AUTOMATED POSTEDITING OF ENGLISH

KEVIN KNIGHT, ISI
ISHWAR CHANDER, USC
POST-EDITING

DETACHABLE PE. MODULE
TWO TYPES OF POST-EDITING MODULES

1. ONE-SIZE-FITS-ALL. Helps solve routine problems faced by many systems in many domains.

2. ADAPTIVE. Improves accuracy of a general-purpose system, as fielded in a particular domain.
ADAPTIVE POST-EDITORS

ADAPTATION ALLOWS GENERAL-PURPOSE SYSTEM TO PERFORM IN SPECIAL-PURPOSE SITUATION.
ONE-SIZE-FITS-ALL POST-EDITORS

EXAMPLE:

INSERTING ARTICLES (A/AN/THE) AND PLURALS INTO ENGLISH TEXT.
Stelco Inc. said it plans to shut down three Toronto-area plant, moving their fastener operation, to a leased facility in Brantford, Ontario.

The company said the fastener business "has been under severe cost pressure for some time." Fasteners, nuts and bolts, are sold to the North American auto market.

The company spokesman declined to estimate the impact of closure on earnings. He said the new facility will employ 500 of existing 600 employees. Steelmaker employs about 16,000 people.

Stelco said it has the option to lease a 350,000-square-foot building in Brantford and proposes to spend 24.5 million Canadian dollar. (US$20.9 million) on the facility. Three existing plants and their land will be sold.
EVALUATION

TASK DEFINITION:
- TAKE DOCUMENT
- REPLACE ALL "THE", "A", "AN" BY DUMMY SYMBOL "XYZ"
- PUT ARTICLES BACK
- HOW MANY DID YOU GET RIGHT?

EASY TO TEST; DON'T NEED FULL MT SYSTEM OR NON-NATIVE TEXTS.
ASSUMPTIONS

- PLURALS ARE GIVEN (AS IN RUSSIAN).
- PLACEHOLDERS ARE GIVEN — NOT YET PREDICTING THE NULL ARTICLE.

EXPERIMENTS CONDUCTED JOINTLY WITH I. CHANDER.
Our first results a respectable 65%

The algorithm always guess "the"!
UPPER BOUNDS

tested human subject

ERIC MEIZ

FULL CONTEXT: 96%

⇒ ARTICLES CONTAIN ALMOST NO INFORMATION (< .25 BITS)

BUT, WE SHOULDN'T EXPECT TO DO BETTER THAN 96%
烟雾缭绕的俱乐部
美国的贴现率
公用事业
年
城市
公司
版权法
Midwest Financial 子公司
第三季度
经济放缓
基准 30 年期债券
食品服务连锁运营商
公司
公司
正确的
30 天简单收益率下降
决策
政府
白宫
多元化

有限的上下文 (V.P)

$\Rightarrow 80\%$
MEANING...

UNIQUE, PREVIOUSLY MENTIONED, GENERIC, INDIRECT ANAPHORA, SPORADIC REFERENCE, NON-REFERRING
→ AI-COMPLETE!
→ WHO CARES.
LET'S LOOK AT
SOME DATA

E.G., 80 MBYTES
OF ONLINE
WALL STREET
JOURNAL
ARTICLES.
DATA ANALYSIS

- Head of noun phrase seems important
  -- White House
  -- Sand
  -- Problems

- Also premodifiers
  -- Biggest X
  -- Next X

- Or both
  -- Federal deficit
• AND THE WORD AFTER
  THE HEAD NOUN
  ___ X AGO

• OR BEFORE
  TRIPLE ___ X

• COMBINATIONS
  REST OF ___ X
  CLEAR ___ WAY
  40% ___ YEAR
  ___ SIGN OF

• ___ YEAR

...
INTEGRATING ALL THE PIECES OF EVIDENCE

NOUN = "white house"  562  99%

PREMOD 2 SUPERLATIVE

NP-1 = PREPOSITION & NOUN = "year"

NOUN = PLURAL

NP-1 = "."

NP+1 = "ago"  881  0%

"pay, — decision that could"
DECISION TREE

- LIST OF QUESTIONS (FEATURES)
- LIST OF NOUN PHRASES (TRAINING INSTANCES)

- CHOOSING A QUESTION

\[ Q_i, Q_j, Q_k \]
SCALE

400,000 training examples

30,000 features

Not feasible to do 400,000 x 30,000 operations at each node.

But, most features only apply to a fraction of instances. Can compute "no" distribution by subtraction.
PHRASES ENDING IN 'PRESIDENT'

# OF NODES IN TREE.

# OF TRAINING Instances
### For Head Noun: **stock**

<table>
<thead>
<tr>
<th>Train-Pct</th>
<th>Questions</th>
<th>Decision-Nodes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>150</td>
<td>13</td>
<td>1253 (the=1212/1272, a-an=41/178)</td>
<td>197</td>
<td>0.86</td>
</tr>
<tr>
<td>20</td>
<td>267</td>
<td>34</td>
<td>1132 (the=1077/1140, a-an=55/149)</td>
<td>157</td>
<td>0.88</td>
</tr>
<tr>
<td>30</td>
<td>370</td>
<td>44</td>
<td>985 (the=931/998, a-an=54/130)</td>
<td>143</td>
<td>0.87</td>
</tr>
<tr>
<td>40</td>
<td>478</td>
<td>51</td>
<td>868 (the=814/854, a-an=54/113)</td>
<td>99</td>
<td>0.90</td>
</tr>
<tr>
<td>50</td>
<td>564</td>
<td>55</td>
<td>721 (the=680/718, a-an=41/88)</td>
<td>85</td>
<td>0.89</td>
</tr>
<tr>
<td>60</td>
<td>619</td>
<td>75</td>
<td>581 (the=539/575, a-an=42/70)</td>
<td>64</td>
<td>0.90</td>
</tr>
<tr>
<td>70</td>
<td>677</td>
<td>88</td>
<td>433 (the=400/430, a-an=33/54)</td>
<td>51</td>
<td>0.89</td>
</tr>
<tr>
<td>80</td>
<td>741</td>
<td>98</td>
<td>285 (the=264/287, a-an=21/36)</td>
<td>38</td>
<td>0.88</td>
</tr>
<tr>
<td>90</td>
<td>775</td>
<td>98</td>
<td>149 (the=138/144, a-an=11/18)</td>
<td>13</td>
<td>0.92</td>
</tr>
</tbody>
</table>

### For Head Noun: **year**

<table>
<thead>
<tr>
<th>Train-Pct</th>
<th>Questions</th>
<th>Decision-Nodes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>438</td>
<td>45</td>
<td>4625 (the=1511/1819, a-an=3114/3339)</td>
<td>533</td>
<td>0.90</td>
</tr>
<tr>
<td>20</td>
<td>755</td>
<td>78</td>
<td>4159 (the=1453/1659, a-an=2706/2926)</td>
<td>426</td>
<td>0.91</td>
</tr>
<tr>
<td>30</td>
<td>1059</td>
<td>97</td>
<td>3704 (the=1318/1449, a-an=2386/2563)</td>
<td>308</td>
<td>0.92</td>
</tr>
<tr>
<td>40</td>
<td>1333</td>
<td>131</td>
<td>3206 (the=1103/1205, a-an=2103/2234)</td>
<td>233</td>
<td>0.93</td>
</tr>
<tr>
<td>50</td>
<td>1589</td>
<td>172</td>
<td>2681 (the=902/983, a-an=1779/1883)</td>
<td>185</td>
<td>0.94</td>
</tr>
<tr>
<td>60</td>
<td>1767</td>
<td>190</td>
<td>2148 (the=690/761, a-an=1458/1532)</td>
<td>145</td>
<td>0.94</td>
</tr>
<tr>
<td>70</td>
<td>1922</td>
<td>226</td>
<td>1616 (the=505/558, a-an=1111/1162)</td>
<td>104</td>
<td>0.94</td>
</tr>
<tr>
<td>80</td>
<td>2079</td>
<td>241</td>
<td>1068 (the=329/363, a-an=739/784)</td>
<td>79</td>
<td>0.93</td>
</tr>
<tr>
<td>90</td>
<td>2140</td>
<td>258</td>
<td>532 (the=157/176, a-an=375/396)</td>
<td>42</td>
<td>0.93</td>
</tr>
</tbody>
</table>

### For Head Noun: **president**

<table>
<thead>
<tr>
<th>Train-Pct</th>
<th>Questions</th>
<th>Decision-Nodes</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>103</td>
<td>21</td>
<td>1001 (the=564/687, a-an=437/596)</td>
<td>282</td>
<td>0.78</td>
</tr>
<tr>
<td>20</td>
<td>187</td>
<td>44</td>
<td>962 (the=532/620, a-an=430/520)</td>
<td>178</td>
<td>0.84</td>
</tr>
<tr>
<td>30</td>
<td>268</td>
<td>59</td>
<td>816 (the=456/538, a-an=360/460)</td>
<td>182</td>
<td>0.62</td>
</tr>
<tr>
<td>40</td>
<td>327</td>
<td>83</td>
<td>714 (the=407/462, a-an=307/393)</td>
<td>141</td>
<td>0.84</td>
</tr>
<tr>
<td>50</td>
<td>410</td>
<td>120</td>
<td>598 (the=329/382, a-an=269/331)</td>
<td>115</td>
<td>0.64</td>
</tr>
<tr>
<td>60</td>
<td>449</td>
<td>135</td>
<td>484 (the=270/311, a-an=214/259)</td>
<td>86</td>
<td>0.85</td>
</tr>
<tr>
<td>70</td>
<td>488</td>
<td>147</td>
<td>365 (the=215/235, a-an=150/193)</td>
<td>63</td>
<td>0.85</td>
</tr>
<tr>
<td>80</td>
<td>536</td>
<td>157</td>
<td>248 (the=137/156, a-an=111/129)</td>
<td>37</td>
<td>0.67</td>
</tr>
<tr>
<td>90</td>
<td>562</td>
<td>171</td>
<td>127 (the=69/75, a-an=58/68)</td>
<td>16</td>
<td>0.69</td>
</tr>
</tbody>
</table>
DISCUSSION

- POWER OF "KNOW NOTHING" STATISTICS

- BUT, WE ALREADY HAVE KNOWLEDGE IN THE SYSTEM:
  - NOUN PHRASE PARSER
  - WORD CLASSES (PLURAL, SUPERLATIVE, PREPOSITION, VERB...)

- IMPROVEMENTS WILL NOW COME FROM ADDING KNOWLEDGE