The (Non)Utility of Semantics for Coreference Resolution

Michael Strube
Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany
The (Non)Utility of Predicate-Argument Frequencies for Pronoun Interpretation

Andrew Kehler* 
UC San Diego
akehler@ucsd.edu

Douglas Appelt 
SRI International
appelt@ai.sri.com

Lara Taylor* 
UC San Diego
lmtaylor@ucsd.edu

Aleksandr Simma† 
UC San Diego
asimma@ucsd.edu

Abstract

State-of-the-art pronoun interpretation systems rely predominantly on morphosyntactic contextual features. While the use of deep knowledge and inference to improve these models would appear technically infeasible, previous work has suggested that predicate-argument statistics mined from naturally-occurring data could provide a useful approximation to such knowledge. We test this idea in several system configurations, and conclude from our results and subsequent error analysis that such statistics offer little or no predictive information above that provided by morphosyntax.

(1) He worries that Glendening’s initiative could push his industry over the edge, forcing it to shift operations elsewhere.

Of course, no well-suited knowledge base and accompanying inference procedure exists that can deliver such a capability robustly in an open domain.

In lieu of this capability, previous authors have suggested that what can be viewed as a more superficial form of semantic information – predicate-argument statistics mined from naturally-occurring data – could be used to capture certain selectional regularities. For instance, such statistics might reveal that forcing_industry is a more likely verb-object combination in naturally-occurring data than forcing_initiative or forcing_edge. Assuming that such statistics imply that industries are more likely
Kehler et al. (2004)

- deep knowledge and inference should improve pronoun resolution but appear to be technically infeasible (back in 2004)
- can predicate-argument frequencies mined from corpora provide an approximation to such knowledge?
- does it actually improve pronoun resolution?
He worries that Glendening’s initiative could push his industry over the edge, forcing it to shift operations elsewhere.

predicate argument frequencies might reveal that FORCING_INDUSTRY is more likely than FORCING_INITIATIVE or FORCING_EDGE
Kehler et al. (2004)

predicate-argument frequencies:

- data: TDT-2 corpus with 1,321,072 subject-verb relationships, 1,167,189 verb-object relationships, 301,477 possessive-noun relationships (formulas after Dagan et al. (1995))

\[
\text{stat}(C) = P(\text{tuple}(C,A)|C) = \frac{\text{freq}(\text{tuple}(C,A))}{\text{freq}(C)}
\]

\[
\ln\left(\frac{\text{stat}(C_2)}{\text{stat}(C_1)}\right) > K \times (\text{salience}(C_1) - \text{salience}(C_2))
\]
Kehler et al. (2004)

- integrated as feature into MaxEnt-based pronoun resolution system
- results disillusioning, improvement of at most 1% accuracy
Kehler et al. (2004)

[...] predicate-argument statistics offer little predictive power to a pronoun interpretation system trained on a state-of-the-art set of morpho-syntactic features. [...] the distribution of pronouns in discourse allows for a system to correctly resolve a majority of them using only morphosyntactic cues. [...] predicate-argument statistics appear to provide a poor substitute for the world knowledge that may be necessary to correctly interpret the remaining cases.
(highly subjective review of research integrating semantics into coreference resolution)
This Talk

(highly subjective review of research integrating “semantics” into coreference resolution)
This is an interesting paper extending Guinaudeau & Strube's work on discourse entity graphs. It's clear and well-written, and the results are useful. I didn't notice any technical problems.

It's not clear to me that this paper is a good fit for *SEM, as readability and coherence are not semantic matters per se, and the use of discourse relations is just on the edge of (what I, at least, think of as) discourse semantics; so I'd give it a lower priority for acceptance for that reason.

Formatting issues: All the hyphens in the tables should be changed to minus signs ($-$ in Latex). A package such as mathptmx should be used so that math mode is in the same Times Roman font as the rest of the paper.
(highly subjective) review of research integrating “semantics” into coreference resolution

• distributional approaches
• semantic role labeling
• WordNet
• Wikipedia
This Talk

...to make a long story short:

• there have been quite a few attempts trying to integrate “semantics” into coreference resolution
• there has been quite a bit of progress in coreference resolution in the last few years (in terms of F-scores, not necessarily in terms of a better understanding of the problem ...)
• none of this progress can be attributed to “semantics”
“Semantics” . . .

... for coreference resolution

- the importance of semantics, world knowledge and inference, common sense knowledge has been recognized early on (Charniak (1973), Hobbs (1978), . . .)
- we reiterate these statements until today
Semantic Role Labeling . . .
Semantic Role Labeling . . .

. . . for coreference resolution (Ponzetto & Strube, 2006b)

*A state commission of inquiry into the sinking of the Kursk will convene in Moscow on Wednesday, the Interfax news agency reported. It said that the diving operation will be completed by the end of next week.*

*If the Interfax news agency is Agent of report and it is the Agent of say, it is more likely that the Interfax news agency is the antecedent of it than Moscow or the Kursk or . . .*
Semantic Role Labeling …

… for coreference resolution (Ponzetto & Strube, 2006b)

semantic role labeling:

- apply ASSERT parser (Pradhan et al., 2004)
- trained on PropBank (Palmer et al., 2005), outputs PropBank labels
- identifies all verb predicates in a sentence together with their arguments
- for ACE2003 data, 11,406 of 32,502 automatically extracted NPs were tagged with 2,801 different predicate-argument pairs
Semantic Role Labeling . . .

... for coreference resolution (Ponzetto & Strube, 2006b)

- integrate as feature (for anaphor and antecedent) into MaxEnt-based coreference resolution system (reimplementation of Soon et al. (2001))
- evaluate on ACE2003 data
- improvement over Soon et al. (2001) 1.5 points MUC F1-score mostly due to improved recall
Semantic Role Labeling . . .

. . . for coreference resolution (Ponzetto & Strube, 2006b)

• similar work by Rahman & Ng (2011)
• they use a semantic parser to label NPs with FrameNet semantic roles
• about 0.5 points (B³, CEAF) F1-score improvement
Exploiting WordNet . . .

. . . for coreference resolution (Soon et al., 2001)

semantic class agreement:

- **PERSON**
  - **MALE**
  - **FEMALE**
- **OBJECT**
  - **ORGANIZATION**
  - **LOCATION**
  - **DATE**
  - **TIME**
  - **MONEY**
  - **PERCENT**
Exploiting WordNet . . .

. . . for coreference resolution (Soon et al., 2001)

• assume that the semantic class of every markable extracted is the first WordNet sense of the head noun of the markable
• if the selected semantic class of a markable is a subclass of one of the defined semantic classes $C$, then the semantic class of the markable is $C$
• the semantic classes of anaphor and antecedent are in agreement,
  • if one is the parent of the other $chairman \rightarrow PERSON$ and $Mr. Lim \rightarrow MALE$, or
  • they are the same $Mr. Lim \rightarrow MALE$ and $he \rightarrow MALE$
• does not appear to have a positive effect on the results
Exploiting WordNet . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

\[ \text{rel}((c_1, c_2)) = \#\text{nodes in path} \]

- \[ \text{rel}(\text{car, auto}) = 1 \]
- \[ \text{rel}(\text{car, bike}) = 0.25 \]
- \[ \text{rel}(\text{car, fork}) = 0.08 \]
Exploiting WordNet . . .

. . . for coreference resolution by computing the semantic relatedness between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

\[
\text{rel}(c_1, c_2) = \#\text{nodes in path}
\]

\[
\text{rel}(\text{car}, \text{auto}) = 1
\]

\[
\text{rel}(\text{car}, \text{bike}) = 0.25
\]

\[
\text{rel}(\text{car}, \text{fork}) = 0.08
\]
Exploiting WordNet . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

e.g. node counting scheme

\[
rel(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}
\]
Exploiting WordNet . . .

. . . for coreference resolution by computing the semantic relatedness between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

e.g. node counting scheme

\[
\text{rel}(c_1, c_2) = \frac{1}{\# \text{nodes in path}}
\]

- \text{rel(car, auto)} = 1
Exploiting WordNet . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

e.g. node counting scheme

\[
\text{rel}(c_1, c_2) = \frac{1}{\# \text{ nodes in path}}
\]

- \( \text{rel}(\text{car}, \text{auto}) = 1 \)
- \( \text{rel}(\text{car}, \text{bike}) = 0.25 \)
Exploiting WordNet . . .

...for coreference resolution by computing the semantic relatedness between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

\[ \text{rel}(c_1, c_2) = \frac{1}{\# \text{nodes in path}} \]

- \( \text{rel}(\text{car}, \text{auto}) = 1 \)
- \( \text{rel}(\text{car}, \text{bike}) = 0.25 \)
- \( \text{rel}(\text{car}, \text{fork}) = 0.08 \)

e.g. node counting scheme
Exploiting WordNet . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- in addition to node counting several different measures for semantic relatedness used
- integrate these as additional features into MaxEnt-based coreference resolution system
- results on ACE 2003 data (MUC score) as reported in Ponzetto & Strube (2007):
Exploiting WordNet . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- in addition to node counting several different measures for semantic relatedness used
- integrate these as additional features into MaxEnt-based coreference resolution system
- results on ACE 2003 data (MUC score) as reported in Ponzetto & Strube (2007):

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F$_1$</th>
<th>A$_p$</th>
<th>A$_{cn}$</th>
<th>A$_{pn}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>54.5</td>
<td>85.4</td>
<td>66.5</td>
<td>40.5</td>
<td>30.1</td>
<td>73.0</td>
</tr>
<tr>
<td>+WordNet</td>
<td><strong>60.6</strong></td>
<td>79.4</td>
<td><strong>68.7</strong></td>
<td><strong>42.4</strong></td>
<td><strong>43.2</strong></td>
<td>66.0</td>
</tr>
</tbody>
</table>
Exploiting Wikipedia . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- extract knowledge from Wikipedia (in analogy to WordNet)
- create a Wikipedia-based semantic network
- map mentions to Wikipedia concepts
- compute semantic relatedness
- integrate Wikipedia-based semantic relatedness measures into MaxEnt-based coreference resolution system
- results (MUC score) as reported in Ponzetto & Strube (2007):
Exploiting Wikipedia . . .

. . . for coreference resolution by computing the **semantic relatedness**
between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- extract knowledge from Wikipedia (in analogy to WordNet)
- create a Wikipedia-based semantic network
- map mentions to Wikipedia concepts
- compute semantic relatedness
- integrate Wikipedia-based semantic relatedness measures into
  MaxEnt-based coreference resolution system
- results (MUC score) as reported in Ponzetto & Strube (2007):

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>P</th>
<th>F₁</th>
<th>Aᵰ</th>
<th>Aᵣcn</th>
<th>Aᵣpn</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>54.5</td>
<td>85.4</td>
<td>66.5</td>
<td>40.5</td>
<td>30.1</td>
<td>73.0</td>
</tr>
<tr>
<td>+WordNet</td>
<td><strong>60.6</strong></td>
<td>79.4</td>
<td><strong>68.7</strong></td>
<td><strong>42.4</strong></td>
<td><strong>43.2</strong></td>
<td>66.0</td>
</tr>
<tr>
<td>+Wikipedia</td>
<td>59.4</td>
<td>82.2</td>
<td><strong>68.9</strong></td>
<td>38.9</td>
<td><strong>41.4</strong></td>
<td>74.5</td>
</tr>
</tbody>
</table>
Exploiting Wikipedia . . .

. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- similar work by Rahman & Ng (2011)
- they use YAGO and its *type* and *means* relations
- 0.7 to 2.8 points (B³, CEAF) F1-score improvement
Recent Work ...
Lee et al. (2011, 2013): “Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules”
Sapena et al. (2011, 2013): “A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution”

see also Cai et al. (2010, 2011): “End-to-end coreference resolution via hypergraph partitioning”

Source: Sapena et al. (2013)
Sapena et al. (2011, 2013): “A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution”

Adding World Knowledge to Coreference Resolution

Source: Sapena et al. (2013)
Sapena et al. (2011, 2013): “A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution”

<table>
<thead>
<tr>
<th>measure</th>
<th>class</th>
<th>baseline</th>
<th></th>
<th>features</th>
<th></th>
<th>constraints</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre</td>
<td>Rec</td>
<td>F1</td>
<td>Pre</td>
<td>Rec</td>
<td>F1</td>
</tr>
<tr>
<td>MUC</td>
<td></td>
<td>74.4</td>
<td>59.9</td>
<td>66.4</td>
<td>75.9</td>
<td>59.6</td>
<td>66.8</td>
</tr>
<tr>
<td>CEAFm</td>
<td></td>
<td>83.0</td>
<td>83.0</td>
<td>83.0</td>
<td>83.4</td>
<td>83.4</td>
<td>83.4</td>
</tr>
<tr>
<td>$B^3$</td>
<td></td>
<td>91.8</td>
<td>84.6</td>
<td>88.1</td>
<td>92.6</td>
<td>84.5</td>
<td>88.4</td>
</tr>
</tbody>
</table>

Source: Sapena et al. (2013)
In this work, we tested a methodology that identified the real-world entities referred to in a document, extracted information about them from Wikipedia, and then incorporated this information in two different ways in the model. It seems that neither of the two forms work very well, however, and that the results and errors are in the same direction: The slight improvement of the few new relationships is offset by the added noise.
Berkeley System

Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

\[ [\text{Voters}]_1 \text{ agree when } [\text{they}]_1 \text{ are given a } [\text{chance}]_2 \text{ to decide if } [\text{they}]_1 \ldots \]

Figure 1: The basic structure of our coreference model. The \( i \)th mention in a document has \( i \) possible antecedence choices: link to one of the \( i - 1 \) preceding mentions or begin a new cluster. We place a distribution over these choices with a log-linear model. Structurally different kinds of errors are weighted differently to optimize for final coreference loss functions; error types are shown corresponding to the decisions for each mention.

Source: Durrett & Klein (2013)
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“Easy Victories from Surface Features”:

- surface features (mention type, mention string, mention head, first and last word of mention, the word immediately preceding and immediately following the mention, mention length, distance)
- feature conjunctions
Berkeley System

Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

\[
\begin{align*}
\text{MENT DIST} &= 1 \\
\text{MENT DIST} &= 1 \land [they] \\
\text{MENT DIST} &= 1 \land [they] \land \text{NOM} \\
\text{ANT. HEAD} &= \text{Voters} \\
\text{ANT. HEAD} &= \text{Voters} \land [they] \\
\text{ANT. HEAD} &= \text{Voters} \land [they] \land \text{NOM} \\
\text{NEW} \land \text{LEN} &= 1 \\
\text{NEW} \land \text{LEN} &= 1 \land [they] \\
\\text{NEW} &\rightarrow \text{a}_2 \\
\text{[Voters]}_1 \text{ generally agree when [they]}_1 ... \\
\end{align*}
\]

Source: Durrett & Klein (2013)
Berkeley System

Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“Easy Victories from Surface Features”:

• surface features (mention type, mention string, mention head, first and last word of mention, the word immediately preceding and immediately following the mention, mention length, distance)
• feature conjunctions
• data-driven features capturing linguistic intuitions at a fine level of granularity
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features of the SURFACE system</td>
<td>418704</td>
</tr>
<tr>
<td>Features on the current mention</td>
<td></td>
</tr>
<tr>
<td>[ANAPHORIC] + [CURRENT ANCESTRY]</td>
<td>46047</td>
</tr>
<tr>
<td>Features on the antecedent</td>
<td></td>
</tr>
<tr>
<td>[ANTECEDENT ANCESTRY]</td>
<td>53874</td>
</tr>
<tr>
<td>[ANTECEDENT GENDER]</td>
<td>338</td>
</tr>
<tr>
<td>[ANTECEDENT NUMBER]</td>
<td>290</td>
</tr>
<tr>
<td>Features on the pair</td>
<td></td>
</tr>
<tr>
<td>[HEAD CONTAINED (T/F)]</td>
<td>136</td>
</tr>
<tr>
<td>[EXACT STRING CONTAINED (T/F)]</td>
<td>133</td>
</tr>
<tr>
<td>[NESTED (T/F)]</td>
<td>355</td>
</tr>
<tr>
<td>[DOC TYPE] + [SAME SPEAKER (T/F)]</td>
<td>437</td>
</tr>
<tr>
<td>[CURRENT ANCESTRY] + [ANT. ANCESTRY]</td>
<td>2555359</td>
</tr>
</tbody>
</table>

Source: Durrett & Klein (2013)
Berkeley System

Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“Uphill Battles on Semantics”

“semantic” features:

- WordNet hypernymy and synonymy
- number and gender for common nouns and proper names
- named entity types
- latent Gigaword clusters, e.g. president and leader, i.e. things which announce
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“Uphill Battles on Semantics”

The main reason that weak semantic cues are not more effective is the small fraction of positive coreference links present in the training data. ... Our weak cues do yield some small gains, so there is hope that better weak indicators of semantic compatibility could prove more useful. ... we conclude that capturing semantics in a data-driven, shallow manner remains an uphill battle.
Berkeley System


- integrate knowledge into coreference resolution system by linking mentions to entities in a knowledge base
- integrate coreference resolution into entity linking system
- does not appear to have positive effect on coreference resolution

- ranking model outperforms mention pair model by large margin (identical systems, just different latent structures)
- no sophisticated semantic features
- state-of-the-art results (1% improvement over Durrett & Klein (2013), Björkelund & Kuhn (2014), 2% improvement over Fernandes et al. (2014))
- any attempt to integrate semantic or world knowledge resulted in failure (gains in recall offset by loss in precision)
Conclusions

...to make a long story short:

• there have been quite a few attempts trying to integrate “semantics” into coreference resolution
• there has been quite a bit of progress in coreference resolution in the last few years (in terms of F-scores, not necessarily in terms of a better understanding of the problem . . . )
• none of this progress can be attributed to “semantics”
Conclusions

...to make a long story short:

• earlier (slight) successes in integrating “semantics” into coreference resolution could not be replicated in recent work
• systems are better, it is much more difficult to make improvements
• progress is due to better mention detection, preprocessing, Berkeley-style features, and, in particular, better algorithms and architectures
Conclusions
Conclusions

• forget about “semantics”
• go to a math class
• study algorithms
Conclusions

• forget about “semantics”
• go to a math class
• study algorithms

Thank You!
References

Learning structured perceptrons for coreference resolution with latent antecedents and non-local features.

Unrestricted coreference resolution via global hypergraph partitioning.

Cai, Jie & Michael Strube (2010).
End-to-end coreference resolution via hypergraph partitioning.

Jack and Janet in search of a theory of knowledge.

Syntax and lexical statistics in anaphora resolution.

Easy victories and uphill battles in coreference resolution.

A joint model for entity analysis: Coreference, typing, and linking.

Fernandes, Eraldo, Cícero dos Santos & Ruy Milidiú (2014).
Latent trees for coreference resolution.

Resolving pronomial references.


Rahman, Altaf & Vincent Ng (2011).
Coreference resolution with world knowledge.

RelaxCor participation in CoNLL shared task on coreference resolution.

A constraint-based hypergraph partitioning approach to coreference resolution.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001).
A machine learning approach to coreference resolution of noun phrases.