Entities, Coreference, Discourse, etc.

Hal Daumé III

School of Computing
University of Utah
me@hal3.name
About Summarization Research

- Sentence Extraction
  - Given a document, identify and extract *important* sentences
  - Special models to handle redundancy
  - Typically trained on manually annotated or automatically annotated extracts

There are a wealth of document/abstract pairs that statistical summarization systems could leverage to learn how to create novel abstracts. Detailed studies of such pairs [Jing et al.] show that human abstractors perform a range of very sophisticated operations when summarizing texts, which include reordering, fusion, and paraphrasing. Unfortunately, existing document/abstract alignment models are not powerful enough to capture these operations. To get around directly tackling this problem, researchers in text summarization have employed one of several techniques.

Some researchers [Bano et al.] have developed simple statistical models for aligning documents and headlines. These models, which implement IBM Model 1 [Brown et al.], treat documents and headlines as simple bags of words and learn probabilistic word-based mappings between the words in the documents and the words in the headlines. As our results show, these models are too weak for capturing the operations that are employed by humans in summarizing texts beyond the headline level.

Other researchers have developed models that make unreasonable assumptions about the data, which lead to the utilization of a very small percent of available data. For instance, the document and sentence compression models of Daumé III, Knight, and Marcu [Knight-Marcu02, Daumé-Marcu02] assume that sentences/documents can be summarized only through deletion of contiguous text segments. Knight and Marcu found that from a corpus of $39,060$ abstract sentences, only $1067$ sentence extracts existed: a recall of only $2.7\%$.
About Actual Summaries

➢ Are *not* extracts
  ➢ The sentence is not an appropriate level of granularity
➢ Are *not* compressions
  ➢ Involve lots of rewriting and reordering
➢ Are *not* bags of words
  ➢ Are fluent, grammatical, etc.
➢ **So why do we focus on these unrealistic problems?**

DATA!
Learning Transformations

➢ Document/Abstract pairs:
Connecting Point has become the single largest Mac retailer.
CP Systems tripled its sales of Macintosh systems; it is now the single largest seller of Macintosh.

➢ English/French pairs:
Connecting Point has become the single largest Mac retailer.
L' Pointe de Connecting bécomé l' retailerese oné-most largezze Macintosh.

➢ How does MT solve this problem?

Alignments!!!
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human 1</td>
<td>0.727</td>
<td>0.746</td>
<td>0.736</td>
</tr>
<tr>
<td>Human 2</td>
<td>0.680</td>
<td>0.695</td>
<td>0.687</td>
</tr>
<tr>
<td>GIZA-HMM</td>
<td>0.120</td>
<td>0.260</td>
<td>0.164</td>
</tr>
<tr>
<td>GIZA-Model 4</td>
<td>0.117</td>
<td>0.260</td>
<td>0.161</td>
</tr>
<tr>
<td>GIZA-HMM (flipped)</td>
<td>0.295</td>
<td>0.250</td>
<td>0.271</td>
</tr>
<tr>
<td>GIZA-Model 4 (flipped)</td>
<td>0.280</td>
<td>0.247</td>
<td>0.262</td>
</tr>
<tr>
<td>Decomposition</td>
<td>0.349</td>
<td>0.379</td>
<td>0.363</td>
</tr>
<tr>
<td>PBHMM</td>
<td><strong>0.456</strong></td>
<td><strong>0.686</strong></td>
<td><strong>0.548</strong></td>
</tr>
</tbody>
</table>
Sources of Error

... and OS/2 must provide additional capabilities that justify the expense and effort of migrating to it from DOS.

The transition from DOS to OS/2 has no real precedents. OS/2 must provide additional capabilities that are sufficient to justify the expense and effort required to migrate to it.

The DMP 300 produces good-quality graphics and highly readable text.

Graphics quality is good, with the printer producing a remarkably smooth curved line... but still eminently readable text... The DMP 300...
Entity Detection and Tracking

In February 1990, cardiac arrest deprived Terri Schiavo of oxygen to her brain for five minutes - five minutes that have led to years of emotional distress and legal battles. There was initial hope for recovery, but there came a point at which the views of Terri's future diverged. In 1998, her husband, Michael Schiavo, filed the first petition to remove Terri's feeding tube and allow her to die. Since then, Terri's future has been fought over in the courts until a judge once again ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.
Entity Detection and Tracking

➢ Official formulation:
  ➢ Identify all *entities* appearing in a document and the textual spans (*mentions*) that refer to these entities

➢ Typical interpretation:
  ➢ Identify all *mentions* appearing in a document and discern which mentions refer to the same *entity*

➢ Identifying mentions also involves mention types:
  ➢ Name (NAM), Nominal (NOM), Pronoun (PRO), Premodifier (PRE)

➢ Identifying entities also involves entity types:
  ➢ Person, Organization (+5 subtypes), GPE (+10), Location (+6), Facility (+8), Vehicle (+5), Weapon (+9)
Mention Detection

➢ Use BIO-encoding to obtain sequence labeling problem:

ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

... Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

turn to the Florida Gov. Jeb Bush and then ...
Coreference Resolution

Today, the President of the United States addressed Congress. He told them that he planned on giving a speech tomorrow, during which he would announce to the country that he had decided to...

Original Training Data

Test Data

Binary Training Pairs

Binary Test Pairs

Your Favorite ML Algorithm

Your Favorite Clustering Algorithm

Today, the President of the United States addressed Congress. He told them that he planned on giving a speech tomorrow, during which he would announce to the country that he had decided to...
Coreference Resolution

➢ Choosing a classifier
  ➢ How should we tune it?
  ➢ Can we train on all pairs?

➢ Choosing training instances
  ➢ Use all pairs? Most recent negatives only? Samples?
  ➢ What about the i.i.d. assumption?

➢ Choosing a clustering algorithm
  ➢ How does this choice interact with the classifier?
  ➢ How can we tune parameters?
Two Successful Approaches

- **Classifier:**
  - Maxent multilabel classifier

- **Instances:**
  - All pairs

- **Clustering:**
  - Use a max-link beam search (the Bell-tree algorithm)

[Florian et al., 2004]

- **Classifier:**
  - Perceptron-trained CRF

- **Instances:**
  - All pairs

- **Clustering:**
  - Use generic graph partitioning algorithms

[McCallum & Wellner, 2004]
Summary of Previous Approaches

- **Mention Detection:**
  - Tractable under Markov assumption
  - Inference requires evaluation of forward/backward (sum-product) algorithm for likelihood or margin-based training
    - For perceptron training, requires evaluation of Viterbi algorithm
  - Prediction requires evaluation of Viterbi (max-product) algorithm

- **Coreference Resolution:**
  - Full inference is never tractable
  - Only the McCallum & Wellner model solves the problem directly
    - But has to resort to very simple Perceptron-style updates
  - For the most part, non-integrated classification + clustering
  - Features are usually simple pairwise-comparisons
Features

Lexical: unigrams (words); the bigrams; the two character prefixes and suffixes; the word stem; the case of the word, computed by regular expressions.

Syntactic: unigrams and bigrams of part of speech as well as shallow-parse features.

Semantic: two most common synsets; all hypernyms; for coreference, distance in the WordNet graph between pairs of head words whether one is a part of the other; synset and hypernym information of the preceding and following verbs.

Lists: about 40 lists of common places, organization, names, etc.

Class: word clusters

Inference: models to predict number and gender; output of MEMMs trained off of the MUC6, the MUC7 and ACE data.

String: string match; substring match; string overlap; pronoun match; and normalized edit distance; string nationality match; linguistically-motivated string edit distance; Jaro distance; acronym match.
## Performance

<table>
<thead>
<tr>
<th>Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name to Name</td>
<td>mid to high 90s</td>
</tr>
<tr>
<td>Name to Pronoun</td>
<td>mid 80s</td>
</tr>
<tr>
<td>Name to Nominal</td>
<td>40s-50s</td>
</tr>
</tbody>
</table>
The castle in Camelot remained the residence of the king until he moved it to London.
Discourse “Hobbs' Distance”

John Doe has already secured the vote of most democrats in his constituency, which is already almost enough to win. But without the support of the governor, he is still on shaky ground.
Does Discourse Distance hold?

Holds perfectly in ~ 55%

60% of remaining are ATTRIBUTION

85% of remaining binuclear
Distribution of Entities

Expectation

Truth
<table>
<thead>
<tr>
<th>Name/Instance Data</th>
<th>12416 Palestinian leader</th>
<th>Yasser Arafat</th>
</tr>
</thead>
<tbody>
<tr>
<td>5772 Bosnian Serb leader</td>
<td>Radovan Karadzic</td>
<td></td>
</tr>
<tr>
<td>3839 White House spokesman</td>
<td>Mike McCurry</td>
<td></td>
</tr>
<tr>
<td>3660 Foreign editor</td>
<td>Rick Christie</td>
<td></td>
</tr>
<tr>
<td>3654 News editor</td>
<td>Art Dalglish</td>
<td></td>
</tr>
<tr>
<td>3228 State Department spokesman</td>
<td>Nicholas Burns</td>
<td></td>
</tr>
<tr>
<td>3089 White House spokesman</td>
<td>Marlin Fitzwater</td>
<td></td>
</tr>
<tr>
<td>2528 PLO leader</td>
<td>Yasser Arafat</td>
<td></td>
</tr>
<tr>
<td>2157 first lady</td>
<td>Hillary Rodham Clinton</td>
<td></td>
</tr>
<tr>
<td>2069 spokesman</td>
<td>Alexander Ivanko</td>
<td></td>
</tr>
<tr>
<td>1677 Soviet leader</td>
<td>Mikhail Gorbachev</td>
<td></td>
</tr>
<tr>
<td>1646 envoy</td>
<td>Richard Holbrooke</td>
<td></td>
</tr>
<tr>
<td>1585 Libyan leader</td>
<td>Moammar Gadhafi</td>
<td></td>
</tr>
<tr>
<td>1196 envoy</td>
<td>Dennis Ross</td>
<td></td>
</tr>
<tr>
<td>1195 Communist leader</td>
<td>Gennady Zyuganov</td>
<td></td>
</tr>
<tr>
<td>1152 White House spokesman</td>
<td>Joe Lockhart</td>
<td></td>
</tr>
<tr>
<td>1109 Turkish Cypriot leader</td>
<td>Rauf Denktash</td>
<td></td>
</tr>
<tr>
<td>1057 White House press secretary</td>
<td>Mike McCurry</td>
<td></td>
</tr>
</tbody>
</table>
Using Name/Instance Data

ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri’s parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo’s feeding tube be reinserted.
Gazetteers

- census data and baby name books
- standard gazetteers (countries, cities, islands, ports, provinces and states)
- airport locations
- company names (NASDAQ and NYSE)
- semantically plural words
- list of persons, organizations and locations that were identified by IdentifiFinder.
Using Gazetteers for Coref

list_of(anaphor) + lexeme_of(antecedent)

list_of(anaphor) + list_of(antecedent) + same_word?
WordNet

Distance in graph?

Japan  Russia

Hyponym/Hypernym?

Nearest-verb hypernyms
Count-based Features

- total # of **entities** detected
- total # of **mentions**
- Ratios:
  - **entity:mention** ratio
  - **entity:word** ratio
  - **mention:word** ratio
  - # of **mentions in the current chain to the total # of mentions**
- **size** of the hypothesized chain
- # of **intervening** mentions
- # of **intervening** mentions of the same type; # of intervening **sentence** breaks
- decayed density
Feature Contributions

Str > Lex > Cou > KB > Inf > Lst > Pat > Cla > Sem > Disc
Linkage Types

➢ When hypothesizing merging a mention into a chain, to which chain element do we 'attach'?

Sum/Average link (default)
Min link
Max link (commonly used)
Last link (commonly used)
First link

Intelligent link:

NAM: first(NAM) + last(NOM) + max
NOM: max(NOM) + last(NAM) + max
PRO: \( \text{avg} (\text{PRO+NAM}) + \text{max} \)
Linkage Types

Intelligence | Min | Avg | Max | Last | First
---|---|---|---|---|---
80 | 81 | 82 | 83 | 84 | 85
86 | 87 | 88 | 89 | 90 |
Errors Pre-Engineering

- NAM:
- NOM:
- PRO:
Errors Post-Engineering

- NAM
- NOM
- PRO

<table>
<thead>
<tr>
<th></th>
<th>NAM:</th>
<th>NOM:</th>
<th>PRO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

Colors:
- :NAM
- :NOM
- :PRO
Discussion

Learning makes a difference (LaSO $\rightarrow$ Searn)

Feature engineering makes a bigger one

Many features not obvious
$\Rightarrow$ LOOK at outputs!

Be clever!