Learning Compression & Linguistic Quality

Ling 573
Systems and Applications
May 14, 2015
Roadmap

- Sentence Compression:
  - Learning compression: Tree-based approach
  - Results & Discussion

- Linguistic Quality:
  - Corpus study and analysis
  - Automatic evaluation
  - Improvements for MDS
Tree-based Compression

- Given a phrase-structure parse tree,
- Determine if each node is: removed, retained, or partial
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial
- Issues:
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

- Issues:
  - # possible compressions exponential
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

- Issues:
  - # possible compressions exponential
  - Need some local way of scoring a node
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

Issues:

- # possible compressions exponential
- Need some local way of scoring a node
- Need some way of ensuring consistency
  - I.e. can’t have retain over remove
  - Need to ensure grammaticality
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

- Issues & Solutions:
  - # possible compressions exponential
    - Order parse tree nodes (here post-order)
    - Do beam search over candidate labelings
  - Need some local way of scoring a node
    - Use MaxEnt to compute probability of label
  - Need some way of ensuring consistency
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

Issues & Solutions:
- # possible compressions exponential
  - Order parse tree nodes (here post-order)
  - Do beam search over candidate labelings
- Need some local way of scoring a node
  - Use MaxEnt to compute probability of label
- Need some way of ensuring consistency
  - Restrict candidate labels based on context
- Need to ensure grammaticality
Tree-based Compression

- Given a phrase-structure parse tree,
  - Determine if each node is: removed, retained, or partial

Issues & Solutions:
- # possible compressions exponential
  - Order parse tree nodes (here post-order)
  - Do beam search over candidate labelings
- Need some local way of scoring a node
  - Use MaxEnt to compute probability of label
- Need some way of ensuring consistency
  - Restrict candidate labels based on context
- Need to ensure grammaticality
  - Rerank resulting sentences using n-gram LM
Features

- Basic features:
  - Analogous to those for sequence labeling
Features

- **Basic features:**
  - Analogous to those for sequence labeling

- **Enhancements:**
  - Context features: decisions about child, sibling nodes
Features

• Basic features:
  • Analogous to those for sequence labeling

• Enhancements:
  • Context features: decisions about child, sibling nodes

• Head-driven search:
  • Reorder so head nodes at each level checked first
    • Why?
Features

- Basic features:
  - Analogous to those for sequence labeling

- Enhancements:
  - Context features: decisions about child, sibling nodes

- Head-driven search:
  - Reorder so head nodes at each level checked first
    - Why? If head is dropped, shouldn’t keep rest
    - Revise context features
Summarization Features

- (aka MULTI in paper)
- Calculated based on current decoded word sequence W
- Linear combination of:
Summarization Features

- (aka MULTI in paper)
- Calculated based on current decoded word sequence $W$
- Linear combination of:
  - Score under MaxEnt
  - Query relevance:
    - Proportion of overlapping words with query
  - Importance: Average sumbasic score over $W$
  - Language model probability
  - Redundancy: $1 - -$ proportion of words overlapping summ
## Summarization Results

<table>
<thead>
<tr>
<th>System</th>
<th>DUC 2006</th>
<th>DUC 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C Rate</td>
<td>R-2</td>
</tr>
<tr>
<td>Best DUC system</td>
<td>–</td>
<td>9.56</td>
</tr>
<tr>
<td>Davis et al. (2012)</td>
<td>–</td>
<td>10.2</td>
</tr>
<tr>
<td>SVR</td>
<td>100%</td>
<td>7.78</td>
</tr>
<tr>
<td>LambdaMART</td>
<td>100%</td>
<td>9.84</td>
</tr>
<tr>
<td>Rule-based</td>
<td>78.99%</td>
<td>10.62*↑</td>
</tr>
<tr>
<td>Sequence</td>
<td>76.34%</td>
<td>10.49↑</td>
</tr>
<tr>
<td>Tree (BASIC + Score&lt;sub&gt;Basic&lt;/sub&gt;)</td>
<td>70.48%</td>
<td>10.49↑</td>
</tr>
<tr>
<td>Tree (CONTEXT + Score&lt;sub&gt;Basic&lt;/sub&gt;)</td>
<td>65.21%</td>
<td>10.55*↑</td>
</tr>
<tr>
<td>Tree (HEAD + Score&lt;sub&gt;Basic&lt;/sub&gt;)</td>
<td>66.70%</td>
<td>10.66*↑</td>
</tr>
<tr>
<td>Tree (HEAD + MULTI)</td>
<td>70.20%</td>
<td>11.02*↑</td>
</tr>
</tbody>
</table>
## Compression Results

<table>
<thead>
<tr>
<th>System</th>
<th>C Rate</th>
<th>Uni-Prec</th>
<th>Uni-Rec</th>
<th>Uni-F1</th>
<th>Rel-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HedgeTrimmer</td>
<td>57.64%</td>
<td>0.72</td>
<td>0.65</td>
<td>0.64</td>
<td>0.50</td>
</tr>
<tr>
<td>McDonald (2006)</td>
<td>70.95%</td>
<td>0.77</td>
<td>0.78</td>
<td>0.77</td>
<td>0.55</td>
</tr>
<tr>
<td>Martins and Smith (2009)</td>
<td>71.35%</td>
<td>0.77</td>
<td>0.78</td>
<td>0.77</td>
<td>0.56</td>
</tr>
<tr>
<td>Rule-based</td>
<td>87.65%</td>
<td>0.74</td>
<td>0.91</td>
<td>0.80</td>
<td>0.63</td>
</tr>
<tr>
<td>Sequence</td>
<td>70.79%</td>
<td>0.77</td>
<td>0.80</td>
<td>0.76</td>
<td>0.58</td>
</tr>
<tr>
<td>Tree (BASIC)</td>
<td>69.65%</td>
<td>0.77</td>
<td>0.79</td>
<td>0.75</td>
<td>0.56</td>
</tr>
<tr>
<td>Tree (CONTEXT)</td>
<td>67.01%</td>
<td>0.79</td>
<td>0.78</td>
<td>0.76</td>
<td>0.57</td>
</tr>
<tr>
<td>Tree (HEAD)</td>
<td>68.06%</td>
<td>0.79</td>
<td>0.80</td>
<td>0.77</td>
<td>0.59</td>
</tr>
</tbody>
</table>
Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features
Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features

- Rule-based approach surprisingly competitive
  - Though less aggressive in terms of compression
Discussion

- Best system incorporates:
  - Tree structure
  - Machine learning
  - Summarization features

- Rule-based approach surprisingly competitive
  - Though less aggressive in terms of compression

- Learning based approaches enabled by sentence compression corpus
General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:
General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:
    - Heuristic vs Learned
    - Surface patterns vs parse trees vs SRL

- Even with linguistic constraints
  - Often negatively impact linguistic quality
General Discussion

- Broad range of approaches:
  - Informed by similar linguistic constraints
  - Implemented in different ways:
    - Heuristic vs Learned
    - Surface patterns vs parse trees vs SRL

- Even with linguistic constraints
  - Often negatively impact linguistic quality
  - Key issue: errors in linguistic analysis
    - POS taggers → Parsers → SRL, etc
General Discussion

- Compression has range of uses:
  - Removing irrelevant information for selection
  - Improving readability
  - Allowing inclusion of more information
  - Slightly different strategies for each
Linguistic Quality
Evaluation

- Shared tasks:
  - Take content as primary evaluation measure
    - ROUGE, Pyramid, (manual) Responsiveness
  - Linguistic quality also part of formal evaluation

- TAC “Readability”:
  - Scored manually on 5-point Likert scale
  - Aims to capture readability, fluency
    - Independent of summary content
What is “Readability”?

- According to TAC,
- Assessors consider (and rate 1-5) each of:
  - Grammaticality:
    - No fragments, datelines, ill-formed sentences, etc
  - Non-redundancy:
    - No unnecessary repetition: includes content, sentences, or full NPs when pronoun is better
  - Referential clarity:
    - Both presence/salience of antecedents, relevance of items
  - Focus:
    - Only content related to summary
  - Coherence: “Well-structured”
What is “Readability”? II

- Definition subsumes many phenomena, errors
- What types of errors do these systems make?
- What errors, issues are reflected in the scores?
- LVQSumm (Friedrich et al, 2013)
  - Annotate linguistic “violations” in automatic summaries
    - TAC2011 data: 2000 “peer” summaries; GLOW system
  - Categorize and tabulate
  - Assess correlation with Readability scores
Charles Carl Roberts IV may have planned to molest the girls at the Amish school, but police have no evidence that he actually did. Charles Carl Roberts IV entered the West Nickel Mines Amish School in Lancaster County and shot 10 girls, killing five. The suspect apparently called his wife from a cell phone shortly before the shooting began, saying he was “acting out in revenge for something that happened 20 years ago, Miller said. The gunman, a local truck driver Charles Roberts, was apparently acting in “revenge” for an incident that happened to him 20 years ago.
Violation Categories

- **Entity mentions:**
  - Affect coreference and readability
  - 1st mention w/o explanation; subseq. Mention w/expl
  - Def NP w/o prev mention; indef NP w/ prev mention
  - Pron w/missing, misleading antecedent; Acronym

- **Clausal level:**
  - Arbitrary spans – up to sentence level
  - Incomplete sent, dateline info, other ungrammatical
  - No semantic relation, wrong discourse rel’n, redundancy
<table>
<thead>
<tr>
<th>violation type</th>
<th>count</th>
<th>avg/doc</th>
<th>Pearson’s $r$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Readability</td>
<td>Pyramid</td>
<td>Respons.</td>
<td></td>
</tr>
<tr>
<td><strong>entity level violations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNP-REF</td>
<td>958</td>
<td>0.50</td>
<td>-0.122</td>
<td>-0.166</td>
<td>-0.133</td>
<td></td>
</tr>
<tr>
<td>FM-EXPL</td>
<td>792</td>
<td>0.41</td>
<td>0.006</td>
<td>-0.050</td>
<td>-0.066</td>
<td></td>
</tr>
<tr>
<td>INP+REF</td>
<td>430</td>
<td>0.22</td>
<td>-0.052</td>
<td>0.235</td>
<td>0.109</td>
<td></td>
</tr>
<tr>
<td>PRN+MISSA</td>
<td>361</td>
<td>0.19</td>
<td>-0.191</td>
<td>-0.140</td>
<td>-0.156</td>
<td></td>
</tr>
<tr>
<td>SM+EXPL</td>
<td>162</td>
<td>0.08</td>
<td>0.020</td>
<td>0.089</td>
<td>0.045</td>
<td></td>
</tr>
<tr>
<td>PRN+MILSA</td>
<td>27</td>
<td>0.01</td>
<td>-0.065</td>
<td>-0.073</td>
<td>-0.089</td>
<td></td>
</tr>
<tr>
<td>ACR-EXPL</td>
<td>11</td>
<td>0.01</td>
<td>-0.038</td>
<td>-0.056</td>
<td>-0.006</td>
<td></td>
</tr>
<tr>
<td>sum(DNP-REF, PRN+MISSA)</td>
<td>1319</td>
<td>0.68</td>
<td>-0.204</td>
<td>-0.208</td>
<td>-0.192</td>
<td></td>
</tr>
<tr>
<td>sum(entity level violations)</td>
<td>2741</td>
<td>1.42</td>
<td>-0.167</td>
<td>-0.074</td>
<td>-0.127</td>
<td></td>
</tr>
<tr>
<td><strong>clause level violations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCOMPLSN</td>
<td>1,044</td>
<td>0.54</td>
<td>-0.210</td>
<td>0.000</td>
<td>-0.029</td>
<td></td>
</tr>
<tr>
<td>OTHRUNGR</td>
<td>793</td>
<td>0.41</td>
<td>-0.180</td>
<td>0.007</td>
<td>-0.016</td>
<td></td>
</tr>
<tr>
<td>INCLUDATE</td>
<td>412</td>
<td>0.21</td>
<td>-0.090</td>
<td>0.039</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>REDUNDINF</td>
<td>504</td>
<td>0.26</td>
<td>-0.160</td>
<td>0.156</td>
<td>0.077</td>
<td></td>
</tr>
<tr>
<td>NOSEMREL</td>
<td>142</td>
<td>0.07</td>
<td>-0.148</td>
<td>-0.102</td>
<td>-0.132</td>
<td></td>
</tr>
<tr>
<td>NODISREL</td>
<td>91</td>
<td>0.05</td>
<td>-0.025</td>
<td>-0.081</td>
<td>-0.062</td>
<td></td>
</tr>
<tr>
<td>misleading discourse connectives</td>
<td>114</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sum(clause level violations)</td>
<td>2,986</td>
<td>1.54</td>
<td>-0.325</td>
<td>0.041</td>
<td>-0.016</td>
<td></td>
</tr>
<tr>
<td>sum(clause level violations,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNP-REF, PRN+MISSA)</td>
<td>4,305</td>
<td>2.22</td>
<td>-0.385</td>
<td>-0.084</td>
<td>-0.122</td>
<td></td>
</tr>
<tr>
<td>sum(all violations)</td>
<td>5,727</td>
<td>2.96</td>
<td>-0.356</td>
<td>-0.022</td>
<td>-0.101</td>
<td></td>
</tr>
</tbody>
</table>
Further Analysis

- Linear model investigates the relationship of particular errors to readability
- Most significant factors: Missing/Misleading refs, fragments, redundant content, poor coherence
- Total # of errors well-correlated with system ranks

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.407</td>
<td>DNP-REF</td>
<td>-0.157</td>
</tr>
<tr>
<td>ACR-EXPL</td>
<td>-0.361</td>
<td>OTHRUNGR</td>
<td>-0.155</td>
</tr>
<tr>
<td>PRN+MISLA</td>
<td>-0.355</td>
<td>INCLUDATE</td>
<td>-0.151</td>
</tr>
<tr>
<td>INCOMPLSN</td>
<td>-0.275</td>
<td>INP+REF</td>
<td>-0.067</td>
</tr>
<tr>
<td>NOSEMREL</td>
<td>-0.262</td>
<td>NODISREL</td>
<td>-0.046</td>
</tr>
<tr>
<td>REDUNDINF</td>
<td>-0.259</td>
<td>FM-EXPL</td>
<td>-0.023</td>
</tr>
<tr>
<td>PRN+MISSA</td>
<td>-0.236</td>
<td>SM+EXPL</td>
<td>0.038</td>
</tr>
</tbody>
</table>
Automatic Evaluation of Linguistic Quality

• Motivation:
  • No focus on linguistic quality b/c no way to tune to it
  • Everyone uses ROUGE b/c you can tune
    • Explicitly tuned in many ML models

• Alternative strategies:
  • Micro: Learn to predict component scores
  • Macro: Learn to predict overall readability score
    • Intuitively: error count (LVQSumm) predicts well, but...
      • Errors manually derived
Micro-Quality Prediction

- (Pitler et al, 2010) via SVM ranking
- Evaluate multiple measures aimed to model LQ
  - General word choice, sequence: Language Models
  - Reference form:
    - Named Entities: modification for 1\textsuperscript{st} mention of PER
    - NP syntax: POS, phrase tags in NPs
  - Local coherence
    - Devices: counts of pron, dem, connectives,...
    - Continuity: adjacency in source, coref w/prev, same, cosine
  - Sentence fluency: features from MT eval
  - Coh-Metrix: set of psycho-ling motivated feats, LSA sim
  - Word coherence: cross-sentence word cooccurrence patterns
  - Entity coherence: via Entity-grids (Brown toolkit)
Results

- System level
- Summary level

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang. models</td>
<td>87.6</td>
<td>83.0</td>
<td>91.2</td>
<td>85.2</td>
<td>86.3</td>
<td>66.3</td>
<td>57.6</td>
<td>62.2</td>
<td>60.5</td>
<td>62.5</td>
</tr>
<tr>
<td>Named ent.</td>
<td>78.5</td>
<td>83.6</td>
<td>82.1</td>
<td>74.0</td>
<td>69.6</td>
<td>52.9</td>
<td>54.4</td>
<td>60.0</td>
<td>54.1</td>
<td>52.5</td>
</tr>
<tr>
<td>NP syntax</td>
<td>85.0</td>
<td>83.8</td>
<td>87.0</td>
<td>76.6</td>
<td>79.2</td>
<td>59.0</td>
<td>50.8</td>
<td>59.1</td>
<td>54.5</td>
<td>55.1</td>
</tr>
<tr>
<td>Coh. devices</td>
<td>82.1</td>
<td>79.5</td>
<td>82.7</td>
<td>82.3</td>
<td>83.7</td>
<td>56.8</td>
<td>54.4</td>
<td>55.2</td>
<td>52.7</td>
<td>53.6</td>
</tr>
<tr>
<td>Continuity</td>
<td>88.8</td>
<td>88.5</td>
<td>92.9</td>
<td>89.2</td>
<td>91.4</td>
<td>61.7</td>
<td>62.5</td>
<td>69.7</td>
<td>65.4</td>
<td>70.4</td>
</tr>
<tr>
<td>Sent. fluency</td>
<td>91.7</td>
<td>78.9</td>
<td>87.6</td>
<td>82.3</td>
<td>84.9</td>
<td>69.4</td>
<td>52.5</td>
<td>64.4</td>
<td>61.9</td>
<td>62.6</td>
</tr>
<tr>
<td>Coh-Metrix</td>
<td>87.2</td>
<td>86.0</td>
<td>88.6</td>
<td>83.9</td>
<td>86.3</td>
<td>65.5</td>
<td>67.6</td>
<td>67.9</td>
<td>63.0</td>
<td>62.4</td>
</tr>
<tr>
<td>Word coh.</td>
<td>81.7</td>
<td>76.0</td>
<td>87.8</td>
<td>81.7</td>
<td>79.0</td>
<td>54.7</td>
<td>55.5</td>
<td>53.3</td>
<td>53.2</td>
<td>53.7</td>
</tr>
<tr>
<td>Entity coh.</td>
<td>90.2</td>
<td>88.1</td>
<td>89.6</td>
<td>85.0</td>
<td>87.1</td>
<td>61.3</td>
<td>62.0</td>
<td>64.3</td>
<td>64.2</td>
<td>63.6</td>
</tr>
<tr>
<td>Meta ranker</td>
<td>92.9</td>
<td>87.9</td>
<td>91.9</td>
<td>87.8</td>
<td>90.0</td>
<td>71.0</td>
<td>68.6</td>
<td>73.1</td>
<td>67.4</td>
<td>70.7</td>
</tr>
</tbody>
</table>
Findings

- Overall accuracies quite good
- Systems overall easier to rank than particular input
  - Smoothes variance, larger sample
- Continuity related features best across components
  - Ensemble of ordering, coref, cosine similarity cues
  - Though LSA-based system detects redundancy well
- Specifically tuned fluency scorer works on fluency
Macro-Quality Prediction

- (Lin et al, 2012) Downloadable

- High-level idea:
  - Discourse version of entity grid
    - Columns: entities (same head)
    - Rows: sentences
    - Cell values: PDTB Relation.Arg# tuples

- Variants:
  - Inter-cell sequence frequencies
    - + Additional tuples: {Non--}Explicit.Relation.Arg#
    - + Intra-cell “sequences”
### Results

- Very strong correlations w/manual readability score
- Beats prior predictors

<table>
<thead>
<tr>
<th></th>
<th>Initial</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>S</td>
<td>K</td>
<td>P</td>
<td>S</td>
<td>K</td>
</tr>
<tr>
<td>R-2</td>
<td>0.7524</td>
<td>0.3975</td>
<td>0.2925</td>
<td>0.6580</td>
<td>0.3732</td>
<td>0.2635</td>
</tr>
<tr>
<td>R-SU4</td>
<td>0.7840</td>
<td>0.3953</td>
<td>0.2925</td>
<td>0.6716</td>
<td>0.3627</td>
<td>0.2540</td>
</tr>
<tr>
<td>BE</td>
<td>0.7171</td>
<td>0.4091</td>
<td>0.2911</td>
<td>0.5455</td>
<td>0.2445</td>
<td>0.1622</td>
</tr>
<tr>
<td>4</td>
<td>0.8194</td>
<td>0.4937</td>
<td>0.3658</td>
<td>0.7423</td>
<td>0.4819</td>
<td>0.3612</td>
</tr>
<tr>
<td>6</td>
<td>0.7840</td>
<td>0.4070</td>
<td>0.3036</td>
<td>0.6830</td>
<td>0.4263</td>
<td>0.3141</td>
</tr>
<tr>
<td>12</td>
<td>0.7944</td>
<td>0.4973</td>
<td>0.3589</td>
<td>0.6443</td>
<td>0.3991</td>
<td>0.3062</td>
</tr>
<tr>
<td>18</td>
<td>0.7914</td>
<td>0.4746</td>
<td>0.3510</td>
<td>0.6698</td>
<td>0.3941</td>
<td>0.2856</td>
</tr>
<tr>
<td>23</td>
<td>0.7677</td>
<td>0.4341</td>
<td>0.3162</td>
<td>0.7054</td>
<td>0.4223</td>
<td>0.3014</td>
</tr>
<tr>
<td>LIN</td>
<td>0.8556</td>
<td>0.6593</td>
<td>0.4953</td>
<td>0.7850</td>
<td>0.6671</td>
<td>0.5008</td>
</tr>
<tr>
<td>LIN+C</td>
<td>0.8612</td>
<td>0.6703</td>
<td>0.4984</td>
<td>0.7879</td>
<td>0.6828</td>
<td>0.5135</td>
</tr>
<tr>
<td>LIN+E</td>
<td>0.8619</td>
<td>0.6855</td>
<td>0.5079</td>
<td>0.7928</td>
<td>0.6990</td>
<td>0.5309</td>
</tr>
<tr>
<td>DICOMER</td>
<td>0.8666</td>
<td>0.7122</td>
<td>0.5348</td>
<td>0.8100</td>
<td>0.7145</td>
<td>0.5435</td>
</tr>
</tbody>
</table>
Referring to People in News Summaries

- Intuition:
  - Referring expressions common source of errors
  - References to people prevalent in news data, summaries
  - Information status constrains realization
  - Targeted rewriting can improve readability

- Approach:
  - Exploit information status distinctions
    - Automatically identified
  - Use to guide rule-based generation of referring expressions
Challenges

• Lack of training data:
  • No summary data labeled for information status

• Readers sensitive to referring expressions
  • Prior work on NP rewriting has shown mixed results
    • Some improvement, some failures

• Relies on potentially errorful coref, other processing
While the British government defended the arrest, it took no stand on extradition of Pinochet to Spain, leaving it to the courts.

While the British government defended the arrest in London of former Chilean dictator Augusto Pinochet, it took no stand on extradition of Pinochet to Spain, leaving it to British courts.
NP Rewrite: mixed example

• *Duisenberg* has said growth in the euro area countries next year will be about 2.5 percent, lower than the 3 percent predicted earlier.

• *Wim Duisenberg, the head of the new European Central Bank*, has said growth in the euro area countries next year will be about 2.5 percent, lower than just 1 percent in the euro-zone unemployment predicted earlier.
Information Status

• Build on three key distinctions:
  • Discourse-new vs discourse-old:
    • First mention handling vs others
  • Hearer-new vs hearer-old:
    • Distinguish well-known individuals from others
      • Don’t waste space describing well-known individuals
  • Major vs minor character:
    • Salience of the person in the event
Corpus Analysis

- Assess relation between:
  - information status and referring expressions

<table>
<thead>
<tr>
<th></th>
<th>Discourse-new</th>
<th>Discourse-old</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name Form</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full name</td>
<td>0.97</td>
<td>0.08</td>
</tr>
<tr>
<td>Surname only</td>
<td>0.02</td>
<td>0.87</td>
</tr>
<tr>
<td>Other (e.g., Britney, JLo)</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Pre-Modification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any</td>
<td>0.51</td>
<td>0.21</td>
</tr>
<tr>
<td>None</td>
<td>0.49</td>
<td>0.79</td>
</tr>
<tr>
<td><strong>Post-Modification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0.60</td>
<td>0.89</td>
</tr>
<tr>
<td>Apposition</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>Relative clause</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>Other</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>Any Modification</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Either Pre- or Post-)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some Modification</td>
<td>0.76</td>
<td>0.30</td>
</tr>
<tr>
<td>No Modification</td>
<td>0.24</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Generating Discourse-New/Old

- If discourse-new,
  - If the NP head is a person name,
    - If appears with pre-modifier in text, write as:
      - Longest pre-modifier + full name
    - Else if it appears with an apposition modifier
      - Add that to the reference
  - Else don’t rewrite
- Else use surname only
- Significantly preferred over original forms
Hearer & Salience

- Discourse-new status:
  - Obvious from summary

- How do we establish hearer or major/minor status?

- Categorize based on human summaries (gold)
  - Specifically by their referring expressions:
    - Hearer-old (i.e. familiar)
      - Title/role+surname or unmodified fullname
    - Major:
      - Referred to by name in some human summary of topic
        - 258 major/3926 minor by data
Training & Application

- Trained classifiers to recognize – using features in docset
  - H-New/Old: F-measure: 0.75 on both classes
  - Major/Minor: F: Major: 0.6; Minor: 0.98
    - All significantly better than baseline

- Create rules based on classification to:
  - Use names (only) for major characters (o.w. common)
  - Exclude post-modifiers for hearer-old, tune for new
  - Include titles for initial mentions
  - Include affiliation premodifers based on hearer/salience
Evaluation

- Created (nearly) deterministic rule set
  - Based on information status classification
  - To rewrite referring expressions in extractive summaries

- Evaluated in paired preference tests over:
  - Original Extractive and Rewritten Summaries

- Where a preference was expressed,
  - Rewritten summaries rated as more coherent
  - Extractive rated as more informative
    - Why? Rewrite rules generally shrink rather than add content
Summary

- Can identify particular correlates of readability scores

- Can automatically predict linguistic quality scores

- Build systems that focus on frequent violations
  - Yield systematic improvements in linguistic quality