Shine

Linking Named Entities
in Web Text to a Knowledge Base
- An API -

Leili Tavabi and Luis Angel Larios C.

polysemy in web texts

…King [[*Cepheus*]] chained Andromeda to a rock in the sea where the hero saw her. [[*Perseus*]] was still wearing the winged sandals of Hermes that he had used in the task of carefully decapitating Medusa, while watching what he was doing only through a mirror. He asked what had happened to [[*Andromeda*]]; then, when he heard, he promptly offered to rescue her by killing the sea monster, but on condition that her parents give her to him in marriage.

possible mappings

Cepheus_(constellation)
Cepheus,_King_of_Aethiopia
Cepheus,_King_of_Tegea
Cepheus_(crater)
Cepheus_(Chinese_astronomy)

Perseus
Perseus_(constellation)
Perseus_of_Macedon
Bristol_Perseus
Perseus_(geometer)
Perseus_(Pantheon)

Andromeda_(constellation)
Andromeda_(TV_series)
Andromeda_(mythology)
Cepheus,_King_of_Tegea
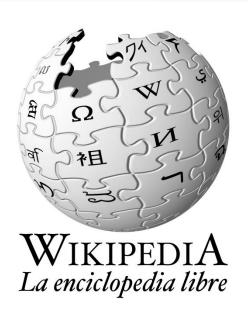Andromeda_(galaxy)
Andromeda_(polifolia)
Andromeda_(Marvel_Comics)

...King [[***Cepheus***]] chained Andromeda to a rock in the sea where the hero saw her. [[***Perseus***]] was still wearing the winged sandals of Hermes that he had used in the task of carefully decapitating Medusa, while watching what he was doing only through a mirror. He asked what had happened to [[***Andromeda***]]; then, when he heard, he promptly offered to rescue her by killing the sea monster, but on condition that her parents give her to him in marriage.

objective

To develop an API that exposes the functionality of linking surface forms in web texts to their corresponding entity mappings in Wikipedia.

WIKIPEDIA
La enciclopedia libre

previous work

- LIEGE: Linking web entities in web lists
  - Introduces the **Linking Quality Metric**, which includes prior probability, type hierarchy based similarity, and distributional context similarity.
- LINDEN: Linking named entities with knowledge base via semantic knowledge
  - Exploits Wikipedia Miner's **Wikify! Service** to detect context concepts inside web texts.
- KAURI: Linking named entities in Tweets
  - Proposes a graph approach that unveils the real mappings through the **User Interest Propagation Algorithm**.

our approach

- Named entities discovery
  - User's named entities may be complemented by the ***Wikify!*** Wikipedia Miner's service to enrich semantics.

- Candidate mappings generation
  - Use of a dictionary based on ***Wikipedia structure***.

- Named entities disambiguation
  - Solve for an optimal set of candidate mappings which maximizes a ***Linking Quality Metric***.

Shine

Input Text

$\downarrow$

Named Entities
$$[[\ m_1\ ]]...[[m_2]]...[[\ m_3\ ]]$$

$\downarrow$

Candidate Mappings

$$
\begin{array}{ccc}
r_{1,1} & r_{2,1} & r_{3,1} \\
r_{1,2} & r_{2,2} & r_{3,2} \\
\vdots & \vdots & \vdots \\
r_{1,|R_1|} & r_{2,|R_2|} & r_{3,|R_3|}
\end{array}
$$

$\downarrow$

$$e_1 \quad e_2 \quad e_3$$
Real Mappings

named entities discovery

Input Text

...King [[*Cepheus*]] chained Andromeda to a rock in the sea where the hero saw her. [[*Perseus*]] was still wearing the winged sandals of Hermes that he had used in the task of carefully decapitating Medusa, while watching what he was doing only through a mirror. He asked what had happened to [[*Andromeda*]]; then, when he heard, he promptly offered to rescue her by killing the sea monster, but on condition that her parents give her to him in marriage.
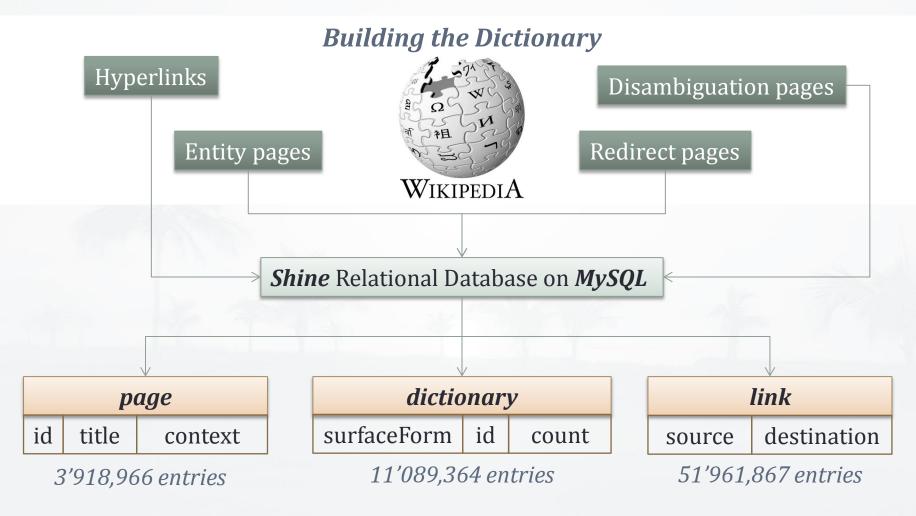
Wikify!

...King [[*Cepheus*]] chained Andromeda to a rock in the sea where the hero saw her. [[*Perseus*]] was still wearing the winged sandals of Hermes that he had used in the task of carefully [[*decapitating*]] [[*Medusa*]], while watching what he was doing only through a mirror. He asked what had happened to [[*Andromeda*]]; then, when he heard, he promptly offered to rescue her by killing the [[*sea monster*]], but on condition that her parents give her to him in marriage.

Named entities M

m_1	m_2	m_3	m_4	m_5	m_6
cepheus	perseus	andromeda	decapitating	sea monster	medusa

candidate mappings generation

Building the Dictionary

Hyperlinks

Entity pages

WIKIPEDIA

Disambiguation pages

Redirect pages

Shine Relational Database on *MySQL*

page		
id	title	context

3'918,966 entries

dictionary		
surfaceForm	id	count

11'089,364 entries

link	
source	destination

51'961,867 entries

named entities disambiguation

*The **quality** of a candidate $r_{i,j}$ for a named entity m_i is a convex combination of three metrics*

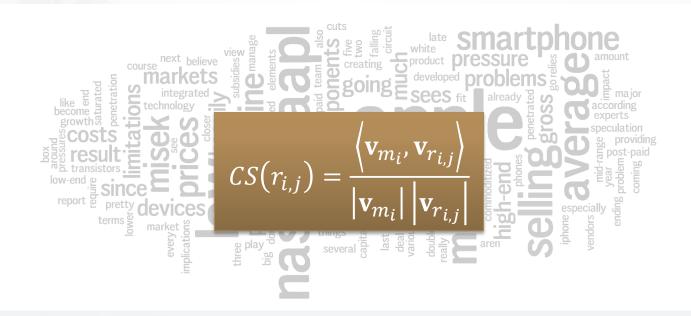Linking Probability   Context Similarity   Global Coherence

- Linking Probability
  – How popular is a candidate mapping $r_{i,j}$ for named entity m_i?

$$LP(r_{i,j}) = \frac{count(r_{i,j})}{\sum_{r_{i,c} \in R_i} count(r_{i,c})}$$

- Context Similarity
  - Extract a **window** of unique terms around each occurrence of m_i in the **purified** query document.
  - Extract **all** unique terms from the purified text of the Wikipedia article that corresponds to a candidate mapping $r_{i,j}$.
  - Use the vector model and **TF-IDF** to compute the cosine similarity between document vectors $\mathbf{v}_u$.

$$CS(r_{i,j}) = \frac{\left\langle \mathbf{v}_{m_i}, \mathbf{v}_{r_{i,j}} \right\rangle}{\left| \mathbf{v}_{m_i} \right| \left| \mathbf{v}_{r_{i,j}} \right|}$$

- Global Coherence
  - Two Wikipedia concepts are topically related if they **share** a non-empty set of concepts pointing to them.
  - Exploit Wikipedia inter-link structure via the **Wikipedia Linking Measure** between two concepts.

$$TR(u_1, u_2) = 1 - \frac{\log(\max(|U_1|, |U_2|)) - \log(|U_1 \cap U_2|)}{\log(|W|) - \log(\min(|U_1|, |U_2|))}$$

$$GC(r_{i,j}) = \frac{1}{|M| - 1} \sum_{c \neq i} TR(r_{i,j}, e_c)$$

- ## The Linking Quality Metric
  - Given the importance weights α, β, and γ, the quality of a candidate is

$$LQ(r_{i,j}) = \alpha * LP(r_{i,j}) + \beta * CS(r_{i,j}) + \gamma * GC(r_{i,j})$$

Overall Quality of Chosen Candidate Mappings E

$$LQ(E) = \sum_{i=1}^{|M|} LQ(e_i)$$

*The goal is to choose the **best combination** of candidate mappings so that $LQ(E)$ is maximal. We use the **Iterative Substitution Algorithm** for this purpose.*

results

Parsing Wikipedia Dump of June 4, 2013	*Storing dictionary and context*	*Shine disambiguation system and service*	*API tests and deployment*

*Shine is a blend of techniques described in **LIEGE**, **LINDEN**, and **KAURI***

- Demanding tasks:
  - Building the dictionary, page, and link tables consumed lots of space and time – **Prone to errors! Manual corrections!**
  - Each optimization of the system required **re-building these tables** several times.

optimization

- **Space reduction**
  - Use of integer-based primary keys for Wikipedia concepts.
  - Preprocessing and storing Wikipedia pages' term-frequency vectors instead of whole text.

- **Accelerate response time**
  - Indexing frequently used columns in dictionaries, page, and link tables.
  - Preprocessing Wikipedia text into term-frequency vectors stored in page table.
  - IDF computation based on current query.

the api

- RESTful
  - **useWikify**: *true* or ***false***. Specify if input web text will be further enriched with named entities discovered by Wikipedia Miner's Wikify! Service.

  - **source**: *Web text* of at least 128 characters, with named entities enclosed by [[and]].

Example

Disambiguate?**useWikify**=true&**source**=An oracle told the king that the only way to get rid of the sea monster was to surrender his virgin daughter Andromeda to the sea monster; so he did. King [[**Cepheus**]] chained Andromeda to a rock in the sea where the hero saw her. [[**Perseus**]] was still wearing the winged sandals of Hermes that he had used in the task of carefully decapitating Medusa, while watching what he was doing only through a mirror. He asked what had happened to [[**Andromeda**]]; then, when he heard, he promptly offered to rescue her by killing the sea monster, but on condition that her parents give her to him in marriage.

http://localhost/Shine/Disambiguate.php?XDEBUG_SESSION_START=ECLIPSE_DBGP&KEY=14009880944682

```xml
<?xml version="1.0" encoding="UTF-8"?>
<message service="Disambiguate" totalScore="2.0643587607016">
    <request>An oracle told the king that the only way to get rid of the [[sea monster]] was to surrender his [[virgin]]
        daughter Andromeda to the [[sea monster]]; so he did. King [[Cepheus]] chained Andromeda to a rock in the sea where
        the hero saw her. [[Perseus]] was still wearing the winged sandals of Hermes that he had used in the task of carefully
        [[decapitating]] [[Medusa]], while watching what he was doing only through a mirror. He asked what had happened to
        [[Andromeda]]; then, when he heard, he promptly offered to rescue her by killing the [[sea monster]], but on condition
        that her parents give her to him in marriage.</request>
    <errors/>
    <warnings/>
    <mappings>
        <mapping wikiID="261293" score="0.31886742665469" surfaceForm="sea monster">Sea_monster</mapping>
        <mapping wikiID="89775" score="0.15108482584396" surfaceForm="virgin">Virginity</mapping>
        <mapping wikiID="3676637" score="0.26556217067654" surfaceForm="cepheus">Cepheus,_King_of_Aethiopia</mapping>
        <mapping wikiID="209446" score="0.37894023019512" surfaceForm="perseus">Perseus</mapping>
        <mapping wikiID="70802" score="0.22685913512266" surfaceForm="decapitating">Decapitation</mapping>
        <mapping wikiID="392192" score="0.36844465642236" surfaceForm="medusa">Medusa</mapping>
        <mapping wikiID="1925" score="0.3546003157863" surfaceForm="andromeda">Andromeda_(mythology)</mapping>
    </mappings>
</message>
```

XML output generated by Shine

conclusion

- Shine
  - Combines **Natural Language Processing** and **Big Data** techniques.
  - Is based on state-of-the-art deterministic methods that have achieved up to **95% accuracy** in the disambiguation task.
  - Can be the foundation of a search engine that captures **semantics** rather than just keywords in user queries.

샤인

thank you!
any questions?