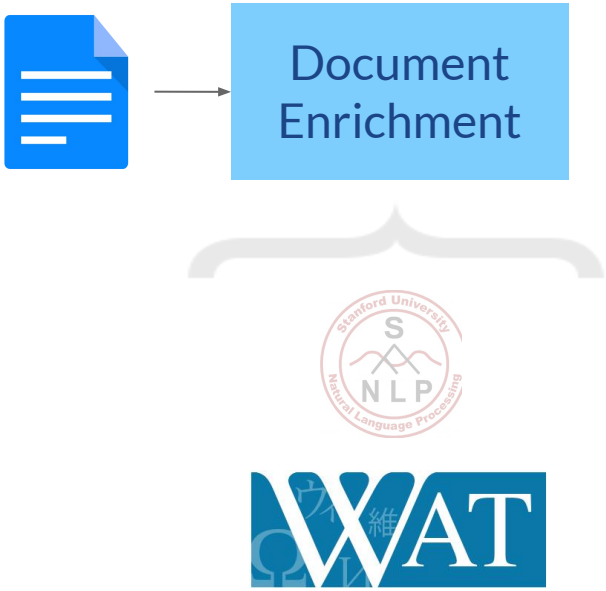
- Pos tagging
- Mention detection
- Dependency parsing
- Co-reference resolver.



Document Aboutness

Our Proposal

- Entity Salience Approach

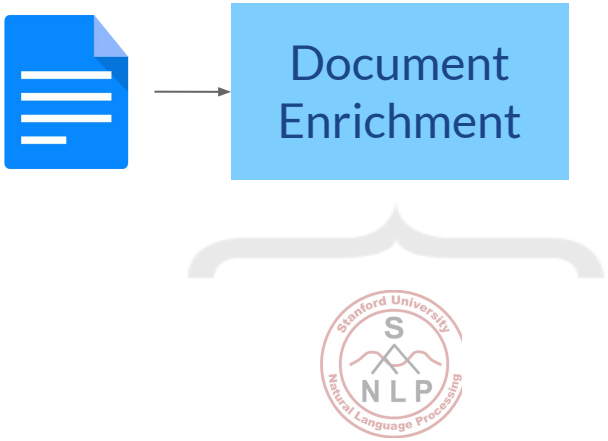


- Entity annotation
- Graph of entities
- Relatedness

Document Aboutness

Our Proposal

- Entity Salience Approach



- Document Summarizer
- Sentence Ranker

Document Aboutness

Our Proposal

- Entity Salience Approach

Document

Document Enrichment

Feature Generation

- Entity →Feature Vector
- Classical syntactic features
  - Frequency
  - Position
  - ...
- New syntactic and semantic features:
  - Position-based
  - Dependency-based
  - Relatedness-based
  - Centrality-based
  - ...

Stanford University S NLP Natural Language Processing

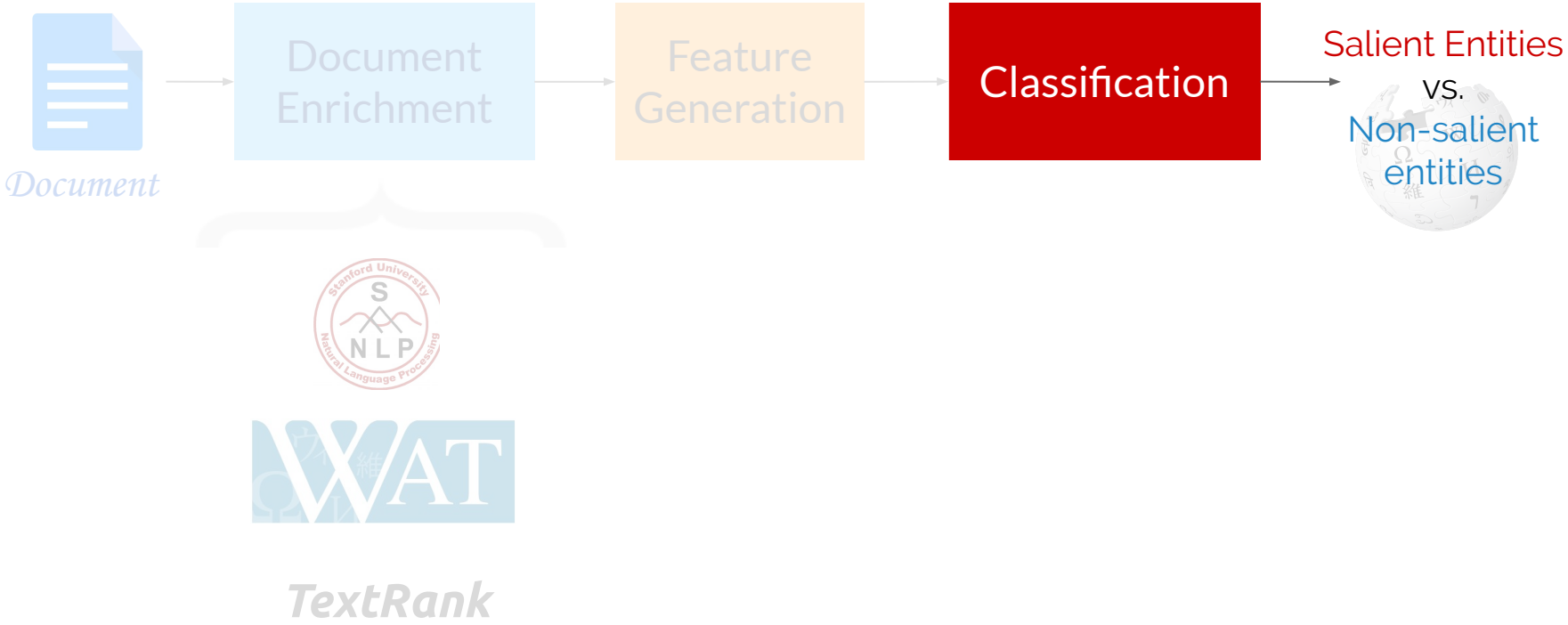
WAT

TextRank

Document Aboutness

Our Proposal

- ▷ Entity Salience Approach



Document Aboutness

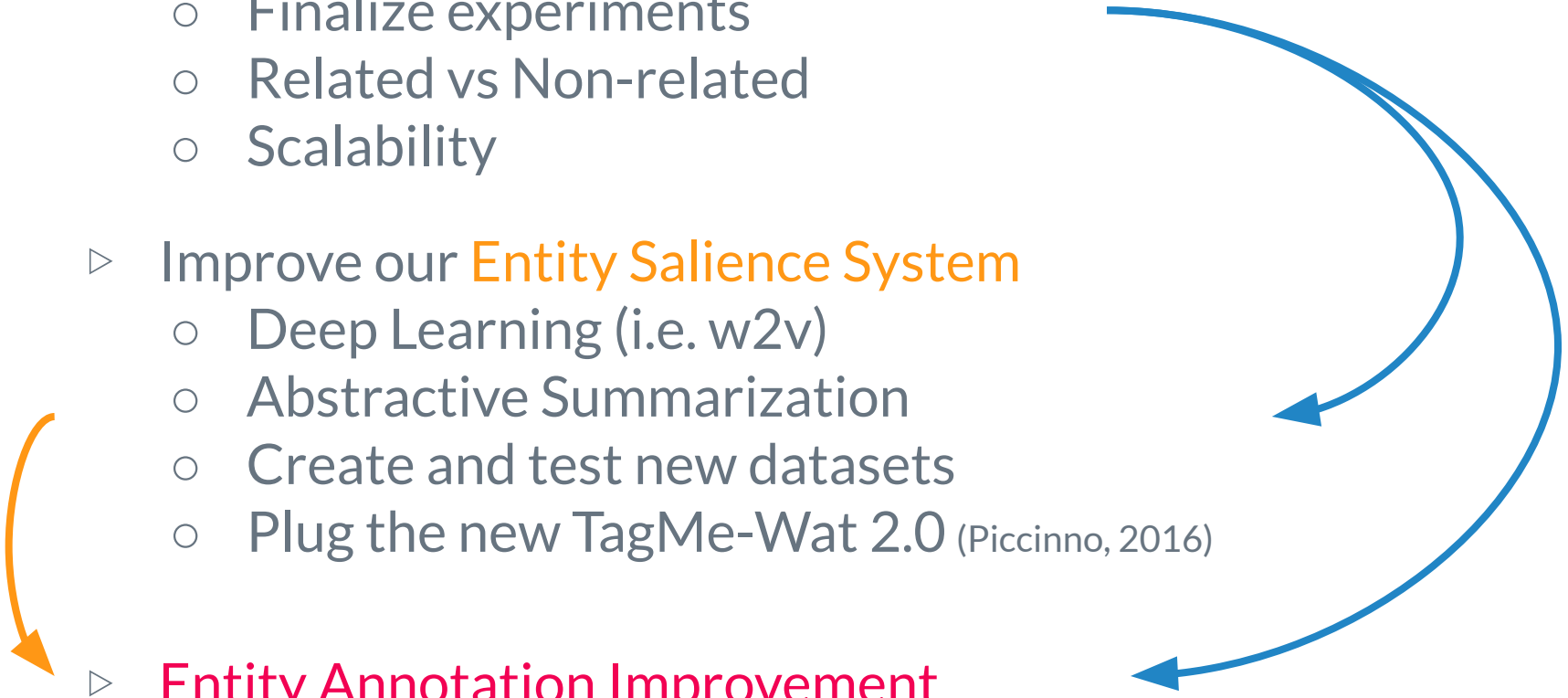
Our Proposal: Main Contributions

- Fully documented system
- Public available via Web-API
- Improvement of state-of-the-art (Cmu-Google, F1: 61.5)
  - New York Times' dataset (110,000 news, 1.3M entities)
  - *62.6 micro-F1 (+1.1%) and 59.5 macro-F1 (+2.5%)*
  - More robust when entities are not biased at the beginning (+9%)
- Deep Feature and Error Analysis

3.
Future Work

Future Work

- Conclude Entity Relatedness
  - Finalize experiments
  - Related vs Non-related
  - Scalability
- Improve our Entity Salience System
  - Deep Learning (i.e. w2v)
  - Abstractive Summarization
  - Create and test new datasets
  - Plug the new TagMe-Wat 2.0 (Piccinno, 2016)
- Entity Annotation Improvement



Thanks!

Any questions?