Algorithms and Applications for Web-Scale Knowledge Graphs

Marco Ponza

Supervisor
Prof. Paolo Ferragina
Menu

1. Document Aboutness  

2. Entity Relatedness  

What next? Last year of PhD
1. Document Aboutness
Introduction
The Document Aboutness Task

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

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
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!

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**

▶ **SEL** *(Trani, DocEng ‘16)*

- Supervised Entity Annotator
- Large Feature Space + Decision Tree

Limitations
- No comparison with CMU-Google System
- Benchmark on small dataset
- Experimented on the WIKINeWS Dataset (365 news, 4747 entities)
- 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

Input Document

Document Enrichment
1. CoreNLP
2. TextRank
3. WAT

Feature Generation
- Basics
- Syntactic
- Semantic

Classification
Classify entities in Salient or Non-Salient

Salient Entities
1. Document Enrichment

- CoreNLP (Manning, ACL '14)
1. Document Enrichment

Module

Sentence Splitting

CoreNLP (Manning, ACL '14)
Entity Salience
Three-Stage System

1. Document Enrichment

CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
Entity Salience
Three-Stage System

1. Document Enrichment

CoreNLP (Manning, ACL '14)

Module
Sentence Splitting
Tokenization
POS-Tagging
Entity Salience
Three-Stage System

1. Document Enrichment

CoreNLP (Manning, ACL '14)

Module

Sentence Splitting
Tokenization
POS-Tagging
Named Entity Recognition

Images via http://corenlp.run
Entity Salience
Three-Stage System

1. Document Enrichment

CoreNLP (Manning, ACL '14)

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

Images via http://corenlp.run
Entity Salience
Three-Stage System

1. Document Enrichment
   ▶ CoreNLP (Manning, ACL '14)

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

Images via http://corenlp.run
Entity Salience
Three-Stage System

1. Document Enrichment
   - CoreNLP (Manning, ACL '14)
     - Graph-Based Summarizer
       - Nodes = Sentences
       - Weights = Normalized Token Overlap
     - Sentence Ranking via PageRank
   - TextRank (Mihalcea, EMNLP '04)

---

*The New York Times*

**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.
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

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

1. Document Enrichment
   - **CoreNLP**  (Manning, ACL ’14)
     - Named Entities + Proper/Common Nouns
   - **WAT**  (Piccinno, SIGIR ’14)
     - Annotates them with Wikipedia Entities

The New York Times

**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, 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.

**Barack Obama**

**Hilary Clinton**

**George Walker Bush**
**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

---

*The New York Times*

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 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.
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

1. Document Enrichment

- **CoreNLP** (Manning, ACL ’14)
  - Named Entities + Proper/Common Nouns

- **WAT** (Piccinno, SIGIR ’14)
  - Annotates them with Wikipedia Entities
  - Relatedness 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
## Feature Generation

### Standard Entity Features
- Frequency
- Positions
- ...

### CMU-Google Features
- POS-Tags, Coreference Freq.
- PageRank on a graph whose weights are based on co-occurrence.
- ...

### 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
- ...

### Novel Features Introduced

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spread(...)</td>
<td>Difference between the maximum and minimum sentence- (resp. token-) positions of $e$ in $d$.</td>
</tr>
<tr>
<td>bucketed-freq(...)</td>
<td>Vector of bucketed frequencies through sentence- (resp. token-) positions of $e$ in $d$.</td>
</tr>
<tr>
<td>textrank-stats(...)</td>
<td>Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of TextRank scores of sentences where $e$ appears in $d$.</td>
</tr>
<tr>
<td>dep-freq(...)</td>
<td>Frequency of $e$ in $d$ when it appears as dependent of the dependency relation $dep$.</td>
</tr>
<tr>
<td>dep-bucketed-freq(...)</td>
<td>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.</td>
</tr>
<tr>
<td>dep-position-stats(...)</td>
<td>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.</td>
</tr>
<tr>
<td>dep-textrank-stats(...)</td>
<td>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.</td>
</tr>
<tr>
<td>rel-stats(...)</td>
<td>Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the relatedness scores between $e$ and all other entities annotated in $d$.</td>
</tr>
<tr>
<td>rel-bucketed-stats(...)</td>
<td>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).</td>
</tr>
<tr>
<td>common-stats(...)</td>
<td>Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the commonness values of $e$ in $d$ (see \textsc{Wat}).</td>
</tr>
<tr>
<td>p-stats(...)</td>
<td>Minimum, maximum, arithmetic mean, median, standard deviation and harmonic mean of the $p$-score values of $e$ in $d$ (see \textsc{Wat}).</td>
</tr>
<tr>
<td>rel-centrality(...)</td>
<td>Degree, PageRank, Betweenness, Katz, HITS, Closeness and Harmonic \cite{2} scores of $e$ computed on the entity graph of $d$. Details in text.</td>
</tr>
</tbody>
</table>
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

Salient Entities

(Chen, SIGKDD '16)
Experiments
Results

<table>
<thead>
<tr>
<th>System</th>
<th>New York Times</th>
<th></th>
<th></th>
<th>Wikinews</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro</td>
<td>Macro</td>
<td></td>
<td>Macro</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>CMU-Google (Dunietz, EACL '14)</td>
<td>60.5</td>
<td>63.5</td>
<td>61.5</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>CMU-Google-ours</td>
<td>58.8</td>
<td>62.6</td>
<td>60.7</td>
<td>35.7</td>
<td>48.6</td>
<td>38.0</td>
</tr>
<tr>
<td>SEL (Trani, DocEng '16)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>n.a.</td>
<td>41.0</td>
<td>51.0</td>
<td>43.0</td>
</tr>
<tr>
<td>SWAT</td>
<td><strong>62.2</strong></td>
<td><strong>63.0</strong></td>
<td><strong>62.6</strong></td>
<td><strong>47.3</strong></td>
<td><strong>55.2</strong></td>
<td><strong>47.8</strong></td>
</tr>
</tbody>
</table>
## Experiments Results

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

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

Results

Independence from position of salient entities

Dataset

New York Times
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 infrastructure SoBigData
2. Entity Relatedness
Entity Relatedness

Motivation

Proliferation of the usage of Knowledge Graphs

- 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

$Relatedness : Entities \times Entities \rightarrow \text{Real}$
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: [leˈoːnardo dî ˈsɛr 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". 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". 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

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 dì ser 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]
Terminology

▷ Our Knowledge Graph (KG):

- **Entity** = Wikipedia Page (a node of KG)
- **Label** = Textual Description of the Wikipedia Page
- **Edges**?
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 sɛr ˈpjɛːro da (v)ˈvintʃi], 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, botany, astronomy, and cartography. He has been variously called the father of paleontology, 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, he epitomised the Renaissance humanist ideal.

Many historians and scholars regard Leonardo as a Genius or "Renaissance Man".

Invention

Entry 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 must be new and not obvious to others skilled in the same field. An inventor may be taking a big step forward in terms of invention. Some inventions can be patented. A patent legally protects the intellectual property rights of the claimed invention is actually an invention. The rules and requirements for patenting an invention process of obtaining a patent are often expensive.

Another meaning of invention is cultural invention, which is an innovative and useful device, method, composition or process. The invention must be new and not obvious to others skilled in the same field. An inventor may be taking a big step forward in terms of invention. Some inventions can be patented. A patent legally protects the intellectual property rights of the claimed invention is actually an invention. The rules and requirements for patenting an invention process of obtaining a patent are often expensive.

Astronomy

Entry from Wikipedia, the free encyclopedia

Astronomy, a natural science, is the study of celestial objects (such as stars, galaxies, planets, nebulae) and objects in space that are not on Earth. The Sun is a star and the Moon is a natural satellite of Earth. The study of the physical properties of celestial objects and phenomena, is one of the oldest sciences. Early civilizations recorded the positions and movements of celestial objects, and performed observational and mathematical calculations to predict their future positions and movements. The scientific field of astronomy is concerned with the study of celestial objects, such as stars, galaxies, planets, and comets.

Leonardo was, and is, renowned primarily as a painter. Among his works, the Mona Lisa and the Last Supper are the most famous and most parodied portrait paintings of all time, their fame approached only by Michelangelo's Creation of Adam. Leonardo's drawing of the Vitruvian Man is also regarded as a cultural icon.
The Wikipedia Knowledge Graph

- **Our Knowledge Graph (KG):**
  - **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

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

Tiger

Cat
Our Two-Stage Framework

A small and weighted subgraph is dynamically grown around the two query entities Tiger and Cat.

How can we populate the subgraph?
Our Two-Stage Framework

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

Populating the subgraph. Choosing the top-k nodes most related to the query entities.
Our Two-Stage Framework

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

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

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

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

\[
\text{CoSimRank (Rothe, ACL '14)}
\]

\[
\text{relatedness} (\text{tiger}, \text{cat}) = 0.65
\]
# Experiments

## Intrinsic Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>WikiSim</th>
<th>WiRe</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>Spearman</td>
<td>Harmonic</td>
</tr>
<tr>
<td>ESA</td>
<td>0.61</td>
<td>0.72</td>
<td>0.67</td>
</tr>
<tr>
<td>Milne-Witten</td>
<td>0.62</td>
<td>0.65</td>
<td>0.63</td>
</tr>
<tr>
<td>DeepWalk</td>
<td>0.71</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>Entity2Vec</td>
<td>0.68</td>
<td>0.70</td>
<td>0.69</td>
</tr>
<tr>
<td>Two-Stage Framework</td>
<td>0.74</td>
<td>0.75</td>
<td>0.74</td>
</tr>
</tbody>
</table>

- **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: 
  ○ 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 $\epsilon$-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)

<table>
<thead>
<tr>
<th></th>
<th>Uncompressed</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Space</strong></td>
<td>5 GB</td>
<td>445 MB</td>
</tr>
<tr>
<td><strong>Average Time</strong></td>
<td>0.5 ms</td>
<td>3 ms</td>
</tr>
</tbody>
</table>

▷ 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?