Automatic Single-Document Key Fact Extraction from Newswire Articles

Itamar Kastner  Christof Monz

Department of Computer Science
Queen Mary, University of London
itkl@dcs.qmul.ac.uk

ISLA, University of Amsterdam
Amsterdam, The Netherlands
c.monz@uva.nl

EACL 2009
The Task: CNN.com Story Highlights

PimpThisBum.com employs irony on homeless man's behalf

STORY HIGHLIGHTS
- Father-son team launch PimpThisBum.com as stunt to help homeless man
- Homeless man, Timothy Edwards, has received more than $50,000 in donations
- Edwards: "To everybody that thinks I am being exploited, I ask you to think again"
- "This isn't a good strategy to address the problem of homelessness," critic says

Next Article in U.S. »

Thirty years later, Sadat's widow still hopes for peace

STORY HIGHLIGHTS
- Thirty years after the Camp David accords, Anwar Sadat's widow hopes for peace
- Jehan Sadat's husband was assassinated two years after the treaty signing
- She says Islam is a religion of peace and the people of the Mideast want peace

Next Article in World »
STORY HIGHLIGHTS

- Father-son team launch PimpThisBum.com as stunt to help homeless man
- Homeless man, Timothy Edwards, has received more than $50,000 in donations
- Edwards: "To everybody that thinks I am being exploited, I ask you to think again"
- "This isn't a good strategy to address the problem of homelessness," critic says

Next Article in U.S. »

STORY HIGHLIGHTS

- Thirty years after the Camp David accords, Anwar Sadat's widow hopes for peace
- Jehan Sadat's husband was assassinated two years after the treaty signing
- She says Islam is a religion of peace and the people of the Mideast want peace

Next Article in World »
Americans celebrate, honor the fallen on Memorial Day

POSTED: 6:09 p.m. EDT, May 28, 2007

NEW YORK (AP) - Veterans and active soldiers unfurled a 90-by-100-foot U.S. flag as the nation's top commander in the Middle East spoke to a Memorial Day crowd gathered in Central Park on Monday.

Navy Adm. William Fallon, commander of U.S. Central Command, said America should remember those whom the holiday honors.

"Their sacrifice has enabled us to enjoy the things that we do," he said in many cases, take for granted." Fallon said.

Across the nation, flags snapped in the wind over decorated gravestones as relatives and friends paid tribute to their fallen soldiers. (Watch the sights and sounds of Memorial Day weekend)

Millions more kicked off summer with trips to beaches or their backyard grills.

In the nation's capital, thousands of motorcycles driven by military veterans and their loved ones roared through Washington to the Vietnam Veterans Memorial. It was the 20th year the group, Rolling Thunder, has taken to the streets of the nation's capital in support of U.S. military members past and present.

President Bush spoke at nearby Arlington National Cemetery, honoring U.S. troops who have fought and died for freedom and expressing his resolve to succeed in the war in Iraq. (Watch President Bush pay tribute to the troops)

"From their deaths must come a world where the cruel dreams of tyrants and terrorists are frustrated and foiled -- where our nation is more secure from attack, and where the gift of liberty is secured for millions who have never known it," the president said.

Helen Velasquez stood over her husband's grave in Farmington, New Mexico, and said he never second guessed his decision to enlist in the Marines. Frank Velasquez, 47, was disabled and died in March.

"I think a lot of people don't think of it until someone they love passes away," she said. "Sometimes I feel they're forgotten. I want them to know they're not forgotten."

A Memorial Day tribute at Brooklyn's Green-Wood Cemetery was marked with 1,200 new gravestones for Civil War veterans whose graves were recently rediscovered after a nearly five-year search. The cemetery, founded in 1836, is the burial place of nearly 800,000 people, including veterans from every American war, said Richard J. Moylan, its president.

At Fort Bragg, North Carolina, the commanding general of the Army's Green Berets said the troops who have died around the world were "soldiers we could count on when the going got tough." Thirteen Green Berets who died during the past year were honored as were 100 former Green Berets who died during the period. (Watch soldiers hold ceremonies for lost comrades in Iraq)

"We always could count on our Special Forces brothers to be there with us, regardless of the odds against us," Maj. Gen. Thomas R. Osmo said at a ceremony at the headquarters of the Army's Special Forces Command.

Elsewhere, Alabama's Gulf Coast was once again packed with holiday-goers after the damage from hurricanes Ivan and Katrina in 2004 and 2005 kept the tourists away.

"It couldn't have been better -- use all the superlatives you want," said Johnny Fisher, general manager of Lucy Buffett's Lulu's at Homeport Marina in Gulf Shores. "It was all perfect"
The Task: CNN.com Story Highlights

- The Story Highlights are written by a human editor.
- Encapsulate the most important pieces of knowledge in the article.
- Culled from various points in the text.
- How can we create these Highlights automatically?
  - **AURUM**: Automatic Retrieval of Unique information with Machine learning.
The Story Highlights are written by a human editor.
Encapsulate the most important pieces of knowledge in the article.
Culled from various points in the text.
How can we create these Highlights automatically?
**AURUM**: Automatic Retrieval of Unique information with Machine learning.
Previous Work

Svore et al. (2007)

- Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources
- Encouraging ROUGE results: +6% for unigram recall, +3% for bigram recall
- Neural networks used for ranking
- Extract exactly 3 Highlights from every article
- 2nd experiment: try to match the order as well

External Data

- Microsoft Live News search logs
- Wikipedia articles
Previous Work

Svore et al. (2007)

- Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources
- Encouraging ROUGE results: +6% for unigram recall, +3% for bigram recall
- Neural networks used for ranking
- Extract exactly 3 Highlights from every article
- 2nd experiment: try to match the order as well

External Data

- Microsoft Live News search logs
- Wikipedia articles
AURUM

Determine Features
1. Inspect the training set and see what the characteristics are
2. Use additional features already existing in summarization and IR
3. Infer other features statistically
4. Assign a weight to each feature using machine learning

Score Sentences
1. For a given article, score each sentence against the features
   \[ \sigma(s) = w_{pos} p_{pos(s)} + \sum_{k=1}^{n} w_k f_k + \sum_{j=1}^{\lfloor s \rfloor} \sum_{k=1}^{m} w_k g_{jk} \]
2. Rank sentences by score
3. Use top \( n \) sentences as Highlights
4. Use first \( n \) sentences as baseline
## Determine Features

1. Inspect the training set and see what the characteristics are
2. Use additional features already existing in summarization and IR
3. Infer other features statistically
4. Assign a weight to each feature using machine learning

## Score Sentences

1. For a given article, score each sentence against the features
   \[
   \sigma(s) = w_{pos} p_{pos}(s) + \sum_{k=1}^{n} w_k f_k + \sum_{j=1}^{s} \sum_{k=1}^{m} w_k g_{jk}
   \]
2. Rank sentences by score
3. Use top \( n \) sentences as Highlights
4. Use first \( n \) sentences as baseline
Sentence Features Evaluated per Sentence

**Sentence Position**

Sentence position score

\[ p_i = 1 - (\log i/\log N) \]

- \( i \) – position of the sentence in the article
- \( N \) – total number of sentences in the article

**Temporal Adverbs**

- Simple temporal descriptors, the kind used in previous summarization work: *Thursday*
- Longer phrases which are typical to the training data: *after less than, for two weeks*
Sentence Features Evaluated per Sentence

Sentence Position

Sentence position score

\[ p_i = 1 - \left( \log \frac{i}{\log N} \right) \]

- \( i \) – position of the sentence in the article
- \( N \) – total number of sentences in the article

Temporal Adverbs

- Simple temporal descriptors, the kind used in previous summarization work: Thursday
- Longer phrases which are typical to the training data: after less than, for two weeks
Proper Nouns

Mention of a person or organisation as an indicator of important or new information (using Charniak’s parser)

Verb Groups

- Types of verbs which appear often in the website’s journalistic writing
- TalkVerbs: report, mention, accuse
- ActionVerbs: provoke, spend, use
Proper Nouns

Mention of a person or organisation as an indicator of important or new information (using Charniak’s parser)

Verb Groups

- Types of verbs which appear often in the website’s journalistic writing
  - TalkVerbs: report, mention, accuse
  - ActionVerbs: provoke, spend, use
n-gram Features

n-gram Frequency in Article and Highlights

Frequency and probability of unigrams, bigrams and trigrams appearing in both the article body and the highlights of a given story.

Trigger Phrases and Spawned Phrases

- **Trigger Phrases**: cause adjacent words to appear in the Story Highlights, e.g. the word according triggered neighbouring words in 1 out of every 4 appearances.

- **Spawned Phrases**: appear in the Highlights and next to Trigger Phrases.
  - **Highlight (spawned)**: 61 PERCENT OF THOSE POLLED NOW SAY IT WAS NOT WORTH INVADING IRAQ, POLL SAYS
  - **Source (trigger)**: NOW, 61 PERCENT OF THOSE SURVEYED SAY IT WAS NOT WORTH INVADING IRAQ, ACCORDING TO THE POLL.

- Further augmented with WordNet by using the synsets.
n-gram Features

**n-gram Frequency in Article and Highlights**
Frequency and probability of unigrams, bigrams, and trigrams appearing in both the article body and the highlights of a given story.

**Trigger Phrases and Spawned Phrases**
- **Trigger Phrases**: cause adjacent words to appear in the Story Highlights, e.g. the word **according** triggered neighbouring words in 1 out of every 4 appearances.
- **Spawned Phrases**: appear in the Highlights and next to Trigger Phrases.
  - **Highlight (spawned)**: 61 PERCENT OF THOSE POLLED NOW SAY IT WAS NOT WORTH INVAIDING IRAQ, POLL SAYS.
  - **Source (trigger)**: NOW, 61 PERCENT OF THOSE SURVEYED SAY IT WAS NOT WORTH INVAIDING IRAQ, ACCORDING TO THE POLL.
- Further augmented with WordNet by using the synsets.
n-gram Features

**n-gram Frequency in Article and Highlights**

Frequency and probability of unigrams, bigrams and trigrams appearing in both the article body and the highlights of a given story.

**Trigger Phrases and Spawned Phrases**

- **Trigger Phrases**: cause adjacent words to appear in the Story Highlights, e.g. the word according triggered neighbouring words in 1 out of every 4 appearances
- **Spawned Phrases**: appear in the Highlights and next to Trigger Phrases
  - **Highlight (spawned)**: 61 PERCENT OF THOSE POLLED NOW SAY IT WAS NOT WORTH INVADING IRAQ, POLL SAYS
  - **Source (trigger)**: NOW, 61 PERCENT OF THOSE SURVEYED SAY IT WAS NOT WORTH INVADING IRAQ, ACCORDING TO THE POLL.

Further augmented with WordNet by using the synsets.
**n-gram Frequency in Article and Highlights**

Frequency and probability of unigrams, bigrams and trigrams appearing in both the article body **and** the highlights of a given story.

**Trigger Phrases and Spawned Phrases**

- **Trigger Phrases**: cause adjacent words to appear in the Story Highlights, e.g. the word *according* triggered neighbouring words in 1 out of every 4 appearances.
- **Spawned Phrases**: appear in the Highlights **and** next to Trigger Phrases.
  - **Highlight (spawned)**: 61 PERCENT OF THOSE POLLED NOW SAY IT WAS NOT WORTH *INVADING IRAQ*, *POLL* SAYS.
  - **Source (trigger)**: NOW, 61 PERCENT OF THOSE SURVEYED SAY IT WAS NOT WORTH *INVADING IRAQ*, *ACCORDING* TO THE *POLL*.

Further augmented with WordNet by using the synsets.
n-gram Frequency in Article and Highlights

Frequency and probability of unigrams, bigrams and trigrams appearing in both the article body and the highlights of a given story.

Trigger Phrases and Spawned Phrases

- **Trigger Phrases**: cause adjacent words to appear in the Story Highlights, e.g. the word *according* triggered neighbouring words in 1 out of every 4 appearances.
- **Spawned Phrases**: appear in the Highlights and next to Trigger Phrases.
  - **Highlight (spawned)**: 61 PERCENT OF THOSE POLLED NOW SAY IT WAS NOT WORTH INVADING IRAQ, POLL SAYS
  - **Source (trigger)**: NOW, 61 PERCENT OF THOSE SURVEYED SAY IT WAS NOT WORTH INVADING IRAQ, ACCORDING TO THE POLL.

Further augmented with WordNet by using the synsets.
## Training Set and Development/Test Set

<table>
<thead>
<tr>
<th>Corpus subset</th>
<th>Dev/Test</th>
<th>Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>Documents</td>
<td>300</td>
<td>1220</td>
</tr>
<tr>
<td>Avg. sentences per article</td>
<td>33.26</td>
<td>31.02</td>
</tr>
<tr>
<td>Avg. sentence length</td>
<td>20.62</td>
<td>20.50</td>
</tr>
<tr>
<td>Avg. number of highlights</td>
<td>3.71</td>
<td>3.67</td>
</tr>
<tr>
<td>Avg. number of highlight sources</td>
<td>4.32</td>
<td>-</td>
</tr>
<tr>
<td>Avg. highlight length in words</td>
<td>10.26</td>
<td>10.28</td>
</tr>
</tbody>
</table>

Development/Test set used five-fold cross-validation due to small number of manually annotated articles.
Experimental Setup

Machine Learning of Feature Weights

1. YASMET, a maximum entropy classifier, was used to automatically learn feature importance.
2. Highlight Sources were manually annotated in 300 articles (the development/test set).
3. In each of the five folds, YASMET determined weights for the different features anew.
4. The weights were used for the test set of the relevant fold.
## Results: Recall, Precision, F-Measure

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-fixed</td>
<td>40.69</td>
<td>44.14</td>
<td>42.35</td>
</tr>
<tr>
<td>AURUM-fixed</td>
<td>41.88 (+2.96%*)</td>
<td>45.40 (+2.85%)</td>
<td>43.57 (+2.88%*)</td>
</tr>
<tr>
<td>Baseline-thresh</td>
<td>42.91</td>
<td>41.82</td>
<td>42.36</td>
</tr>
<tr>
<td>AURUM-thresh</td>
<td>44.49 (+3.73%*)</td>
<td>43.30 (+3.53%)</td>
<td>43.88 (+3.59%*)</td>
</tr>
</tbody>
</table>

- **AURUM-fixed**: always returns 4 sentences
- **AURUM-thresh**: returns between 3 and 6 sentences according to a threshold learned automatically from the development data
  - Anywhere between 1 and 7 sentences as Highlight Sources in an article – usually between 3 and 6
  - Therefore, impossible to know in advance how many Highlight Sources needed
  - Threshold learned from the development set and then used on the test set
- **Baseline**: always returns the same number of sentences as AURUM
## Results: Recall, Precision, F-Measure

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-fixed</td>
<td>40.69</td>
<td>44.14</td>
<td>42.35</td>
</tr>
<tr>
<td>AURUM-fixed</td>
<td>41.88 (+2.96%*)</td>
<td>45.40 (+2.85%)</td>
<td>43.57 (+2.88%*)</td>
</tr>
<tr>
<td>Baseline-thresh</td>
<td>42.91</td>
<td>41.82</td>
<td>42.36</td>
</tr>
<tr>
<td>AURUM-thresh</td>
<td>44.49 (+3.73%*)</td>
<td>43.30 (+3.53%)</td>
<td>43.88 (+3.59%*)</td>
</tr>
</tbody>
</table>

- **AURUM-fixed**: always returns 4 sentences
- **AURUM-thresh**: returns between 3 and 6 sentences according to a threshold learned automatically from the development data
  - Anywhere between 1 and 7 sentences as Highlight Sources in an article – usually between 3 and 6
  - Therefore, impossible to know in advance how many Highlight Sources needed
  - Threshold learned from the development set and then used on the test set
- **Baseline**: always returns the same number of sentences as AURUM
## Results: Recall, Precision, F-Measure

<table>
<thead>
<tr>
<th>System</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-fixed</td>
<td>40.69</td>
<td>44.14</td>
<td>42.35</td>
</tr>
<tr>
<td>AURUM-fixed</td>
<td>41.88 (+2.96%*)</td>
<td>45.40 (+2.85%)</td>
<td>43.57 (+2.88%*)</td>
</tr>
<tr>
<td>Baseline-thresh</td>
<td>42.91</td>
<td>41.82</td>
<td>42.36</td>
</tr>
<tr>
<td>AURUM-thresh</td>
<td>44.49 (+3.73%*)</td>
<td>43.30 (+3.53%)</td>
<td>43.88 (+3.59%*)</td>
</tr>
</tbody>
</table>

- **AURUM-fixed**: always returns 4 sentences
- **AURUM-thresh**: returns between 3 and 6 sentences according to a threshold learned automatically from the development data
  - Anywhere between 1 and 7 sentences as Highlight Sources in an article – usually between 3 and 6
  - Therefore, impossible to know in advance how many Highlight Sources needed
  - Threshold learned from the development set and then used on the test set
- **Baseline**: always returns the same number of sentences as AURUM
### Results: ROUGE-1, ROUGE-2 (Recall Metric)

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-1 (unigrams)</th>
<th>ROUGE-2 (bigrams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-fixed</td>
<td>47.73</td>
<td>15.98</td>
</tr>
<tr>
<td>AURUM-fixed</td>
<td>49.20 (+3.09%*)</td>
<td>16.53 (+3.63%*)</td>
</tr>
<tr>
<td>Baseline-thresh</td>
<td>55.11</td>
<td>19.31</td>
</tr>
<tr>
<td>AURUM-thresh</td>
<td><strong>56.73 (+2.96%*)</strong></td>
<td><strong>19.66 (+1.87%)</strong></td>
</tr>
</tbody>
</table>

**ROUGE**: Recall-Oriented Understudy for Gisting Evaluation

Current standard evaluation metric for automatic summarization

**ROUGE-N** =

\[
\frac{\sum_{S \in \text{References}} \sum_{\text{ngram}_n \in S} \text{Match}(\text{ngram}_n)}{\sum_{S \in \text{References}} \sum_{\text{ngram}_n \in S} \text{Count}(\text{ngram}_n)}
\]
## Results: ROUGE-1, ROUGE-2 (Recall Metric)

<table>
<thead>
<tr>
<th>System</th>
<th>ROUGE-1 (unigrams)</th>
<th>ROUGE-2 (bigrams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-fixed</td>
<td>47.73</td>
<td>15.98</td>
</tr>
<tr>
<td>AURUM-fixed</td>
<td>49.20 (+3.09%*)</td>
<td>16.53 (+3.63%*)</td>
</tr>
<tr>
<td>Baseline-thresh</td>
<td>55.11</td>
<td>19.31</td>
</tr>
<tr>
<td>AURUM-thresh</td>
<td>56.73 (+2.96%*)</td>
<td>19.66 (+1.87%)</td>
</tr>
</tbody>
</table>

- **ROUGE**: Recall-Oriented Understudy for Gisting Evaluation
- Current standard evaluation metric for automatic summarization
- \[ \text{ROUGE-N} = \frac{\sum_{S \in \text{References}} \sum_{ngram_n \in S} \text{Match}(ngram_n)}{\sum_{S \in \text{References}} \sum_{ngram_n \in S} \text{Count}(ngram_n)} \]
### Individual Feature Importance

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence position</td>
<td>10.23</td>
<td>Spawn 2-gram</td>
<td>3.73</td>
</tr>
<tr>
<td>Proper noun</td>
<td>5.18</td>
<td>Trigger 3-gram</td>
<td>3.70</td>
</tr>
<tr>
<td>Superlative</td>
<td>4.15</td>
<td>1-gram score</td>
<td>2.74</td>
</tr>
<tr>
<td>3-gram score</td>
<td>3.75</td>
<td>Temporal adjective</td>
<td>1.75</td>
</tr>
<tr>
<td>Trigger 2-gram</td>
<td>3.74</td>
<td>Mention of CNN</td>
<td>1.30</td>
</tr>
</tbody>
</table>
Conclusions

- Combination of features
  - Structural: sentence position
  - Word-class: proper nouns, word groups
  - Lexical frequency
- Feature weights learned by machine learning techniques
- Limited use of external resources
- AURUM outperforms a simple yet robust baseline
- Linguistic features are useful and comparable to external data sources
- Annotated data set available soon at the following URL: http://www.science.uva.nl/~christof/data/hl/
Conclusions

- Combination of features
  - Structural: sentence position
  - Word-class: proper nouns, word groups
  - Lexical frequency
- Feature weights learned by machine learning techniques
- Limited use of external resources
- AURUM outperforms a simple yet robust baseline
- Linguistic features are useful and comparable to external data sources

Annotated data set available soon at the following URL:
http://www.science.uva.nl/~christof/data/hl/
Conclusions

- Combination of features
  - Structural: sentence position
  - Word-class: proper nouns, word groups
  - Lexical frequency

- Feature weights learned by machine learning techniques

- Limited use of external resources
- AURUM outperforms a simple yet robust baseline
- Linguistic features are useful and comparable to external data sources

- Annotated data set available soon at the following URL:
  http://www.science.uva.nl/~christof/data/hl/
Americans celebrate, honor the fallen on Memorial Day

STORY HIGHLIGHTS
- Memorial Day marked by parades, tributes, commemorations.
- ABA estimated 90 million Americans would take at least 90 miles during the weekend.
- President Bush gives speech at Arlington National Cemetery.
- Gulf Coast voters abstain from voting with people suffering holiday weekend.

NEW YORK—Veterans and active duty soldiers gathered Monday to honor those who gave their lives.

VA deputy secretary of state, Mayor Bloomberg, and other officials paid tribute to the fallen heroes.

Thank you!

ikt1@dcs.qmul.ac.uk