SEMANTIC GRAPHS DERIVED FROM TRIPPLETS WITH APPLICATION IN DOCUMENT SUMMARIZATION

Delia Rusu, Blaž Fortuna, Dunja Mladenić, Marko Grobelnik
Jožef Stefan Institute
Department of Knowledge Technologies

October 17, 2008
Overview

- Introduction
- System Pipeline
- System Evaluation
- Conclusions
Overview

- Introduction
- System Pipeline
- System Evaluation
- Conclusions
Introduction

- Given a document, perform the following:
  - Annotate it with **named entities** (people, places, organizations)
  - Extract **(subject, verb, object)** – triplets
  - Refine the triplets and obtain a graphical representation of the document – a **semantic graph**
  - **Automatically summarize the document** based on the semantic graph
Overview

- Introduction

- System Pipeline
  - Named Entities and Triplets
  - Semantic Graphs
  - Document Summarization

- System Evaluation

- Conclusions
System Pipeline

Document(s)
**Berlin Hotel Unsure How to Deal with Raccoon Guest**

A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser. For North Americans, having a raccoon in your garage is hardly a big deal. In Germany, however -- where the animals are something of a rarity -- the same can't be said.

For the past few weeks, a raccoon has been causing a stir in Germany's capital Berlin after making its home in the garage of one of the city's landmark hotels, the 39-story Park Inn. The furry creature has been spotted scurrying around the hotel's garage and scavenging for food in the trash cans of a near-by Burger King fast food restaurant. Hotel staff have christened the little guest Alex, after the nickname for Alexanderplatz, the square where the hotel is located.

Over the last few weeks, the little animal has become a minor celebrity after Berlin's media got wind of the raccoon's unusual choice of home. The city's newspapers have been reporting about the latest sightings of Alex and the hotel's failed attempt to find it a new home. Although welcoming their unusual guest, the hotel's management seems ambivalent about having a raccoon living in its garage.

When the news first broke, the hotel's general manager Thomas Hattenberger told reporters they were proud the animal had chosen their hotel as its new home. However, at the same time he announced they were looking for a new home -- one more suitable for a raccoon -- for Alex. The hotel even offered to sponsor the animal, if a zoo or a private individual could be found to take Alex in.

But, after seeking the help of wildlife experts at Berlin's Senate Department for Urban Development, the hotel's management discovered it might not be allowed to evict Alex. According to German hunting law, a wild animal cannot be removed from private property -- even from one's own -- unless it presents a danger.

But the raccoon has not yet caused any trouble. "At the moment we are just waiting to see what will happen," hotel spokeswoman Catherina Cora told SPIEGEL ONLINE Monday. "The raccoon is not a danger to anyone and so far it has created no harm." She added that the hotel's only concern was that, because the animal was quite unusual, it might frighten a driver and cause an accident.

To check up on the raccoon's health, Derk Ehler, expert for wild animals at Berlin's Senate Department for Urban Development, paid the animal a visit on Sunday night. Ehler suspects Alex could have grown up at Berlin's most famous tourist attraction. He told newspaper Berliner Morgenpost that a raccoon family living at the Brandenburg Gate had been missing two young raccoons for a while.

Raccoons are native to North America, but have become more widespread in Europe since the beginning of the 20th century after escaping from fur farms. The Berlin raccoon population, which experts estimate to comprise around 100 or 120 family groups, derives from a farm in the town of Strausberg, 30 kilometers east of Berlin. Raccoons are believed to have escaped from the farm at the end of World War II and ended up settling in the German capital.

Although some of them, like Alex, have made their home in buildings -- such as a shopping center in the Berlin district of Spandau -- most have sought to live near rivers, canals and lakes.
Berlin Hotel Unsure How to Deal with Raccoon Guest

A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser. For North Americans, having a raccoon in your garage is hardly a big deal. In Germany, however -- where the animals are something of a rarity -- the same can't be said.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
System Pipeline

Named Entities

Berlin Hotel

Unsure How to Deal with Raccoon Guest. A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser. For North Americans, having a raccoon in your garage is hardly a big deal. In Germany, however -- where the animals are something of a rarity -- the same can’t be said. For the past few weeks, a raccoon has been causing a stir in Germany’s capital Berlin after making its home in the garage of one of the city’s landmark hotels, the 39-story Park Inn. The furry creature has been spotted scurrying around the hotel’s garage and scavenging for food in the trash cans of a near-by Burger King fast food restaurant. Hotel staff have christened the little guest Alex, after the nickname for Alexanderplatz, the square where the hotel is located. Over the last few weeks, the

legend

Named entities have the following color code:

Person, highlighting the person name.

Location, highlighting the location name.

Organization, highlighting the organization name.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
System Pipeline

Triplets (subject, verb, object)

Berlin Hotel Unsure How to Deal with Raccoon Guest. A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser. For North Americans, having a raccoon in your garage is hardly a big deal. In Germany, however -- where the animals are something of a rarity -- the same can't be said. For the past few weeks, a raccoon has been causing a stir in Germany's capital Berlin after making its home in the garage of one of the city's landmark hotels, the 39-story Park Inn. The furry creature has been spotted scurrying around the hotel's garage and scavenging for food in the trash cans of a nearby Burger King fast food restaurant. Hotel staff have christened the little guest Alex, after the nickname for Alexanderplatz, the square

**Legend**

Triplets have the following color code:
- **Subject**, highlighting the subject triplet element.
- **Verb**, highlighting the verb triplet element.
- **Object**, highlighting the object triplet element.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
When the news first broke, the hotel’s general manager Thomas Hattenberger told reporters they were proud the animal had chosen their hotel as its new home. However, at the same time he announced they were looking for a new home -- one more suitable for a raccoon -- for Alex. To check up on the raccoon’s health, Derk Ehlert, expert for wild animals at Berlin’s Senate Department for Urban Development, paid the animal a visit on Sunday night. Ehlert suspects Alex could have grown up at Berlin’s most famous tourist attraction. He told newspaper Berliner Morgenpost that a raccoon family living at the Brandenburg Gate had been missing two young raccoons for a while.
A Berlin hotel is trying to figure out how to deal with an unusual guest -- a raccoon who has moved into its garage. According to German law, the hotel does not have the right to evict the furry trespasser.
Named Entities and Triplets

- **Named Entities:**
  - Names of people, locations, organizations
  - Manually annotated (e.g., REUTERS data set) and/or automatically extracted (e.g., using **GATE** - General Architecture for Text Engineering http://gate.ac.uk/)

- **Triplets:**
  - Extracted with **Stanford Parser**
    http://nlp.stanford.edu/software/lex-parser.shtml
Semantic Graphs

Co-reference
Semantic Graphs

Co-reference

Arrigo Sacchi  Sacchi  Mr. Sacchi  =>  Arrigo Sacchi
Semantic Graphs

- Co-reference
- Anaphora
Gilles Garin, chairman of DHL France, said that the express distribution group still hoped to build the hub at Strasbourg. “As long as they have not said clearly 'no', we are still interested,” he told Reuters by telephone.
Semantic Graphs

- Co-reference
- Anaphora
- Semantic Normalization
Semantic Graphs

Using WordNet synsets:

- sediments
- Tokyo Bay – built – deposits
Semantic Graphs

Co-reference  Anaphora  Semantic Normalization  Semantic Graph
Semantic Graphs

Co-reference  Anaphora  Semantic Normalization  Semantic Graph
HOLY JETPACKS, FUSION MAN!
Swiss Man Flies With Own Wings.
It’s a blast, it’s a blast ... it’s a blast! A Swiss guy strapped four jet engines to his back and flew above the Alps. A dedicated inventor has set a new record for powered personal flight.
The Swissman, dressed in a white flight suit, went up in a plane above the Swiss town of Beromünster and took flight using a jetpack he created.
The Swiss man is a former Swiss military pilot and has been a co-pilot and captain for Switzerland’s two national airlines, Swissair and Swissair.
Dressed in a white flight suit, wearing a white helmet and strapped to his back, the pilot jumped off an airplane 3,349 meters above the Earth. He first unfolded the jetpack, eight-foot wings strapped to his back, then fired up four small jet engines originally intended to power model aircraft.
A helicopter flew nearby to document his five minutes of glory, and an airplane followed to measure his speed. Rossi reached speeds of 3,868 kilometers per hour (2,388 miles per hour).
"The flight was excellent," Rossi told reporters gathered at the airport where he touched down. Rossi wears a heavy, heat-resistant flying suit, similar to those worn by race car drivers and firefighters, to protect himself from the jet engine’s heat.
Should anything go wrong, Rossi has a parachute to rely on. "I’ve had many ‘whoops’ moments," he said Wednesday. "My safety is not at risk."
Rossi has more flights planned for the future — including a 22-mile crossing of the English Channel that may be broadcast live on television. He plans to fly through the Grand Canyon some day.
Jetpack hopefuls, take heart: Rossi plans to market his invention to the public in the future. It won’t come cheap, Rossi and his sponsors say they’ve poured almost €200,000 (330,000) into the project so far.
Document Summarization
Document Summarization

Feature Extractor
Document Summarization

• Linguistic Attributes
• Document Attributes
• Graph Attributes
Document Summarization

- Linguistic Attributes
- Document Attributes
- Graph Attributes
Document Summarization

- Linguistic Attributes
- Document Attributes
- Graph Attributes

Feature Extractor

Linear SVM

• Linear Model

Modeling in programming languages like Java or Python enhances efficiency and accuracy.
Rossy and his sponsors say they poured almost 200,000 (300,000) into the project so far. [12.4989]

Rossy has more flights planned for the future -- including a 22-mile crossing of the English Channel that may be broadcast live on television. [7.50326]

...
Document Summarization

Feature Extractor → Linear SVM

- Linguistic Attributes
- Document Attributes
- Graph Attributes

Linear Model

Rossy and his sponsors say they poured almost 200,000 (300,000) into the project so far. [12.4989]

Rossy has more flights planned for the future -- including a 22-mile crossing of the English Channel that may be broadcast live on television. [7.50326]
# Document Summarization

## EXTRACTED FEATURES

<table>
<thead>
<tr>
<th>Linguistic Attributes (13)</th>
<th>Document Attributes (11)</th>
<th>Graph Attributes (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Logical form tag</td>
<td>• Sentence related: e.g. – location of sentence within doc</td>
<td>• Authority and Hub weight, Page Rank</td>
</tr>
<tr>
<td>• Treebank tag</td>
<td>• Triplet related: e.g. - frequency of triplet elem in sentence, in doc, …</td>
<td>• Node degree</td>
</tr>
<tr>
<td>• Part of speech tag</td>
<td></td>
<td>• Size of weakly connected component</td>
</tr>
<tr>
<td>• Depth of linguistic node</td>
<td></td>
<td>• Size of max length chain</td>
</tr>
<tr>
<td>• 8 semantic tags for named entities</td>
<td></td>
<td>• Frequency of verbs among edges</td>
</tr>
</tbody>
</table>
# Document Summarization

## Features Ranking (Weka)

- Object - Word
- Subject - Word
- Verb - Word
- Location Of Sentence In Document
- Similarity With Centroid
- Number Of Locations In Sentence
- Number Of Named Entities In Sentence
- Authority Weight Object
- Hub Weight Subject
- Size Weakly Conn Comp Object
Some of Mr. Thomas’s opponents have warned that they intend to make him answer very explicit questions about his personal and philosophical beliefs, in order to avoid repeating the example of David Souter, "the stealth candidate."

But if Mr. Thomas does refuse to answer explicit questions, it will not be merely Mr. Souter’s example he will be following -- it will be that of the justice he has been nominated to replace, Thurgood Marshall.

Thurgood Marshall’s example makes it plain, though, that Mr. Thomas may refuse to give that categorical answer. During the Marshall hearings in 1967, the issue of the day was not abortion but crime, and the most controversial case was Miranda v. Arizona.

But such personal experiences are irrelevant to the work of a justice of the Supreme Court, according to Thurgood Marshall.

"A judge is not permitted to use his own personal views under any circumstances," he said, "but if the particular decision-maker, whether he be judge or justice or magistrate, applies the Constitution to the facts and it comes out that way, that is inevitable."

Sen. Kennedy stressed a central point both in the committee room and before the full Senate: The Senate is to assess qualifications; it is not to examine political or judicial philosophy.

With hearings scheduled to begin today, the nomination of Judge Clarence Thomas to the U.S. Supreme Court has again raised questions concerning the responsibilities of senators and nominees during the confirmation process.

Some of Mr. Thomas’s opponents have warned that they intend to make him answer very explicit questions about his personal and philosophical beliefs, in order to avoid repeating the example of David Souter, "the stealth candidate."

But if Mr. Thomas does refuse to answer explicit questions, it will not be merely Mr. Souter’s example he will be following -- it will be that of the justice he has been nominated to replace, Thurgood Marshall.

During the Marshall hearings in 1967, the issue of the day was not abortion but crime, and the most controversial case was Miranda v.

But such personal experiences are irrelevant to the work of a justice of the Supreme Court, according to Thurgood Marshall.

Mr. Thomas, too, has been previously approved by the committee.

As the second black American to serve on the Supreme Court, Judge Thomas will shoulder many responsibilities; not because he is black, but because he will hold a lofty office.
Overview

- Introduction
- System Pipeline
- System Evaluation
  - Gender Retrieval
  - Co-reference Resolution
  - Anaphora Resolution
  - Summary Generation
- Conclusions
### Gender Retrieval Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Masculine</th>
<th>Feminine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System</strong></td>
<td>170/206 (83%)</td>
<td>7/14 (50%)</td>
<td>177/220 (80%)</td>
</tr>
<tr>
<td><strong>Baseline</strong></td>
<td>206/206 (100%)</td>
<td>0/14 (0%)</td>
<td>206/220 (94%)</td>
</tr>
</tbody>
</table>

- 15 random documents taken from Reuters RCV1 data set were manually annotated

**System** – the system we implemented  
**Baseline** – assigns the most frequent gender (masculine) to all named entities labeled as persons
Co-reference Resolution Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Co-References</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
<td>750/783 (96%)</td>
</tr>
<tr>
<td>GATE</td>
<td>646/783 (83%)</td>
</tr>
</tbody>
</table>

• 15 random documents taken from Reuters RCV1 data set were manually annotated

*System* – the system we implemented

*GATE* – the co-reference resolver implemented within the GATE system
Anaphora Resolution Evaluation

- 77 random documents taken from Reuters RCV1 data set were manually annotated
- **System** – the system we implemented
- **Baseline Gender** – takes the closest named entity as a pronoun replacement, having gender information into account
- **Baseline No Gender** – takes the closest named entity as a pronoun replacement, but no gender information is considered
## Anaphora Resolution Evaluation

<table>
<thead>
<tr>
<th>Pronouns</th>
<th>System</th>
<th>Baseline-gender</th>
<th>Baseline-no gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>He</td>
<td>35/42 (83%)</td>
<td>18/42 (43%)</td>
<td>18/42 (43%)</td>
</tr>
<tr>
<td>They</td>
<td>7/20 (35%)</td>
<td>8/20 (40%)</td>
<td>2/20 (10%)</td>
</tr>
<tr>
<td>I</td>
<td>4/15 (27%)</td>
<td>0/15 (0%)</td>
<td>2/15 (13%)</td>
</tr>
<tr>
<td>She</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>Who</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>It</td>
<td>11/35 (31%)</td>
<td>11/35 (31%)</td>
<td>11/35 (31%)</td>
</tr>
<tr>
<td>Other</td>
<td>2/4 (50%)</td>
<td>2/6 (33%)</td>
<td>3/6 (50%)</td>
</tr>
<tr>
<td>Total</td>
<td>59/116 (51%)</td>
<td>39/118 (33%)</td>
<td>36/118 (31%)</td>
</tr>
</tbody>
</table>
Summary Generation Evaluation

- Document Understanding Conference (DUC) 2002 and 2007 data sets

- **DUC 2002 task**
  - automatic summarization of a single newswire/newspaper article

- **DUC 2007 task**
  - update task – produce short (~100 words) multi-document update summaries of newswire articles
  - 10 topics (A-J), each divided in 3 clusters (A-C); each cluster with 7-10 articles
  - Our aim:
    - A summary of documents in cluster A
  - Training data: DUC 2002 & DUC 2007 main task
Summary Generation Evaluation

DUC 2002 Task

<table>
<thead>
<tr>
<th>Our System Results</th>
<th>J. Leskovec et al.’s system results</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Documents</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>31.25%</td>
<td>71.49%</td>
<td>43.43%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Documents</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>26.25%</td>
<td>73.22%</td>
<td>38.64%</td>
</tr>
</tbody>
</table>

Ten-fold cross validation results comparison between the two systems:
- the system proposed in the paper
- J. Leskovec et al.’s system presented at LinkKDD 2004 (Workshop on Link Analysis and Group Detection)
Summary Generation Evaluation

DUC 2007 Update Task

ROUGE 2 Evaluation Results

- Best: Language Computer Corp.
- Microsoft Research
- University of Maryland
- Columbia University
- Our System
- Universität Karlsruhe
- Worst

Average Recall
Summary Generation Evaluation

DUC 2007 Update Task

ROUGE SU4 Evaluation Results

- **Best**: Language Computer Corp.
- University of Maryland
- Microsoft Research
- Columbia University
- **Our System**
- Universität Karlsruhe
- **Worst**

- Average Recall
Conclusions

- **Semantic graph** representation of documents
- **Automatic document summary** obtained by sentence extraction
- Future work
  - system integration
    - online newswire crawler
    - complex ontology
Thank you for your attention!
Applying Co-reference Resolution

-> Graph with 46 nodes, 35 edges
Applying Anaphora Resolution

-> Graph with 43 nodes, 35 edges
Applying Semantic Normalization

-> Graph with 39 nodes, 33 edges