Heidelberg Institute for
Theoretical Studies

The (Non)Utility of Semantics for Coreference Resolution

(CORBON Remix)

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

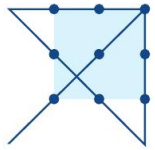
State-of-the-art pronoun interpretation systems rely predominantly on morphosyntactic contextual features. While the use of deep knowledge and inference to improve these models would appear technically infeasible, previous work has suggested that predicate-argument statistics mined from naturally-occurring 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.

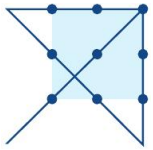
In lieu of this capability, previous authors have suggested that what can be viewed as a more superficial form of semantic information – predicate-argument statistics mined from naturally-occurring data – could be used to capture certain selectional regularities. For instance, such statistics might reveal 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?

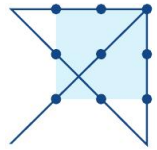
Kehler et al. (2004)



*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)



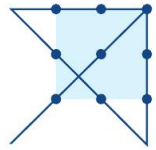
predicate-argument frequencies:

- 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))

$$stat(C) = P(tuple(C, A)|C) = \frac{freq(tuple(C, A))}{freq(C)}$$

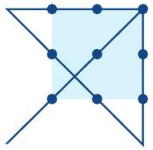
$$ln(\frac{stat(C_2)}{stat(C_1)} > K \times (salience(C_1) - salience(C_2))$$

Kehler et al. (2004)



- integrated as feature into MaxEnt-based pronoun resolution system
- results disillusioning, improvement of at most 1% accuracy

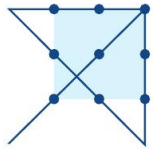
Kehler et al. (2004)



[. . .] 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.

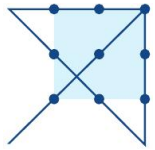
Kehler et al. (2004, p.296)

This Talk



(highly subjective review of research integrating semantics into coreference resolution

This Talk



(highly subjective review of research integrating “semantics” into coreference resolution

Soundness/Correctness: 4
Impact: 4
Substance: 4
Replicability: 4
Overall Score: 4
Reviewer Confidence: 3

Comments

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.

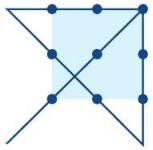
It's not clear to me that this paper is a good fit for *SEM, as readability and coherence are not semantic matters per se, and the use of discourse relations is just on the edge of (what I, at least, think of as) discourse semantics; so I'd give it a lower priority for acceptance for that reason.

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.

==========

REVIEWER #2

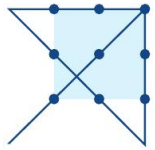
This Talk



(highly subjective) review of research integrating “semantics” into coreference resolution

- distributional approaches
- semantic role labeling
- WordNet
- Wikipedia (Yago, DBpedia, Freebase, . . .)

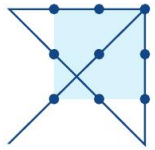
This Talk



. . . 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”**

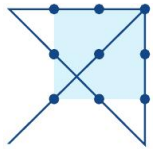
“Semantics” . . .



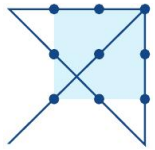
. . . for coreference resolution

- 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 . . .

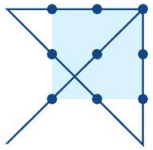


. . . for coreference resolution (Ponzetto & Strube, 2006b)

> ***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 . . .

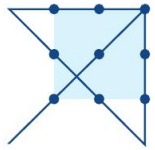


. . . for coreference resolution (Ponzetto & Strube, 2006b)

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

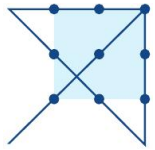
Semantic Role Labeling . . .



. . . 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

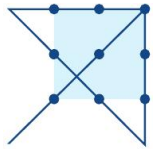
Semantic Role Labeling . . .



. . . 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 (B^3, CEAF) F1-score improvement

Exploiting WordNet . . .

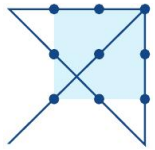


. . . for coreference resolution (Soon et al., 2001)

semantic class agreement:

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

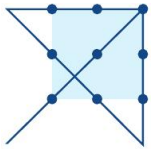
Exploiting WordNet . . .



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

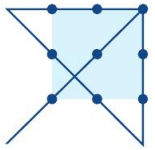
- assume that the semantic class of every markable extracted is the first WordNet sense of the head noun of the markable
- if the selected semantic class of a markable is a subclass of one of the defined semantic classes *C*, then the semantic class of the markable is *C*
- the semantic classes of anaphor 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 . . .

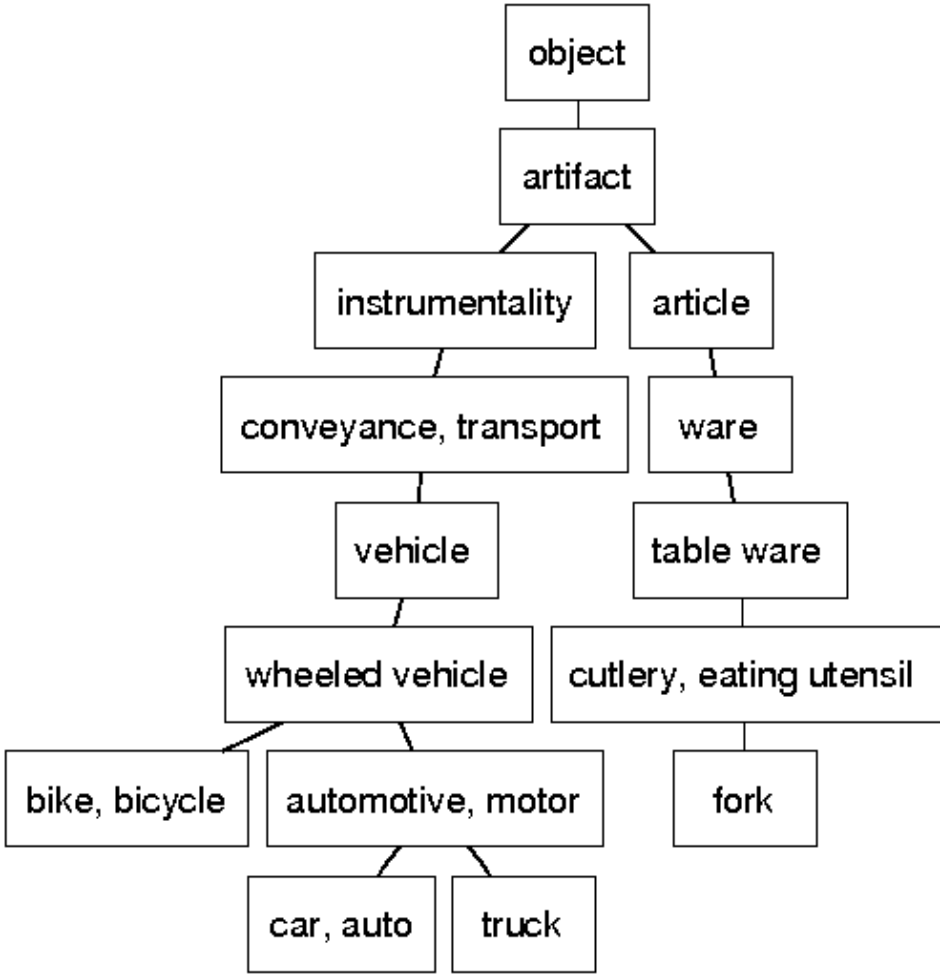


. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

Exploiting WordNet . . .

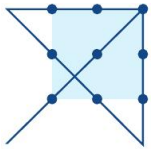


. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

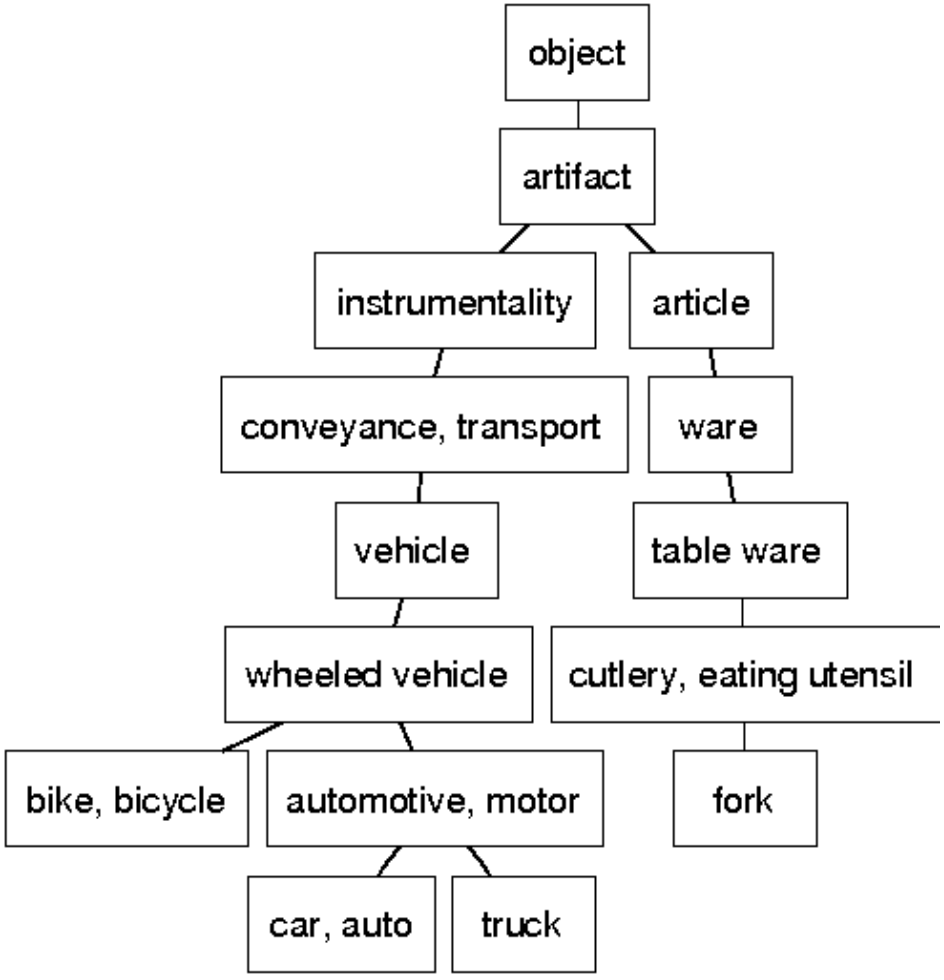


compute relatedness

Exploiting WordNet . . .



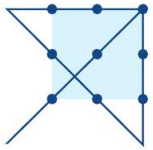
. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)



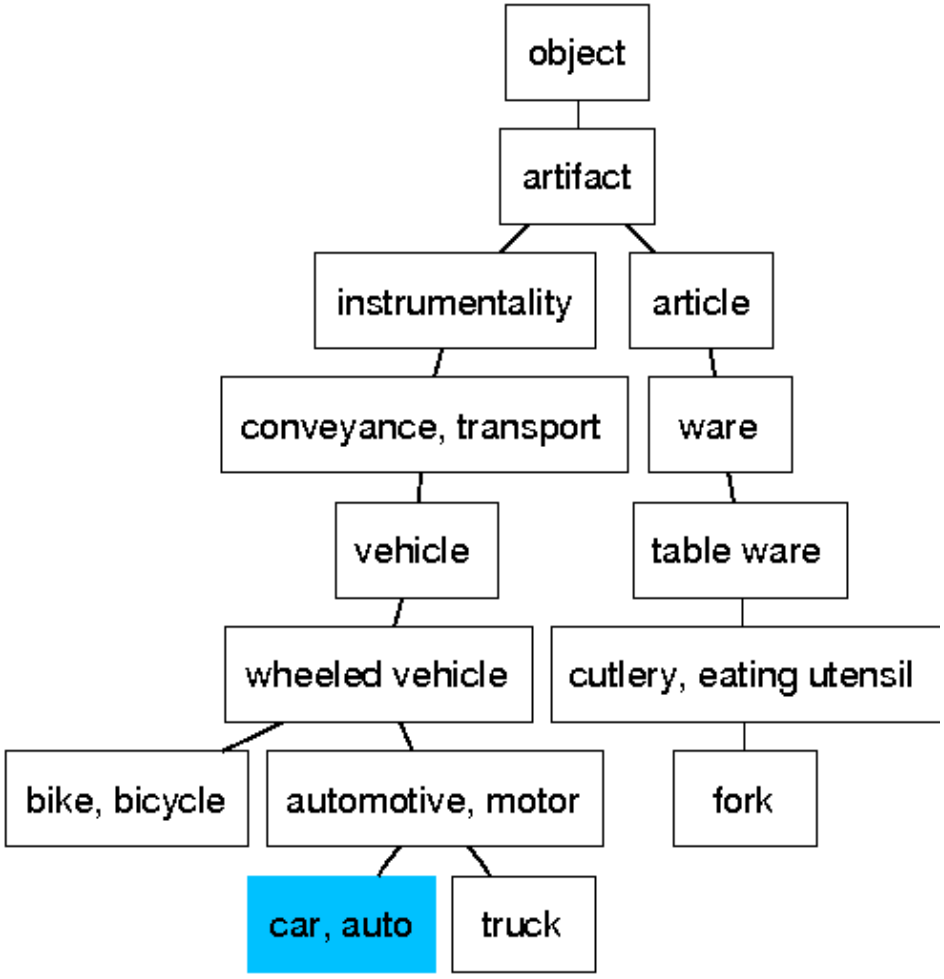
e.g. *node counting scheme*

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

Exploiting WordNet . . .



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

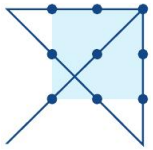


e.g. *node counting scheme*

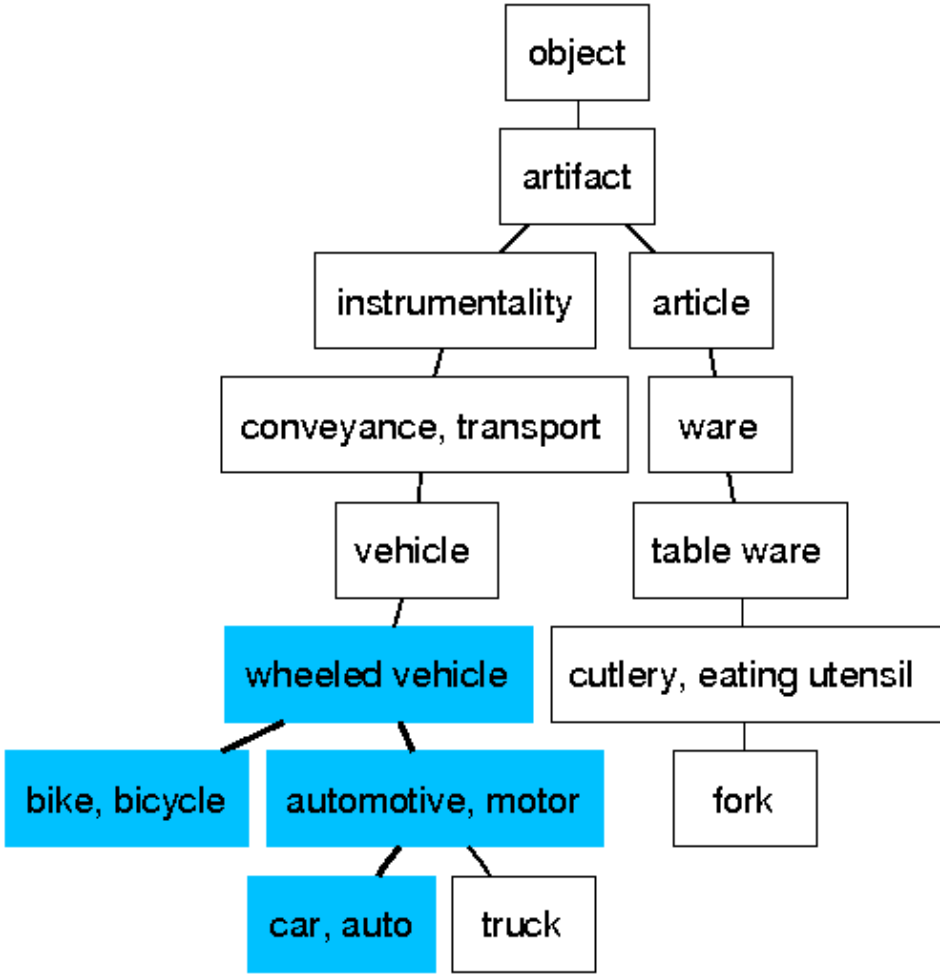
$$rel(c_1, c_2) = \frac{1}{\#\ nodes\ in\ path}$$

- $rel(car, auto) = 1$

Exploiting WordNet . . .



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

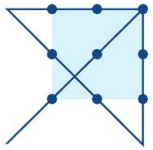


e.g. *node counting scheme*

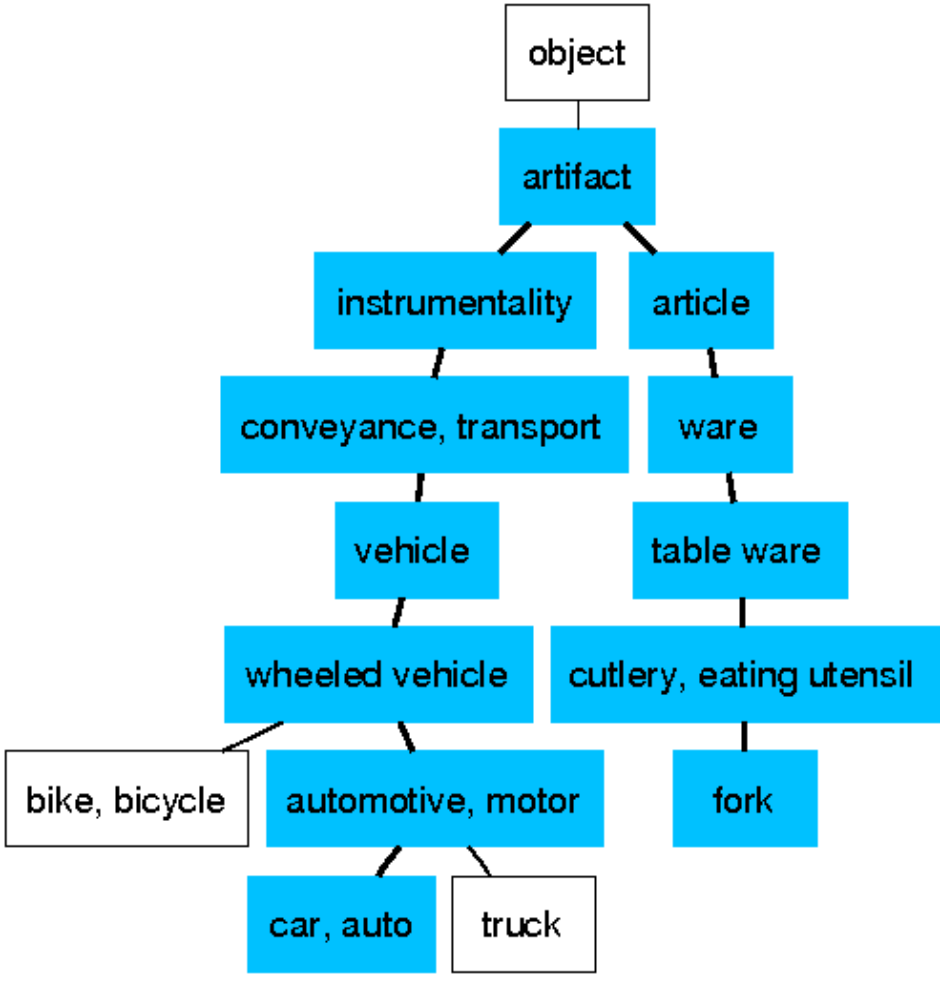
$$rel(c_1, c_2) = \frac{1}{\#\ nodes\ in\ path}$$

- $rel(car, auto) = 1$
- $rel(car, bike) = 0.25$

Exploiting WordNet . . .



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

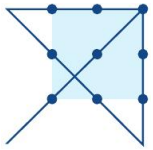


e.g. *node counting scheme*

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

- $rel(car, auto) = 1$
- $rel(car, bike) = 0.25$
- $rel(car, fork) = 0.08$

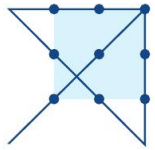
Exploiting WordNet . . .



. . . 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 . . .

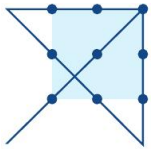


. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- 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):

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	85.4	66.5	40.5	30.1	73.0
+WordNet	**60.6**	79.4	**68.7**	**42.4**	**43.2**	66.0

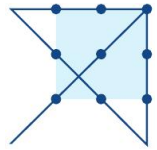
Exploiting Wikipedia . . .



. . . 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 . . .

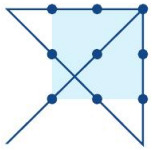


. . . for coreference resolution by computing the **semantic relatedness** between anaphor and antecedent (Ponzetto & Strube, 2006, 2007)

- 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):

	R	P	F_1	A_p	A_{cn}	A_{pn}
baseline	54.5	85.4	66.5	40.5	30.1	73.0
+WordNet	**60.6**	79.4	**68.7**	**42.4**	**43.2**	66.0
+Wikipedia	**59.4**	82.2	**68.9**	38.9	**41.4**	**74.5**

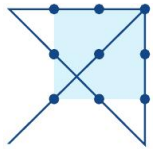
Exploiting Wikipedia . . .



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

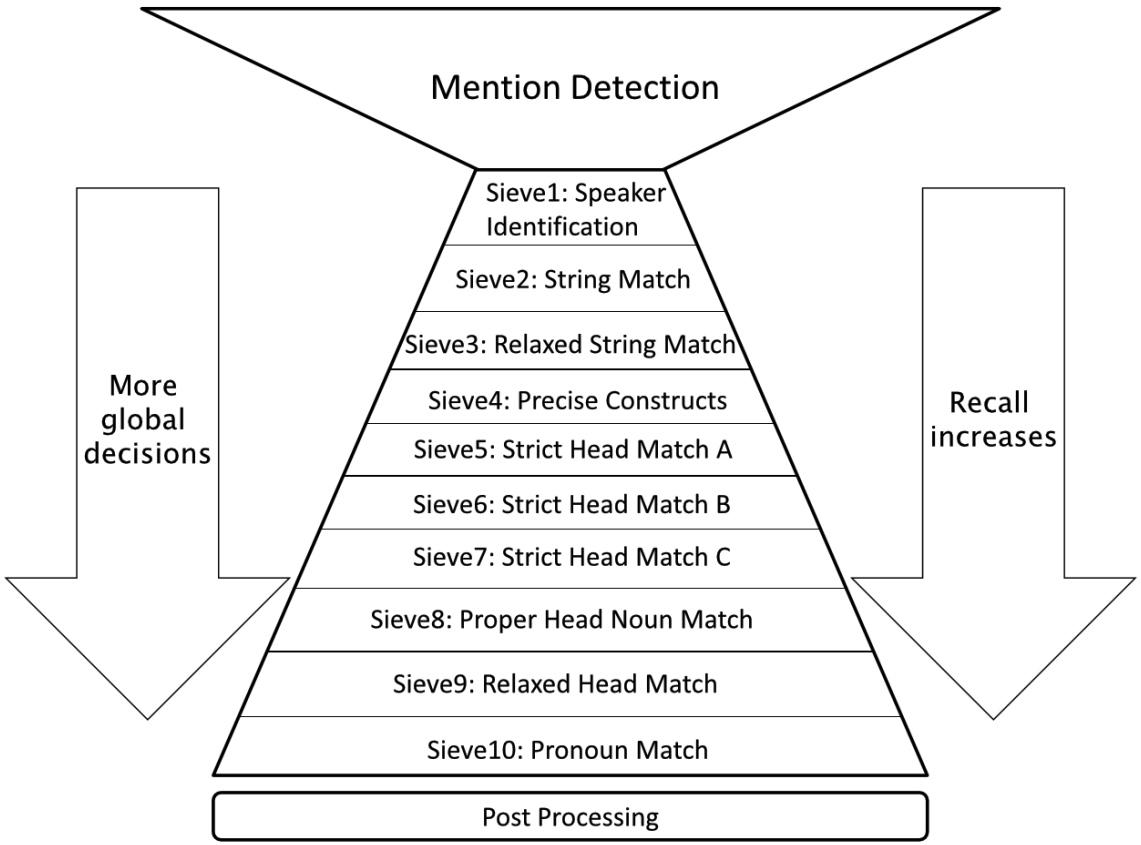
- similar work by Rahman & Ng (2011)
- they use YAGO and its *type* and *means* relations
- 0.7 to 2.8 points (B^3, CEAF) F1-score improvement

Recent Work . . .



Stanford System

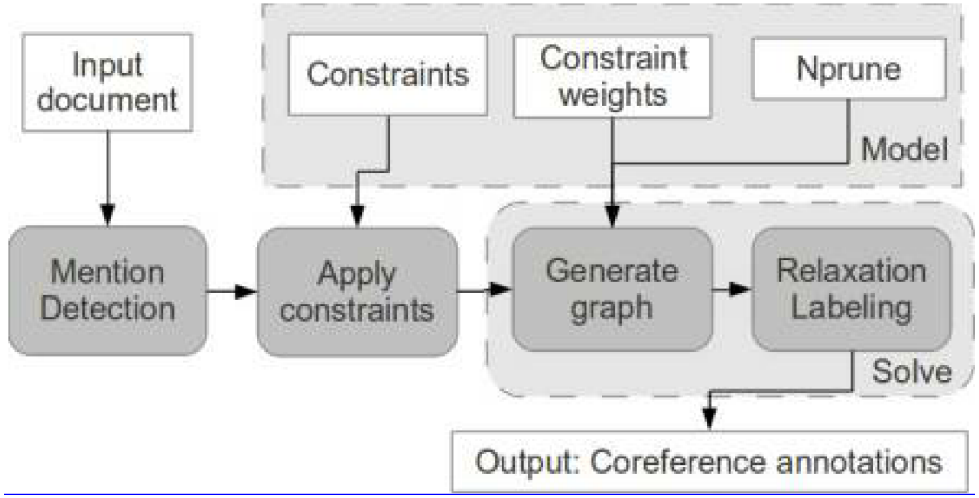
Lee et al. (2011, 2013): “Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules”



Source: Lee et al. (2013)

RelaxCor

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



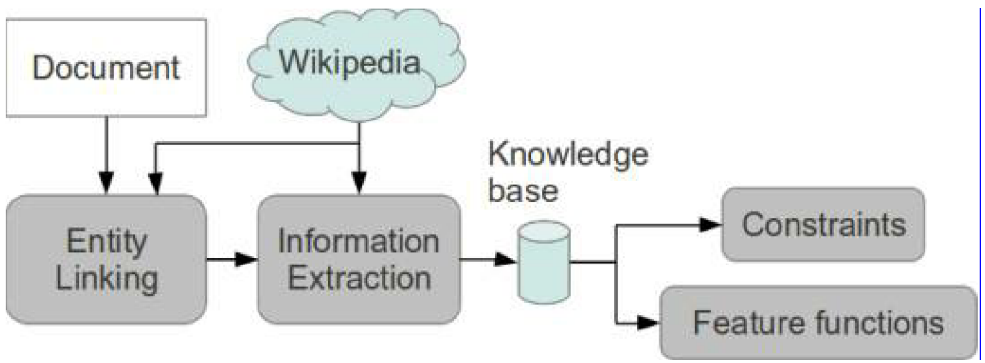
see also Cai et al. (2010, 2011): “End-to-end coreference resolution via hypergraph partitioning”

Source: Sapena et al. (2013)

RelaxCor

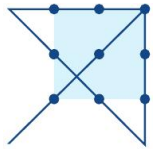
Sapena et al. (2011, 2013): “A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution”

Adding World Knowledge to Coreference Resolution



Source: Sapena et al. (2013)

RelaxCor

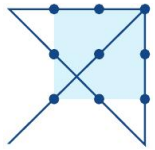


Sapena et al. (2011, 2013): “A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution”

		baseline			features			constraints		
measure	class	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
MUC		74.4	59.9	66.4	**75.9**	*59.6*	**66.8**	**75.4**	**60.3**	**67.0**
CEAFm		83.0	83.0	83.0	**83.4**	**83.4**	**83.4**	**83.5**	**83.5**	**83.5**
B^3		91.8	84.6	88.1	**92.6**	*84.5*	**88.4**	**92.3**	**84.7**	**88.4**

Source: Sapena et al. (2013)

RelaxCor



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

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.***

Sapena et al. (2013)

Berkeley System

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

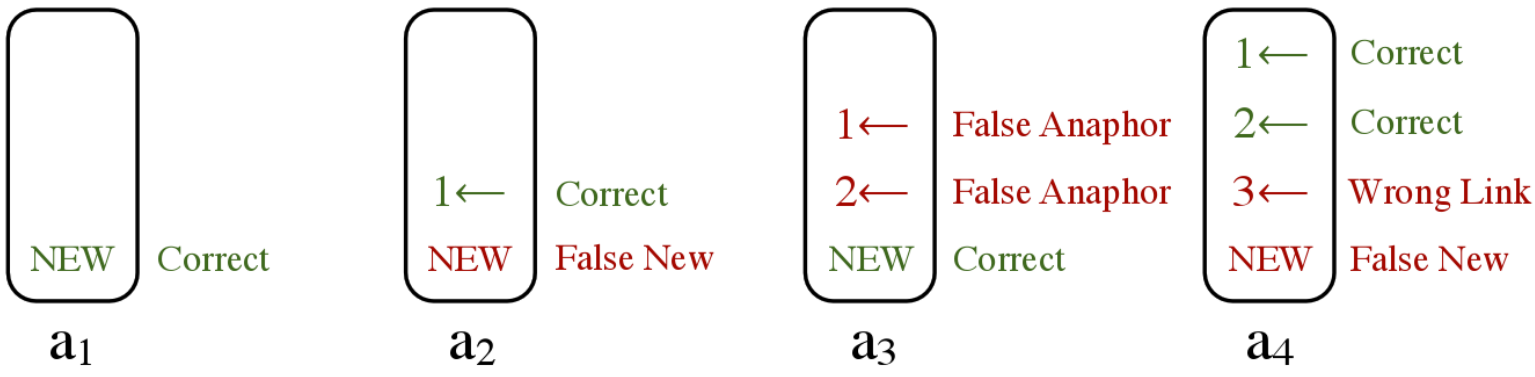
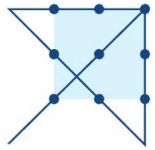


Figure 1: The basic structure of our coreference model. The ith mention in a document has i possible antecedence choices: link to one of the $i-1$ preceding mentions or begin a new cluster. We place a distribution over these choices with a log-linear model. Structurally different kinds of errors are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

Source: Durrett & Klein (2013)

Berkeley System

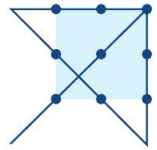


Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

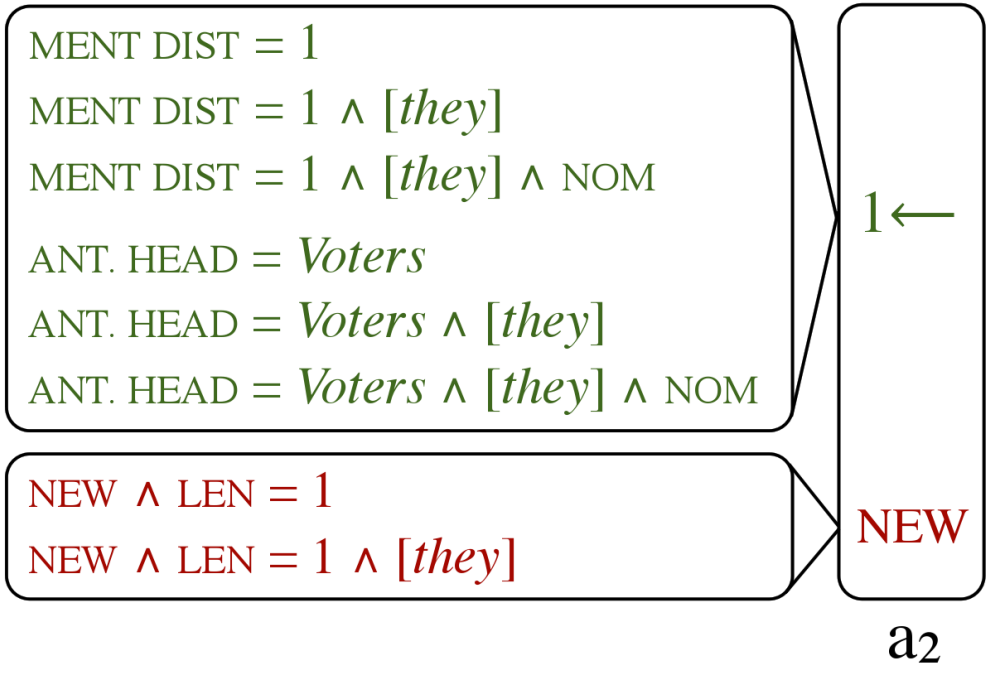
“Easy Victories from Surface Features”:

- 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

Berkeley System



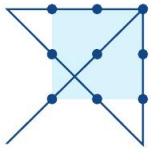
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”



[*Voters*]$_1$ *generally agree when* [*they*]$_1$ ...

Source: Durrett & Klein (2013)

Berkeley System

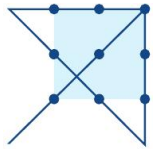


Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“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

Berkeley System

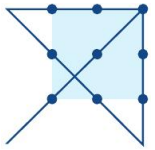


Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

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

Source: Durrett & Klein (2013)

Berkeley System



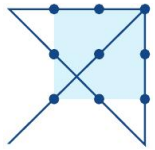
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“Uphill Battles on Semantics”

“semantic” features:

- WordNet hypernymy and synonymy
- number and gender for common nouns and proper names
- named entity types
- ancestry of each mention head (dependency paths)
- latent Gigaword clusters, e.g. *president* and *leader*, i.e. things which *announce*

Berkeley System



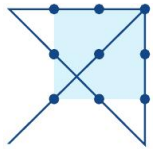
Durrett & Klein (2013): “Easy Victories and Uphill Battles in Coreference Resolution”

“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.*

Source: Durrett & Klein (2013)

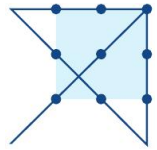
Berkeley System



Durrett & Klein (2014): “A Joint Model for Entity Analysis: Coreference, Typing, and Linking”

- 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

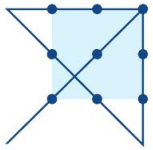
CORT



Martschat & Strube (2015, TACL)

- 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)

YASS



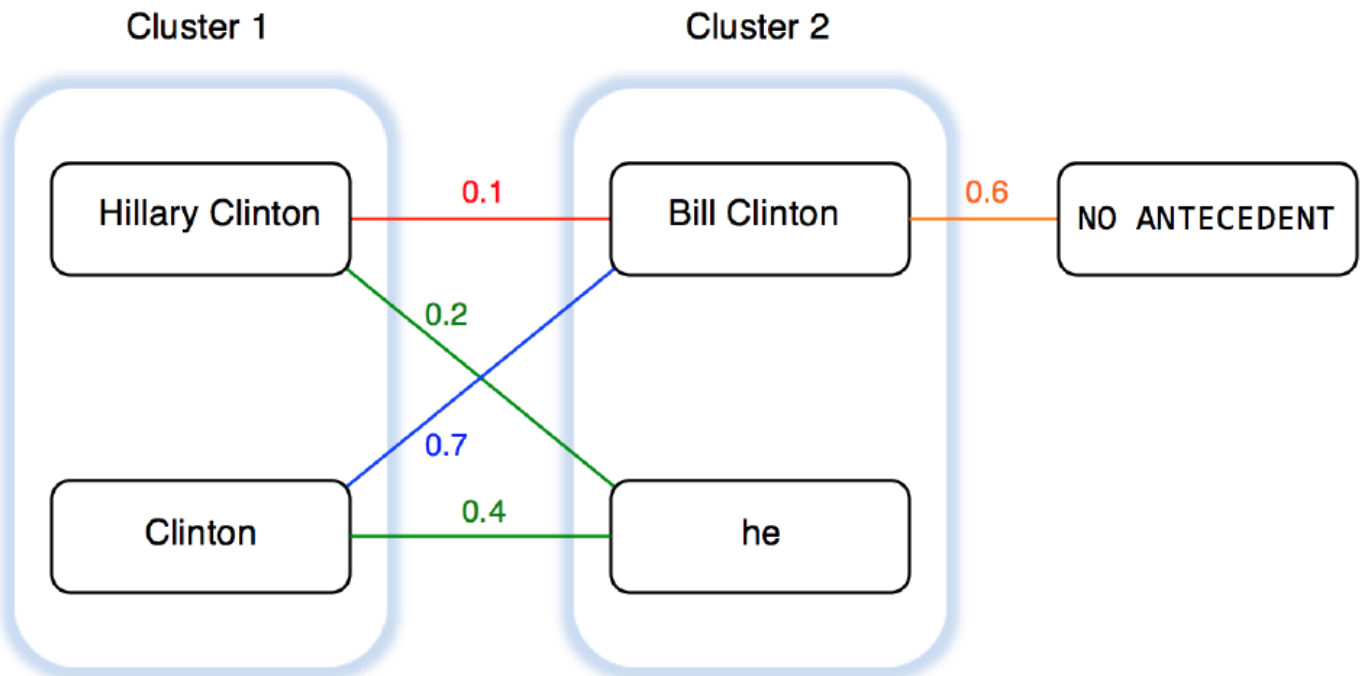
Clark & Manning (2015, ACL): “Entity-Centric Coreference Resolution with Model Stacking”

- aggregates scores produced by mention-pair model and takes these as entity-level features

YASS

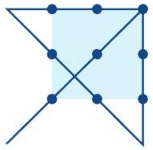
Clark & Manning (2015, ACL): “Entity-Centric Coreference Resolution with Model Stacking”

- aggregates scores produced by mention-pair model and takes these as entity-level features



Source: Clark & Manning (2015)

YASS

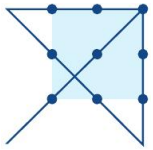


Clark & Manning (2015, ACL): “Entity-Centric Coreference Resolution with Model Stacking”

- aggregates scores produced by mention-pair model and takes these as entity-level features

- pairwise features:
  - distance features . . .
  - syntactic features . . .
  - semantic features, e.g., named entity type, speaker identification,
  - rule-based features, e.g., exact and partial string matching,
  - lexical features, e.g., the first, last, and head word of the current mention.

YASS

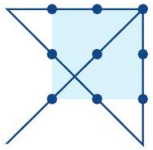


Clark & Manning (2015, ACL): “Entity-Centric Coreference Resolution with Model Stacking”

- aggregates scores produced by mention-pair model and takes these as entity-level features

- very similar to Nicolae & Nicolae (2006):”BestCut: A Graph Algorithm for Coreference Resolution”

Neural . . . Deep . . .

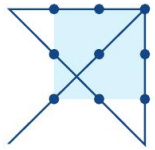


Wiseman et al. (2015, ACL): “Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution”

- mention-ranking model
- learns anaphoricity detection and antecedent ranking together
- uses simple unconjoined features as input
- neural network learns feature representations

- mention and pairwise raw features
- “semantic” features: entity type, animacy, gender, same speaker,

Neural . . . Deep . . . 2nd

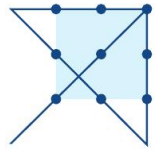


Wiseman et al. (2016, NAACL): “Learning Global Features for Coreference Resolution”

- basically the same as before, but learns global features

- “semantics” creeps in implicitly through a structured representation

Side Note: Evaluation

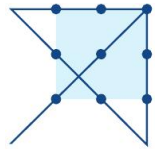


Moosavi & Strube (2016, ACL): “Which Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric”

Problem:

- precision and recall reported by CEAF are sometimes counterintuitive
- difference in scores between MUC, B^3 and CEAF cannot be easily interpreted
- these metrics are quite useless for system development
- MUC, B^3, CEAF are not quite reliable (and neither is BLANC) – dependent on mention identification
- **averaging three unreliable scores, however, does not result in a reliable one**

Side Note: Evaluation

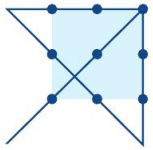


Moosavi & Strube (2016, ACL): “Which Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric”

Solution:

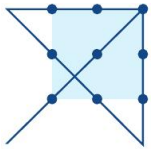
- link-based entity aware metric
- models entity awareness by notion of importance
- handles singletons by self-links
- results in a ranking quite similar to CoNLL score (on CoNLL 2012 results)
- overcomes shortcomings of MUC, B^3, CEAF
- can be used to tune precision and recall
- branch `LEA-scorer` in reference implementation of CoNLL-scorer

Side Note: Data



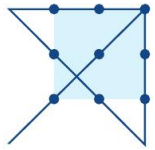
- there is the danger that we overfit to CoNLL/OntoNotes data (e.g. through Berkeley-style lexicalized features)
- also: our systems are quite dependent on annotation guidelines of a particular project (e.g. mention definition)
- I encourage you to evaluate on different datasets
  however, don't throw away OntoNotes:
  OntoNotes is cool
- evaluate extrinsically

Side Note: Data



- there is the danger that we overfit to CoNLL/OntoNotes data (e.g. through Berkeley-style lexicalized features)
- also: our systems are quite dependent on annotation guidelines of a particular project (e.g. mention definition)
- I encourage you to evaluate on different datasets
  however, don't throw away OntoNotes:
  OntoNotes is **Coooooooooooooolllllllllllll!!!!!!!!!!!!!**
- also: go beyond English, go beyond plain old entity coreference (do bridging, event coref, metonymy, . . .), add annotation layers to OntoNotes
- evaluate extrinsically

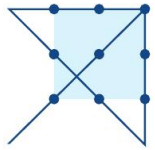
Conclusions



. . . 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”

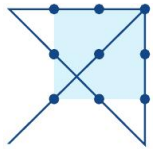
Conclusions



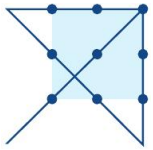
. . . to make a long story short:

- 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** – and not necessarily deeper – **algorithms and architectures**

Conclusions

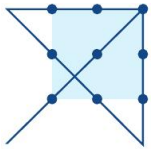


Conclusions



- forget about “semantics”
- go to a maths class
- study algorithms

Conclusions



- forget about “semantics”
- go to a maths class
- study algorithms

Thank You!

References

Björkelund, Anders & Jonas Kuhn (2014).
Learning structured perceptrons for coreference resolution with latent antecedents and non-local features.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Baltimore, Md., 22–27 June 2014, pp. 47–57.

Cai, Jie, Éva Mújdricza-Maydt & Michael Strube (2011).
Unrestricted coreference resolution via global hypergraph partitioning.
In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pp. 56–60.

Cai, Jie & Michael Strube (2010).
End-to-end coreference resolution via hypergraph partitioning.
In *Proceedings of the 23rd International Conference on Computational Linguistics,* Beijing, China, 23–27 August 2010, pp. 143–151.

Charniak, Eugene (1973).
Jack and Janet in search of a theory of knowledge.
In *Advance Papers from the Third International Joint Conference on Artificial Intelligence, Stanford, Cal.*, pp. 337–343. Los Altos, Cal.: W. Kaufmann.

Clark, Kevin & Christopher D. Manning (2015).
Entity-centric coreference resolution with model stacking.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pp. 1405–1415.

Dagan, Ido, John Justeson, Shalom Lappin, Herbert Leass & Ammon Ribak (1995).
Syntax and lexical statistics in anaphora resolution.
Applied Artificial Intelligence, 9(6):633–644.

Durrett, Greg & Dan Klein (2013).
Easy victories and uphill battles in coreference resolution.
In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* Seattle, Wash., 18–21 October 2013, pp. 1971–1982.

Durrett, Greg & Dan Klein (2014).
A joint model for entity analysis: Coreference, typing, and linking.
Transactions of the Association of Computational Linguistics, 2:477–490.

Fernandes, Eraldo Rezende, Cícero Nogueira dos Santos & Ruy Luiz Milidiú (2014).

Latent trees for coreference resolution.
Computational Linguistics, 40(4):801–835.

Hobbs, Jerry R. (1978).
Resolving pronominal references.
Lingua, 44:311–338.

Kehler, Andrew, Douglas Appelt, Lara Taylor & Aleksandr Simma (2004).
The (non)utility of predicate-argument frequencies for pronoun interpretation.
In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pp. 289–296.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky (2013).
Deterministic coreference resolution based on entity-centric, precision-ranked rules.
Computational Linguistics, 39(4):885–916.

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky (2011).
Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task.
In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pp. 28–34.

Martschat, Sebastian & Michael Strube (2015).
Latent structures for coreference resolution.
Transactions of the Association for Computational Linguistics, 3:405–418.

Moosavi, Nafise Sadat & Michael Strube (2016).
Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Berlin, Germany, 7–12 August 2016.
To appear.

Nicolae, Cristina & Gabriel Nicolae (2006).
BestCut: A graph algorithm for coreference resolution.
In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July 2006, pp. 275–283.

Palmer, Martha, Daniel Gildea & Paul Kingsbury (2005).
The proposition bank: An annotated corpus of semantic roles.
Computational Linguistics, 31(1):71–105.

Ponzetto, Simone Paolo & Michael Strube (2006a).
Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution.

In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* New York, N.Y., 4–9 June 2006, pp. 192–199.

Ponzetto, Simone Paolo & Michael Strube (2006b).
Semantic role labeling for coreference resolution.
In *Companion Volume to the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 143–146.

Ponzetto, Simone Paolo & Michael Strube (2007).
Knowledge derived from Wikipedia for computing semantic relatedness.
Journal of Artificial Intelligence Research, 30:181–212.

Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H. Martin & Dan Jurafsky (2004).
Shallow semantic parsing using Support Vector Machines.
In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May 2004, pp. 233–240.

Rahman, Altaf & Vincent Ng (2011).
Coreference resolution with world knowledge.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Portland, Oreg., 19–24 June 2011, pp. 814–824.

Sapena, Emili, Lluís Padró & Jordi Turmo (2011).
RelaxCor participation in CoNLL shared task on coreference resolution.
In *Proceedings of the Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, pp. 35–39.

Sapena, Emili, Lluís Padró & Jordi Turmo (2013).
A constraint-based hypergraph partitioning approach to coreference resolution.
Computational Linguistics, 39(4):847–884.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001).
A machine learning approach to coreference resolution of noun phrases.
Computational Linguistics, 27(4):521–544.

Wiseman, Sam, Alexander M. Rush & Stuart Shieber (2016).
Learning global features for coreference resolution.
In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, Cal., 12–17 June 2016.
To appear.

Wiseman, Sam, Alexander M. Rush, Stuart Shieber & Jason Weston (2015).

Learning anaphoricity and antecedent ranking features for coreference resolution.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pp. 1416–1426.