Coreference & Coherence

Ling571
Deep Processing Techniques for NLP
March 6, 2017
Roadmap

- Coreference algorithms:
  - Data-driven techniques
  - Deterministic sieves

- Discourse structure
  - Cohesion
    - Topic segmentation
  - Coherence
    - Discourse parsing
Data-driven Reference Resolution

- Prior approaches: Knowledge-based, hand-crafted
- Data-driven machine learning approach
  - Coreference as classification, clustering, ranking problem
    - Mention-pair model:
      - For each pair $NP_i, NP_j$, do they corefer?
      - Cluster to form equivalence classes
    - Entity-mention model
      - For each pair $NP_k$ and cluster $C_j$, should the NP be in the cluster?
    - Ranking models
      - For each $NP_k$, and all candidate antecedents, which highest?
NP Coreference Examples

• Link all NPs refer to same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...
Annotated Corpora

- Available shared task corpora
  - MUC-6, MUC-7 (Message Understanding Conference)
    - 60 documents each, newswire, English
  - ACE (Automatic Content Extraction)
    - Originally English newswire
    - Later include Chinese, Arabic; blog, CTS, Usenet, etc

- Treebanks
  - English Penn Treebank (OntoNotes)
  - German, Czech, Japanese, Spanish, Catalan, Medline
Feature Engineering

- Other coreference (not pronominal) features
  - String-matching features:
    - Mrs. Clinton <-> Clinton

- Semantic features:
  - Can candidate appear in same role w/same verb?
  - WordNet similarity
  - Wikipedia: broader coverage

- Lexico-syntactic patterns:
  - E.g. X is a Y
Typical Feature Set

- 25 features per instance: 2NPs, features, class
  - lexical (3)
    - string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - pronoun_1, pronoun_2, demonstrative_2, indefinite_2, ...
    - number, gender, animacy
    - appositive, predicate nominative
    - binding constraints, simple contra-indexing constraints, ...
    - span, maximalnp, ...
  - semantic (2)
    - same WordNet class
    - alias
  - positional (1)
    - distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - naïve pronoun resolution algorithm
Coreference Evaluation

- Key issues:
  - Which NPs are evaluated?
    - Gold standard tagged or
    - Automatically extracted
  
- How good is the partition?
  - Any cluster-based evaluation could be used (e.g. Kappa)
  - MUC scorer:
    - Link-based: ignores singletons; penalizes large clusters
    - Other measures compensate
Clustering by Classification

- Mention-pair style system:
  - For each pair of NPs, classify +/- coreferent
    - Any classifier
  - Linked pairs form coreferential chains
    - Process candidate pairs from End to Start
    - All mentions of an entity appear in single chain
  - F-measure: MUC-6: 62-66%; MUC-7: 60-61%
    - Soon et. al, Cardie and Ng (2002)
Multi-pass Sieve Approach

- Raghunathan et al., 2010

Key Issues:
- Limitations of mention-pair classifier approach
  - Local decisions over large number of features
    - Not really transitive
  
- Can’t exploit global constraints

- Low precision features may overwhelm less frequent, high precision ones
Multi-pass Sieve Strategy

- Basic approach:
  - Apply tiers of deterministic coreference modules
    - Ordered highest to lowest precision
  - Aggregate information across mentions in cluster
    - Share attributes based on prior tiers

- Simple, extensible architecture
  - Outperforms many other (un-)supervised approaches
Multi-Pass Sieve

Mention Detection

Sieve1: Speaker Identification
Sieve2: String Match
Sieve3: Relaxed String Match
Sieve4: Precise Constructs
Sieve5: Strict Head Match A
Sieve6: Strict Head Match B
Sieve7: Strict Head Match C
Sieve8: Proper Head Noun Match
Sieve9: Relaxed Head Match
Sieve10: Pronoun Match

Post Processing

More global decisions
Recall increases
Pre-Processing and Mentions

- Pre-processing:
  - Gold mention boundaries given, parsed, NE tagged

- For each mention, each module can skip or pick best candidate antecedent
  - Antecedents ordered:
    - Same sentence: by Hobbs algorithm
    - Prev. sentence:
      - For Nominal: by right-to-left, breadth first: proximity/recency
      - For Pronoun: left-to-right: salience hierarchy
    - W/in cluster: aggregate attributes, order mentions
    - Prune indefinite mentions: can’t have antecedents
Multi-pass Sieve Modules

- **Pass 1**: Exact match (N): P: 96%

- **Pass 2**: Precise constructs
  - Predicate nominative, (role) appositive, re:; pronoun, acronym, demonym

- **Pass 3**: Strict head matching
  - Matches cluster head noun AND all non-stop cluster wds AND modifiers AND non i-within-l (embedded NP)

- **Pass 4 & 5**: Variants of 3: drop one of above
Multi-pass Sieve Modules

- Pass 6: Relaxed head match
  - Head matches any word in cluster AND all non-stop cluster wds AND non i-within-l (embedded NP)

- Pass 7: Pronouns
  - Enforce constraints on gender, number, person, animacy, and NER labels
Multi-pass Effectiveness

<table>
<thead>
<tr>
<th>Passes</th>
<th>MUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td>{1}</td>
<td>95.9</td>
</tr>
<tr>
<td>{1,2}</td>
<td>95.4</td>
</tr>
<tr>
<td>{1,2,3}</td>
<td>92.1</td>
</tr>
<tr>
<td>{1,2,3,4}</td>
<td>91.7</td>
</tr>
<tr>
<td>{1,2,3,4,5}</td>
<td>91.1</td>
</tr>
<tr>
<td>{1,2,3,4,5,6}</td>
<td>89.5</td>
</tr>
<tr>
<td>{1,2,3,4,5,6,7}</td>
<td>83.7</td>
</tr>
</tbody>
</table>
### Sieve Effectiveness

- ACE Newswire

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>This work (sieve)</td>
<td>83.8</td>
<td>73.2</td>
<td>78.1</td>
</tr>
<tr>
<td>This work (single pass)</td>
<td>82.2</td>
<td>71.5</td>
<td>76.5</td>
</tr>
<tr>
<td>Haghighi and Klein (2009) +S</td>
<td>77.0</td>
<td>75.9</td>
<td>76.5</td>
</tr>
<tr>
<td>Poon and Domingos (2008)</td>
<td>71.3</td>
<td>70.5</td>
<td>70.9</td>
</tr>
<tr>
<td>Finkel and Manning (2008) +G</td>
<td>78.7</td>
<td>58.5</td>
<td>67.1</td>
</tr>
</tbody>
</table>
Questions

- Good accuracies on (clean) text. What about...
  - Conversational speech?
    - Ill-formed, disfluent
  - Dialogue?
    - Multiple speakers introduce referents
  - Multimodal communication?
    - How else can entities be evoked?
    - Are all equally salient?
More Questions

- Good accuracies on (clean) (English) text: What about..
  - Other languages?
    - Salience hierarchies the same
      - Other factors
    - Syntactic constraints?
      - E.g. reflexives in Chinese, Korean,..
    - Zero anaphora?
      - How do you resolve a pronoun if you can’t find it?
Reference Resolution Algorithms

- Many other alternative strategies:
  - Linguistically informed, saliency hierarchy
    - Centering Theory
  - Machine learning approaches:
    - Supervised: Maxent
    - Unsupervised: Clustering
  - Heuristic, high precision:
    - Cogniac
Conclusions

- Co-reference establishes coherence
- Reference resolution depends on coherence
- Variety of approaches:
  - Syntactic constraints, Recency, Frequency, Role
- Similar effectiveness - different requirements
- Co-reference can enable summarization within and across documents (and languages!)
Discourse Structure
Why Model Discourse Structure? (Theoretical)

- Discourse: not just constituent utterances
  - Create joint meaning
  - Context guides interpretation of constituents
  - How????
    - What are the units?
    - How do they combine to establish meaning?
      - How can we derive structure from surface forms?
    - What makes discourse coherent vs not?
    - How do they influence reference resolution?
Why Model Discourse Structure? (Applied)

- Design better summarization, understanding
- Improve speech synthesis
  - Influenced by structure
- Develop approach for generation of discourse
- Design dialogue agents for task interaction
- Guide reference resolution
Discourse Topic Segmentation

- Separate news broadcast into component stories

On "World News Tonight" this Thursday, another bad day on stock markets, all over the world global economic anxiety. ||
Another massacre in Kosovo, the U.S. and its allies prepare to do something about it. Very slowly. ||
And the millennium bug, Lubbock Texas prepares for catastrophe, Bangalore in India sees only profit. ||
Discourse Segmentation

- Basic form of discourse structure
  - Divide document into linear sequence of subtopics

- Many genres have conventional structures:
  - Academic: Into, Hypothesis, Methods, Results, Concl.
  - Newspapers: Headline, Byline, Lede, Elaboration
  - Patient Reports: Subjective, Objective, Assessment, Plan

- Can guide: summarization, retrieval
Cohesion

- Use of linguistics devices to link text units
  - Lexical cohesion:
    - Link with relations between words
      - Synonymy, Hypernymy
      - *Peel, core and slice the pears and the apples. Add the fruit to the skillet.*
  - Non-lexical cohesion:
    - E.g. anaphora
      - *Peel, core and slice the pears and the apples. Add them to the skillet.*
  - Cohesion chain establishes link through sequence of words

- Segment boundary = dip in cohesion
TextTiling (Hearst ‘97)

- Lexical cohesion-based segmentation
  - Boundaries at dips in cohesion score
  - Tokenization, Lexical cohesion score, Boundary ID

- Tokenization
  - Units?
    - White-space delimited words
    - Stopped
    - Stemmed
    - 20 words = 1 pseudo sentence
Lexical Cohesion Score

- Similarity between spans of text
  - \( b \) = ‘Block’ of 10 pseudo-sentences before gap
  - \( a \) = ‘Block’ of 10 pseudo-sentences after gap
  - How do we compute similarity?
    - Vectors and cosine similarity (again!)

\[
\text{sim}_{\text{cosine}}(\vec{b}, \vec{a}) = \frac{\vec{b} \cdot \vec{a}}{\|\vec{b}\| \|\vec{a}\|} = \frac{\sum_{i=1}^{N} b_i \times a_i}{\sqrt{\sum_{i=1}^{N} b_i^2} \sqrt{\sum_{i=1}^{N} a_i^2}}
\]
Segmentation

- Depth score:
  - Difference between position and adjacent peaks
  - E.g., \((y_{a1} - y_{a2}) + (y_{a3} - y_{a2})\)