The Role of Salience Ranking in Anaphora Resolution

Scott Martin
http://coffeeblack.org/

Natural Language Understanding and Artificial Intelligence Laboratory
Nuance Communications

Workshop on Logic and Probabilistic Methods for Dialog
ESSLLI, Barcelona
August 14, 2015

A pervasive aspect of natural language

- Anaphora is so pervasive that people start to use it as soon as they can

A pervasive aspect of natural language

- Anaphora is so pervasive that people start to use it as soon as they can
- If you're not sure, read the first bullet again

A pervasive aspect of natural language

- Anaphora is so pervasive that people start to use it as soon as they can
- If you're not sure, read the first bullet again
- Pronouns and definites (like *the bike I used to have*) are the paradigm cases, but anaphora also occurs with iterative adverbs (*too*, *again*), events, states, questions, and topics of discussion

A pervasive aspect of natural language

- Anaphora is so pervasive that people start to use it as soon as they can
- If you're not sure, read the first bullet again
- Pronouns and definites (like *the bike I used to have*) are the paradigm cases, but anaphora also occurs with iterative adverbs (*too*, *again*), events, states, questions, and topics of discussion
- Giving machines the ability to resolve anaphora would greatly help in many natural language processing (NLP) applications

A pervasive aspect of natural language

- Anaphora is so pervasive that people start to use it as soon as they can
- If you're not sure, read the first bullet again
- Pronouns and definites (like *the bike I used to have*) are the paradigm cases, but anaphora also occurs with iterative adverbs (*too*, *again*), events, states, questions, and topics of discussion
- Giving machines the ability to resolve anaphora would greatly help in many natural language processing (NLP) applications
- Anaphora is especially important (I would argue) for dialog, because ideally systems are built to deal with multi-utterance situations

Multiple angles of attack

The two main types of approaches to computational anaphora resolution:

Theoretical/Algorithmic approaches target specific theoretical principles associated with the phenomenon of anaphora (Bos, 2003; van Eijck and Unger, 2010; Lee et al., 2013)

Statistical approaches attempt to use more shallow surface string features as proxies for complex, theory-based principles (Denis and Baldridge, 2008; Durrett and Klein, 2013)

Multiple angles of attack

The two main types of approaches to computational anaphora resolution:

Theoretical/Algorithmic approaches target specific theoretical principles associated with the phenomenon of anaphora (Bos, 2003; van Eijck and Unger, 2010; Lee et al., 2013)

Statistical approaches attempt to use more shallow surface string features as proxies for complex, theory-based principles (Denis and Baldridge, 2008; Durrett and Klein, 2013)

In this talk, I'll discuss how a system for resolving anaphora ideally should leverage elements of both of these kinds of approach

Making the right choice

Number/Gender It was the CEO's public comments about the board that offended them, not so much her failure to attend the quarterly meeting.

Binding He has a movie about David Letterman where another actor plays him but Johnny Carson plays himself.

That's just historical

One challenge in anaphora resolution is the need to keep around an arbitrarily long discourse history.

A Would you like *The Book of Eli* or *The Town*?

U I would like the horror movie.

A You chose *The Town* on ~~Turner Network Television~~.

U I'll watch that one. Who is ~~the star~~?

A Some possible answers are: Ben Affleck and Rebecca Hall.

U Where was she born?

A ~~Berkeley, CA~~.

U When is it on?

A 7:15 PM.

U I'll watch it then.

That's just historical

One challenge in anaphora resolution is the need to keep around an arbitrarily long discourse history.

A Would you like *The Book of Eli* or *The Town*?

U I would like the horror movie.

A You chose *The Town* on ~~Turner Network Television~~.

U I'll watch that one. Who is ~~the star~~?

A Some possible answers are: Ben Affleck and Rebecca Hall.

U Where was she born?

A ~~Berkeley, CA~~.

U When is it on?

A 7:15 PM.

U I'll watch it then.

The more history, the more distractors are present to complicate the task.

No you're never gonna get it

Famously, not all potential antecedents are accessible for later anaphoric reference.

U Find a movie with James Franco.

A There are several episodes of *Freaks and Geeks* on TBS, but no ~~movies starring him~~.

U Ok, I'll watch one of those.

Knowledge is power

Basic lexical information There were a lot of Tour de France riders staying at our hotel. Several of the athletes even ate in the hotel restaurant.

Knowledge is power

Basic lexical information There were a lot of Tour de France riders staying at our hotel. Several of the athletes even ate in the hotel restaurant.

World knowledge She was staying at the Ritz, but even that hotel didn't offer dog walking service.

Knowledge is power

Basic lexical information There were a lot of Tour de France riders staying at our hotel. Several of the athletes even ate in the hotel restaurant.

World knowledge She was staying at the Ritz, but even that hotel didn't offer dog walking service.

Entailments There was ~~one woman with a dog~~ and another woman without a dog in the elevator. The woman without a dog decided to start a conversation.

Knowledge is power

Basic lexical information There were a lot of Tour de France riders staying at our hotel. Several of the athletes even ate in the hotel restaurant.

World knowledge She was staying at the Ritz, but even that hotel didn't offer dog walking service.

Entailments There was ~~one woman with a dog~~ and another woman without a dog in the elevator. The woman without a dog decided to start a conversation.

Bridging Kim really likes a book her aunt gave her for Christmas, even though she detests the author.

When constraints aren't enough

Unfortunately, all of these factors still don't uniquely determine which antecedent should be chosen.

A Your schedule is Thai Basil in Sunnyvale at 6:30 PM, followed by Dawn of the Planet of the Apes at 8:45 PM at the Mercado theater in Santa Clara.

U Where is it?

When constraints aren't enough

Unfortunately, all of these factors still don't uniquely determine which antecedent should be chosen.

A Your schedule is Thai Basil in Sunnyvale at 6:30 PM, followed by Dawn of the Planet of the Apes at 8:45 PM at the Mercado theater in Santa Clara.

U Where is it?

Another very important aspect is *salience*, roughly, an antecedent's relative likelihood for a given anaphor in a given discourse context, other things (like the constraints just discussed) being equal.

When constraints aren't enough

Unfortunately, all of these factors still don't uniquely determine which antecedent should be chosen.

A Your schedule is Thai Basil in Sunnyvale at 6:30 PM, followed by Dawn of the Planet of the Apes at 8:45 PM at the Mercado theater in Santa Clara.

U Where is it?

Another very important aspect is *salience*, roughly, an antecedent's relative likelihood for a given anaphor in a given discourse context, other things (like the constraints just discussed) being equal. Things that affect salience:

- Background knowledge the interlocutors have about each other and the world
- The topic of the discourse, or the *question under discussion*
- The structure of the discourse itself

When constraints aren't enough

Unfortunately, all of these factors still don't uniquely determine which antecedent should be chosen.

A Your schedule is Thai Basil in Sunnyvale at 6:30 PM, followed by Dawn of the Planet of the Apes at 8:45 PM at the Mercado theater in Santa Clara.

U Where is it?

Another very important aspect is *salience*, roughly, an antecedent's relative likelihood for a given anaphor in a given discourse context, other things (like the constraints just discussed) being equal. Things that affect salience:

- Background knowledge the interlocutors have about each other and the world
- The topic of the discourse, or the *question under discussion*
- The structure of the discourse itself

All of these are complicated to plausibly model in a computational setting.

Structural influences on relative salience

Some aspects of discourse structure that bear on salience:

- Recency of mention
- Distance (number of intervening utterances)
- Number of mentions
- Embedding
- Grammatical role
- Precedence
- The anaphor's descriptive content

Recency

One of the most straightforward factors is recency of mention:

- A cowboy walked in and sat down. Another cowboy came in, and he ordered a double bourbon. The first cowboy recognized him.

Recency

One of the most straightforward factors is recency of mention:

- A cowboy walked in and sat down. Another cowboy came in, and he ordered a double bourbon. The first cowboy recognized him.

The fact that the adjective *first* seems to be required in order to rank the first cowboy mentioned over the second is evidence of the recency effect.

Distance and mentions

Relatedly, it is difficult to select an antecedent mentioned many turns ago.

Distance and mentions

Relatedly, it is difficult to select an antecedent mentioned many turns ago. But repeatedly mentioning the same antecedent raises its salience:

A Do you mean Josh Radnor or Jason Segel?

U Jason Segel. He's the one who plays Marshall.

A Do you want to see *Forgetting Sarah Marshall*? He's in that.

U I just want to know what other movies he was in.

A Are you interested in TV series? He plays in one of those as well, along with ~~James Franco~~.

U Is there anything with him playing right now?

A There is nothing else with him playing right now.

Embedding

Less-embedded antecedents seem to be relatively more salient, even when less recent:

U Is there any thing with James Franco on?

A There is [an episode of [*Freaks and Geeks*]] playing now.

U No, I have already seen that.

Grammatical role

There seems to sometimes be a preference for syntactic parallelism, as in

- Whenever a linguistics PhD runs into another graduate of the same program at ESSLLI, usually he buys him a drink.

Precedence

Normally antecedents actually antecede the anaphor that refers back to them.

- Lance scheduled a meeting with all of the journalists who had accused him of doping. ~~Alberto~~ didn't attend.

Precedence

Normally antecedents actually antecede the anaphor that refers back to them.

- Lance scheduled a meeting with all of the journalists who had accused him of doping. ~~Alberto~~ didn't attend.

This pattern isn't completely general, though:

- If he travels to a different country, Lance always has to watch out for the local authorities.

Descriptive content

The amount of descriptive content associated with the anaphor seems to be inversely correlated with how recent the antecedent must be. As discussed earlier, sometimes more descriptive content seems required:

- A cowboy walked in, and then another, and another, and finally a group of 17 cowboys walked in. The very first cowboy who came in ordered bourbon for the whole bunch.

Descriptive content

The amount of descriptive content associated with the anaphor seems to be inversely correlated with how recent the antecedent must be. As discussed earlier, sometimes more descriptive content seems required:

- A cowboy walked in, and then another, and another, and finally a group of 17 cowboys walked in. The very first cowboy who came in ordered bourbon for the whole bunch.

When descriptive content is relatively impoverished, recency becomes stronger:

U What is on right now?

A There's an episode of *Freaks and Geeks*, one of the *Terminator* movies, and a Stanford football game.

U Ok, I'll watch that.

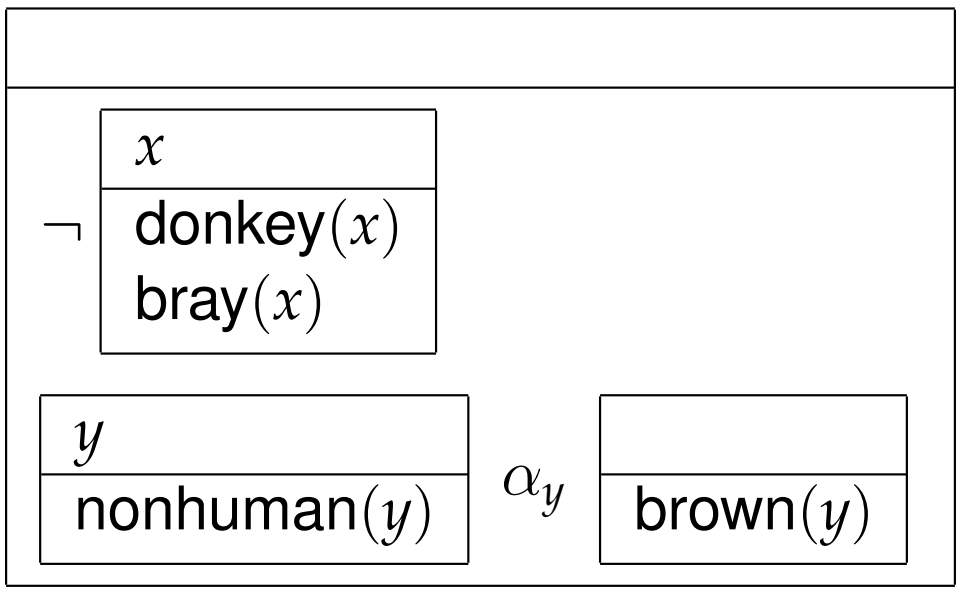
Semantic modeling

Semantics gives a way to explicitly capture the constraints. In Bos's (2003) model, for example, only equally or less deeply embedded referents are available as antecedents.

Semantic modeling

Semantics gives a way to explicitly capture the constraints. In Bos's (2003) model, for example, only equally or less deeply embedded referents are available as antecedents.

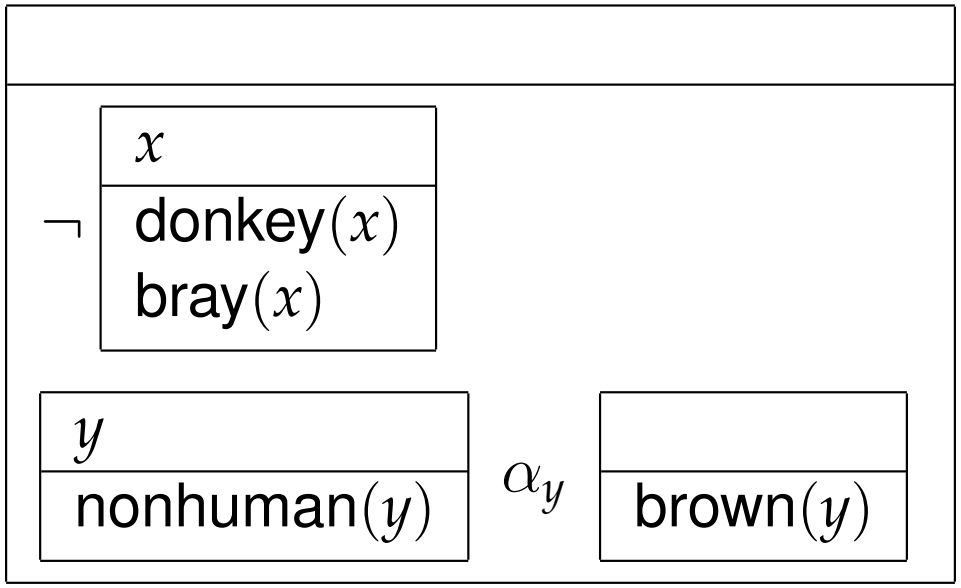
- No donkey was braying. # It was brown.



Semantic modeling

Semantics gives a way to explicitly capture the constraints. In Bos's (2003) model, for example, only equally or less deeply embedded referents are available as antecedents.

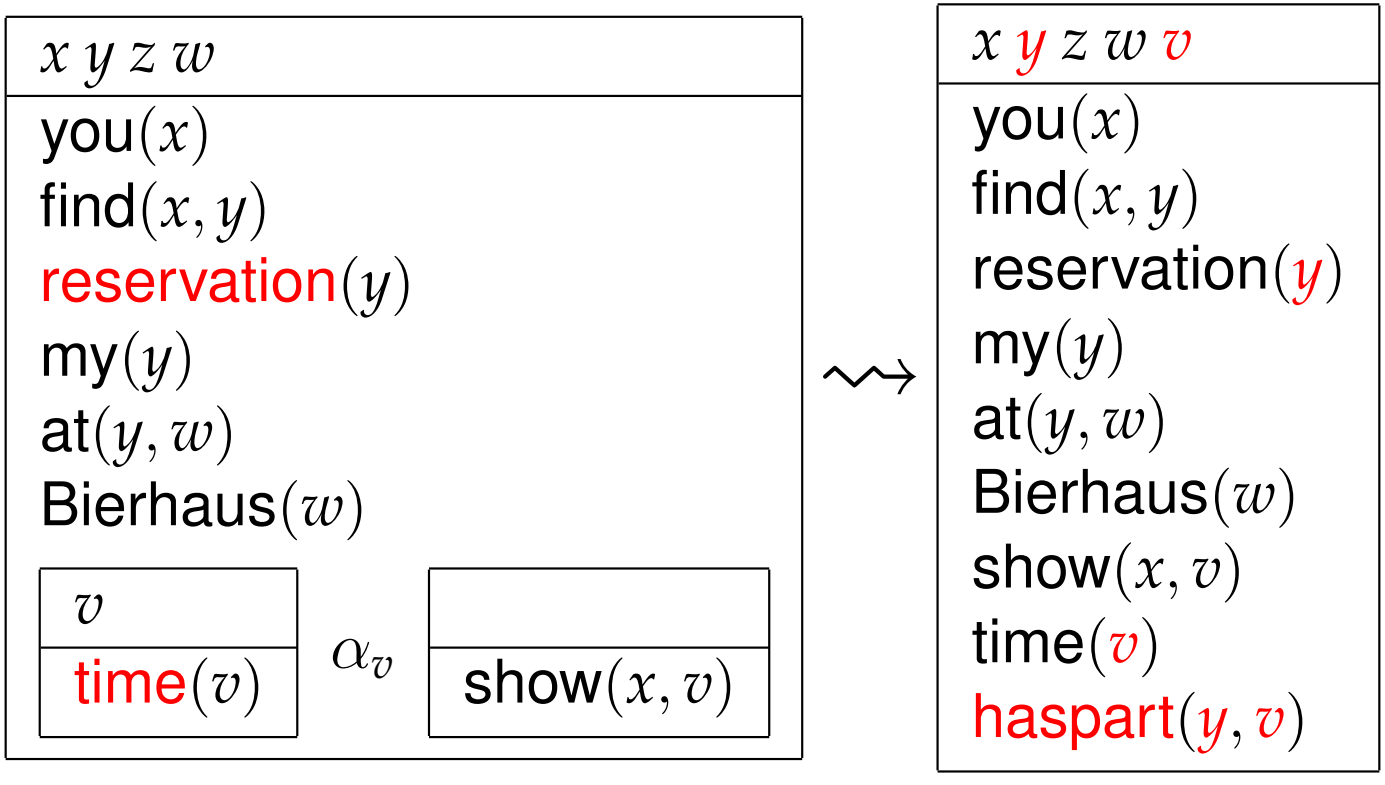
- No donkey was braying. # It was brown.



Here, the referent x is inaccessible—it is trapped in a more deeply embedded representation level.

Anaphora resolution algorithm example

- Find my reservation at ~~Bierhaus~~. Show the time.



Prerequisites of this approach

A fairly detailed syntactic analysis is required to derive these kinds of semantic representations. We also need:

- An implementation of binding constraints

Prerequisites of this approach

A fairly detailed syntactic analysis is required to derive these kinds of semantic representations. We also need:

- An implementation of binding constraints
- A source of lexical and world knowledge in order to get entailments

Prerequisites of this approach

A fairly detailed syntactic analysis is required to derive these kinds of semantic representations. We also need:

- An implementation of binding constraints
- A source of lexical and world knowledge in order to get entailments
- A theorem prover to filter out inconsistent entailments

Prerequisites of this approach

A fairly detailed syntactic analysis is required to derive these kinds of semantic representations. We also need:

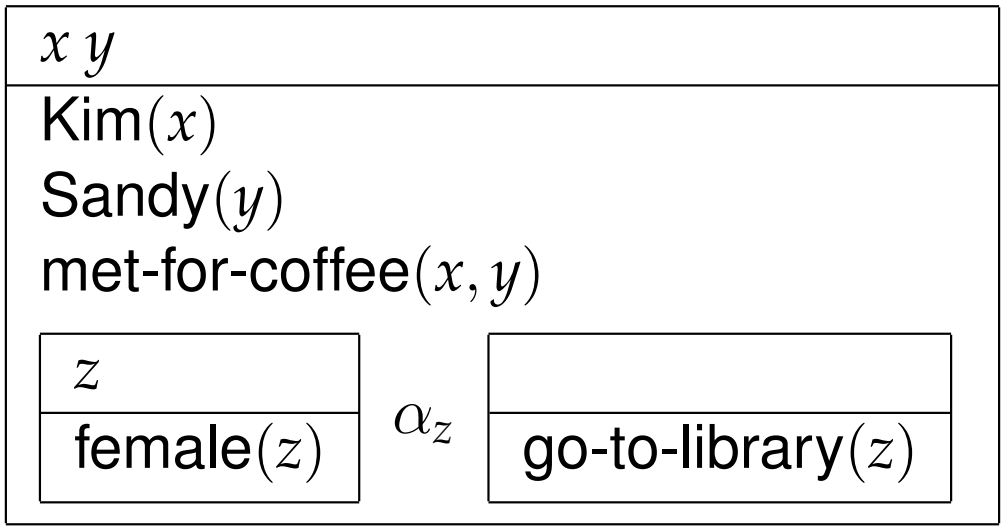
- An implementation of binding constraints
- A source of lexical and world knowledge in order to get entailments
- A theorem prover to filter out inconsistent entailments

Other deep approaches to anaphora resolution (e.g., centering-based approaches) would have similar requirements.

Constraints only go so far

Even for very simple examples, the constraints on anaphora resolution are necessary, but almost never sufficient:

- Kim met Sandy for a coffee, and after that she went to the library.



The algorithm doesn't decide between x or y as z's antecedent.

Ranking antecedents

- So Bos's algorithm, and similar approaches, essentially presents an anaphora resolution system with a list of potential antecedents for each anaphor
- The task of *ranking* them by salience remains

Ranking antecedents

- So Bos's algorithm, and similar approaches, essentially presents an anaphora resolution system with a list of potential antecedents for each anaphor
- The task of *ranking* them by salience remains
- One approach:
  - Implement each feature associated with salience (recency, embedding, etc.) as a little program that generates a score for an anaphor/antecedent pair
  - Rank the list of potential antecedents for a given anaphor by compiling, pairwise, a score for each possible resolution

Ranking antecedents

- So Bos's algorithm, and similar approaches, essentially presents an anaphora resolution system with a list of potential antecedents for each anaphor
- The task of *ranking* them by salience remains
- One approach:
  - Implement each feature associated with salience (recency, embedding, etc.) as a little program that generates a score for an anaphor/antecedent pair
  - Rank the list of potential antecedents for a given anaphor by compiling, pairwise, a score for each possible resolution
- Clearly, not all salience factors have equal impact in all situations, but they interact in complex and hard-to-understand ways

Machine learning

- The programs that generate salience scores can be weighted, and the weights can be assigned by hand (cf. Lappin and Leass 1994)
- But this seems like a task that is well-suited to machine learning instead

Machine learning

- The programs that generate salience scores can be weighted, and the weights can be assigned by hand (cf. Lappin and Leass 1994)
- But this seems like a task that is well-suited to machine learning instead
- That is, instead of a human trying to decide how the salience factors should interact, why not derive the appropriate weights by doing statistics over a lot of data, with the salience factors as features?

Machine learning

- The programs that generate salience scores can be weighted, and the weights can be assigned by hand (cf. Lappin and Leass 1994)
- But this seems like a task that is well-suited to machine learning instead
- That is, instead of a human trying to decide how the salience factors should interact, why not derive the appropriate weights by doing statistics over a lot of data, with the salience factors as features?
- There is of course the nontrivial issue of selecting an appropriate machine learning algorithm, training regime, etc.

Machine learning

- The programs that generate salience scores can be weighted, and the weights can be assigned by hand (cf. Lappin and Leass 1994)
- But this seems like a task that is well-suited to machine learning instead
- That is, instead of a human trying to decide how the salience factors should interact, why not derive the appropriate weights by doing statistics over a lot of data, with the salience factors as features?
- There is of course the nontrivial issue of selecting an appropriate machine learning algorithm, training regime, etc.
- But a more pressing problem is: where does the data come from?

Machine learning

- The programs that generate salience scores can be weighted, and the weights can be assigned by hand (cf. Lappin and Leass 1994)
- But this seems like a task that is well-suited to machine learning instead
- That is, instead of a human trying to decide how the salience factors should interact, why not derive the appropriate weights by doing statistics over a lot of data, with the salience factors as features?
- There is of course the nontrivial issue of selecting an appropriate machine learning algorithm, training regime, etc.
- But a more pressing problem is: where does the data come from?
- Ideally, we'd like a data set with a bunch of anaphors resolved to antecedents in large and small discourses, exhibiting all the salience factors—not always easy to come by

Sparsity

- Pure machine learning approaches essentially take this approach of implementing salience factors as model features, but ignore the accessibility constraints, binding constraints, etc.

Sparsity

- Pure machine learning approaches essentially take this approach of implementing salience factors as model features, but ignore the accessibility constraints, binding constraints, etc.
- Even so, the approach has its limits: trying to capture accessibility, binding, and entailment-based features via machine learning techniques runs into sparsity issues

Sparsity

- Pure machine learning approaches essentially take this approach of implementing salience factors as model features, but ignore the accessibility constraints, binding constraints, etc.
- Even so, the approach has its limits: trying to capture accessibility, binding, and entailment-based features via machine learning techniques runs into sparsity issues
- In fact, even the more basic salience factors, such as precedence and grammatical role, are fairly sparse

Sparsity

- Pure machine learning approaches essentially take this approach of implementing salience factors as model features, but ignore the accessibility constraints, binding constraints, etc.
- Even so, the approach has its limits: trying to capture accessibility, binding, and entailment-based features via machine learning techniques runs into sparsity issues
- In fact, even the more basic salience factors, such as precedence and grammatical role, are fairly sparse
- And some features, such as ones involving discourse goals or a question under discussion, may be very difficult or impossible to reliably compute

Hybrid vigor?

A possible ideal approach: a hybrid system that uses an algorithm like Bos's to implement the constraints, and a learning approach to handle antecedent ranking. The required ingredients:

- A reliable (and fast) syntactic and semantic parser,
- A source of lexical and world knowledge
- A large volume of data annotated for anaphor/antecedent relations (resolved or not)

Hybrid vigor?

A possible ideal approach: a hybrid system that uses an algorithm like Bos's to implement the constraints, and a learning approach to handle antecedent ranking. The required ingredients:

- A reliable (and fast) syntactic and semantic parser,
- A source of lexical and world knowledge
- A large volume of data annotated for anaphor/antecedent relations (resolved or not)

Note that the first two are already needed for the algorithmic approach, and the first is needed for a more fine-grained view of the annotated dataset

Conclusion

- I propose that the most successful anaphora resolution systems will have elements of both algorithmic and data-driven approaches

Conclusion

- I propose that the most successful anaphora resolution systems will have elements of both algorithmic and data-driven approaches
- It's true that both are expensive
  - Algorithmic approaches need deep syntactic and semantic analysis, and require a knowledge source
  - Machine learning approaches need a ton of human-annotated data

Conclusion

- I propose that the most successful anaphora resolution systems will have elements of both algorithmic and data-driven approaches
- It's true that both are expensive
  - Algorithmic approaches need deep syntactic and semantic analysis, and require a knowledge source
  - Machine learning approaches need a ton of human-annotated data
- But then anaphora is pervasive and complex, so we should expect anaphora resolution to be both a difficult and worthwhile task

Thanks

Thanks for listening, and any suggestions for improvement are welcome!

Acknowledgments to several current and former lab members: Chris Brew, Kathleen Dahlgren, Ben Goldsmith, Jiaying Shen, Joel Tetreault

References I

J. Bos. Implementing the binding and accommodation theory for anaphora resolution and presupposition projection. *Computational Linguistics*, 29(2):179–210, 2003. doi: 10.1162/089120103322145306.

P. Denis and J. Baldridge. Specialized models and ranking for coreference resolution. In *Conference on Empirical Methods in Natural Language Processing*, 2008.

G. Durrett and D. Klein. Easy victories and uphill battles in coreference resolution. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

S. Lappin and H. J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

H. Lee, A. Chang, Y. Piersman, N. Chambers, M. Surdeanu, and D. Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4): 885–916, 2013.

References II

J. van Eijck and C. Unger. *Computational Semantics with Functional Programming*. Cambridge University Press, 2010.