

Ethan Roday  
LING 575 SP16  
2016/05/19

# Evaluating Theories of Coreference Resolution

---



## Coreference Resolution: The Task

Bayer AG has approached Monsanto Co. about a takeover that would fuse two of the world's largest suppliers of crop seeds and pesticides, according to people familiar with the matter.

Details of the offer couldn't be learned and it's unclear whether Monsanto would be receptive to it.

Should the bid succeed, a combination of the companies would boast \$67 billion in annual sales and create the world's largest seed and crop-chemical company. A successful deal would ratchet up consolidation in the agricultural sector, after rivals Dow Chemical Co., DuPont Co. and Syngenta AG struck their own deals over the last six months.

<http://www.wsj.com/articles/bayer-makes-takeover-approach-to-monsanto-1463622691>



## Coreference Resolution: The Task

Bayer AG has approached **Monsanto Co.** about **a takeover** that would fuse two of the world's largest suppliers of crop seeds and pesticides, according to people familiar with **the matter**.

Details of **the offer** couldn't be learned and it's unclear whether **Monsanto** will be receptive to it.

Should **the bid** succeed, a combination of the companies would boast \$67 billion in annual sales and create the world's largest seed and crop-chemical company. **A successful deal** would ratchet up consolidation in the agricultural sector, after rivals **Dow Chemical Co.**, DuPont Co. and Syngenta AG struck their own deals over the last six months.

<http://www.wsj.com/articles/bayer-makes-takeover-approach-to-monsanto-1463622691>



# Not Another Machine Learning Problem

Four-step solution is typical:

- > Mention identification
- > Feature extraction
- > Pairwise coreference determination
- > Mention Clustering

Just a machine learning problem, right?



# Not Another Machine Learning Problem

Wrong! Why?

- > Dialogue is incremental
- > Dialogue is intentional
- > Can't keep the whole dialogue in context
- > Tradeoff between *accessibility* and *ambiguity*
- > Different theories of coreference make different predictions



# Theories of Coreference

Major theories have three components:

- > Linguistic structure
- > Intentional structure
- > Attentional state

Two competing theories:

- > The cache model
- > The stack model



# Theories of Coreference

## The Cache Model (Walker, 1996)

- > Linguistic structure governs attentional structure
- > Accessible referents: most recent  $n$  entities

### Parameters:

- > Cache size ( $n$ )
- > Cache update operation
  - Least Frequently Used (LFU)
  - Least Recently Used (LRU)



# Theories of Coreference

## The Stack Model (Grosz and Sidner, 1986)

- > Intentional structure governs attentional structure
- > Accessible referents: all entities in the stack

### Parameters:

- > Pushing operation
- > Popping operation





# Head To Head: Two Analyses

How do we evaluate these theories?

1. Intrinsic: simulation of coreference theories using annotated data (Poesi et al., 2006)
2. Extrinsic: inclusion in an end-to-end ML system (Stent and Bangalore, 2010)



# Head To Head: Intrinsic Analysis

## Setup:

- > Stack Model: three pushing strategies, four popping strategies
  - Twelve total systems
- > Cache Model: three cache sizes, two update strategies
  - Six total systems
- > Simulated attentional structure and compared against annotated data



# Head To Head: Intrinsic Analysis

Two primary evaluation metrics:

- > Accessibility rate (ACC)
- > Average ambiguity (Amb Ave)



# Head To Head: Intrinsic Analysis

## Stack:

Comb.	ACC	Not Acc	Amb Ave	Distr avge	Amb Perc	AmbP Wth Pre
<b>A-I</b>	42.3%(58 )	57.7%(79)	0.66	0.28	8.8%(12)	8.0%(11)
<b>A-DC</b>	48.9%(67 )	51.1%(70)	0.78	0.34	9.5%(13)	8.0%(11)
<b>A-DT</b>	62.0%(85 )	38.0%(52)	0.98	0.47	13.1%(18)	9.5%(13)
<b>A-N</b>	72.2%(99 )	27.8%(38)	1.09	0.50	13.1%(18)	9.5%(13)
<b>S-I</b>	52.5%(72 )	47.5%(65)	0.77	0.30	8.8%(12)	5.1%( 7)
<b>S-DC</b>	55.5%(76 )	44.5%(61)	0.97	0.35	9.5%(13)	5.8%( 8)
<b>S-DT</b>	70.0%(96 )	30.0%(41)	1.07	0.50	13.1%(18)	8.8%(12)
<b>S-N</b>	73.7%(101)	26.3%(36)	1.13	0.52	13.9%(19)	9.5%(13)
<b>I-I</b>	64.2%(88 )	35.8%(49)	0.93	0.39	12.4%(17)	8.0%(11)
<b>I-DC</b>	67.1%(92 )	32.9%(45)	0.97	0.40	12.4%(17)	8.0%(11)
<b>I-DT</b>	72.3%(99 )	27.7%(38)	1.08	0.48	13.9%(19)	9.5%(13)
<b>I-N</b>	73.7%(101)	26.3%(36)	1.10	0.49	13.9%(19)	9.5%(13)

## Cache:

Repl Policy	Cache Size	ACC	AmbAve avge	Distr
<b>LRU</b>	7	30.6%(42 )	0.50	0.1
	12	55.5%(76 )	0.80	0.2
	20	68.6%(94 )	1.02	0.3
	25	78.8%(108)	1.20	0.4
<b>LFU</b>	7	13.9%(19 )	0.46	0.2
	12	29.2%(40 )	0.74	0.3
	20	44.5%(61 )	0.98	0.3
	25	48.2%(66 )	1.08	0.4



# Head To Head: Extrinsic Analysis

## Setup:

- > Three feature sets:
  - Dialogue-related features
  - Task-related features
  - Basic features
- > Two pair construction strategies:
  - Stack-based: mentions in the subtask stack
  - Cache-based: mentions in the previous four turns
- > Five systems in total



# Head To Head: Extrinsic Analysis

Three primary evaluation metrics:

- > MUC-6

- Number of correct *links* in each chain

- > B<sup>3</sup>

- Correctness of chain for each *mention*

- > CEAF

- Similarity between *aligned* chains



# Head To Head: Intrinsic Analysis

Results:

Method	Scoring metric								
	MUC-6			B <sup>3</sup>			CEAF		
	R	P	F	R	P	F	R	P	F
Strong stack-based (variable history)									
Stack, Task+Dialog+Basic, CC	42.0	69.7	52.4	74.2	91.2	81.9	86.3	69.0	76.7
Stack, Task+Dialog+Basic, ILP	41.6	69.8	52.2	74.1	91.4	81.8	86.4	68.9	76.6
Hybrid cache/task-based (4 turns history)									
Cache, Task+Dialog+Basic, CC	38.9	69.7	49.9	73.2	92.1	81.6	86.5	67.6	75.9
Cache, Task+Dialog+Basic, ILP	38.6	69.8	49.7	73.1	92.2	81.6	86.6	67.5	75.9
Strong cache-based (4 turns history)									
Cache, Dialog+Basic, CC	40.0	72.8	51.6	73.6	92.9	82.2	87.1	67.8	76.2
Cache, Dialog+Basic, ILP	39.7	72.9	51.4	73.5	93.1	82.1	87.1	67.7	76.2
Cache-based baseline (4 turns history, no dialog features)									
Cache, Basic, CC	37.0	71.6	48.7	72.6	93.1	81.6	86.8	66.5	75.3
Cache, Basic, ILP	36.6	71.6	48.5	72.5	93.2	81.6	86.9	66.4	75.3
All (ILP not shown to save space)									
All, Basic, CC	38.1	69.3	49.2	72.7	91.9	81.2	86.3	67.2	75.5
All, Task+Dialog+Basic, CC	42.8	68.0	52.6	74.2	90.2	81.4	85.7	69.5	76.8



# Discussion

---

- > Stack seems to perform better overall
- > Intrinsic analysis shows:
  - Accessibility limitation of the stack
  - Ambiguity explosion with cache size
- > Extrinsic analysis shows:
  - Stack model finds more correct links
  - Stack model finds fewer and more accurate chains






# Discussion

---

## Limitations:

- > Small dataset on intrinsic evaluation
- > Extrinsic evaluation did not test cache sizes
- > Maintenance of attentional structure is non-probabilistic





UNIVERSITY *of* WASHINGTON

# Appendix



# Theories of Coreference

## The Stack Model (Grosz and Sidner, 1986)

- > Intentional structure governs attentional structure
- > Accessible referents: all entities in the stack
  
- > What is counted as a stack element?
  - Depends on theory of discourse units
    - > Clause, turn, Discourse Segment Purpose
- > When do stack elements get pushed and popped?
  - Depends on theory of discourse structure
    - > RST, DRT, RDA, ...



# Reference and Anaphora in Dialog

---

LING 575

Vinay Ramaswamy

# Reference and Anaphora

- Which words/phrases refer to some other word/phrase?
- How are they related?

Anaphora: An **anaphor** is a word/phrase that refers back to another phrase: the **antecedent** of the anaphor.

**Mary** thought that she lost **her** keys.

**her** refers to **Mary**

# Hobb's Algorithm

- Intuition:
  - Start with target pronoun
  - Climb parse tree to S root
  - For each NP or S
    - Do breadth-first, left-to-right search of children
      - Restricted to left of target
    - For each NP, check agreement with target
  - Repeat on earlier sentences until matching NP found

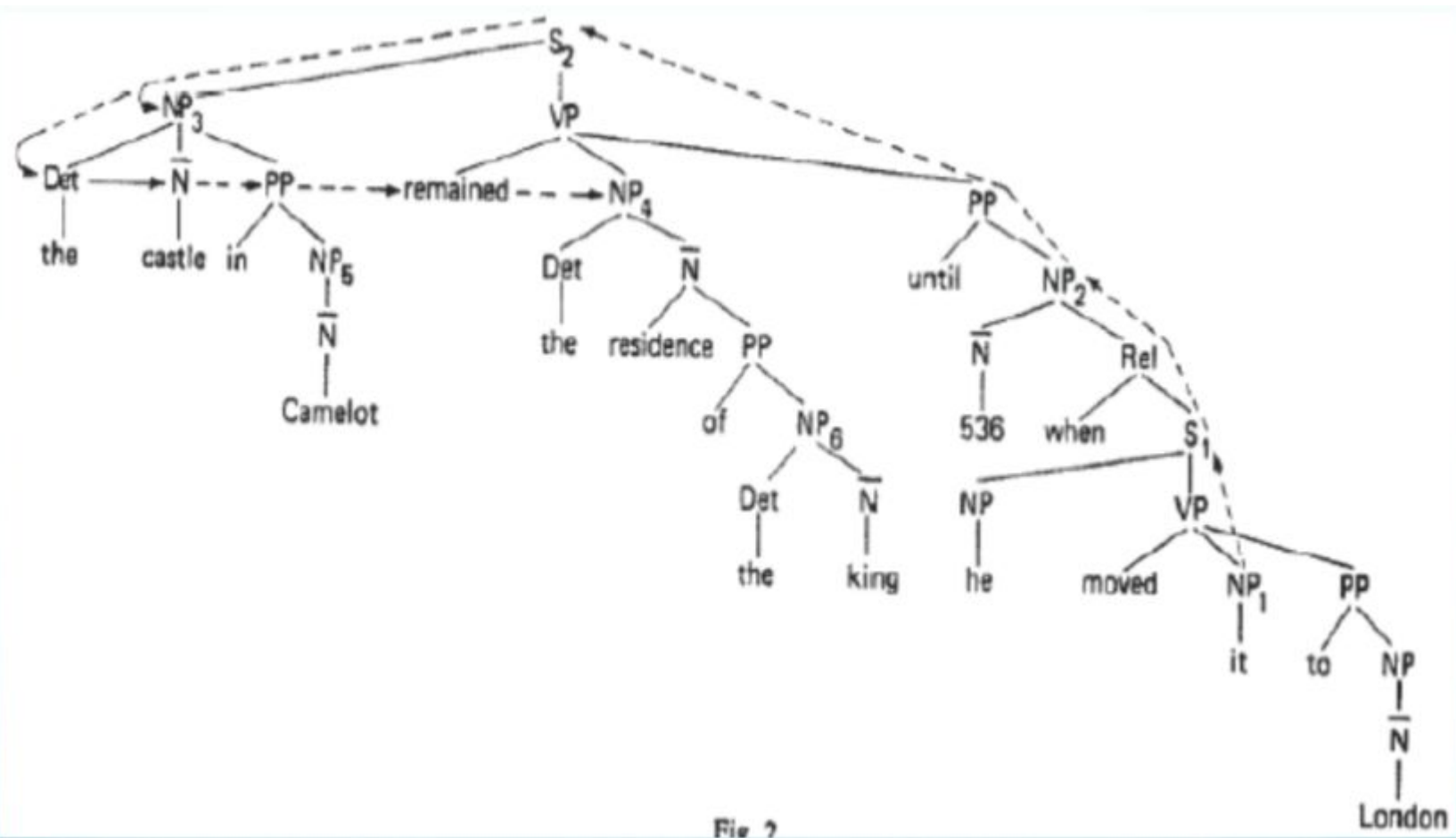


Fig 2

# Reference Resolution in Dialog

- Dialog forces us to think more globally about the process of reference.
- Speech uses lot more references than written communication.
- Reference is collaborative.
- Evidence of failure of reference attempts is typically immediate.

Example 2 (implicit grounding):

*B: [uh] i would like that in the antique ivory*

*A: and how many?*

Example 3 (evidence of misunderstanding):

*A: that is a brief tall package of two size 40 white totals one*

*A: next item?*

*B: could you repeat that was what tall?*



- Constructing a referring expression is incremental.
- Most evident when a hearer completes a referring expression started by a speaker

**Example 4 (completion):**

*A: what was that number, t a 0 0 7*

*B: 0 0 4*

- Reference is hearer-oriented
- No reference attempt can succeed without the understanding and agreement of the hearer.
- For ex. In an instruction giving task a speaker may make a referring expression less technical if the hearer is not a domain expert

# A Machine Learning Approach to Pronoun Resolution

Michael Strube and Christoph Muller

- Decision tree based approach to pronoun resolution in spoken dialogue.
- Works with pronouns with NP- and non-NP-antecedents.
- Features designed for pronoun resolution in spoken dialogue.
- Evaluate the system on twenty Switchboard dialogues.
- Corpus-based methods and machine learning techniques have been applied to anaphora resolution in written text with considerable success.
- Describes the extensions and adaptations needed for applying their anaphora resolution system from their earlier paper to pronoun resolution in spoken dialogue.

# NP and non-NP Antecedents

A1: ... [he]<sub>i</sub>'s nine months old. ...

A2: [He]<sub>i</sub> likes to dig around a little bit.

A3: [His]<sub>i</sub> mother comes in and says, why did you let [him]<sub>i</sub>  
[play in the dirt]<sub>j</sub>,

A:4 I guess [[he]<sub>i</sub>'s enjoying himself]<sub>k</sub>.

B5: [That]<sub>k</sub>'s right.

B6: [It]<sub>j</sub>'s healthy, ...

# NP and non-NP Antecedents

- Abundance of (personal and demonstrative) pronouns with non-NP-antecedents or no antecedents at all.
- Corpus studies have shown - a significant amount (50%) of pronouns have non-NP-antecedents, in dialog.
- Performance of a pronoun resolution algorithm can be improved considerably by resolving pronouns with non-NP-antecedents.
- NP-markables identify referring expressions like noun phrases, pronouns and proper names.
- VP-markables are verb phrases, S-markables sentences.

# Data Generation

- All markables were sorted in document order
- Markables - contain member attribute with the ID of the coreference class they are part of.
- If the list contained an NP-markable at the current position and if this markable was not an indefinite noun phrase, it was considered a potential anaphor.
- In that case, pairs of potentially co-referring expressions were generated by combining the potential anaphor with each compatible NP-markable preceding it in the list.
- The resulting pairs were labelled P if both markables had the same (non-empty) value in their member attribute, N otherwise.
- Non-NP-antecedents -Potential non-NP-antecedents generated by selecting S- and VP-markables from the last two valid sentences preceding the potential anaphor.

# Features

NP-Level :

Grammatical Function, NP Form, case etc.

Coreference-Level : (Relation between Antecedent and Anaphor)

Distance, compatibility in terms of agreement

Dialog Features :

Expression type, importance of expression in dialog, information content

---

**NP-level features**

1. ante\_gram\_func grammatical function of antecedent
2. ante\_npform form of antecedent
3. ante\_agree person, gender, number
4. ante\_case grammatical case of antecedent
5. ante\_s\_depth the level of embedding in a sentence
6. ana\_gram\_func grammatical function of anaphor
7. ana\_npform form of anaphor
8. ana\_agree person, gender, number
9. ana\_case grammatical case of anaphor
10. ana\_s\_depth the level of embedding in a sentence

---

**Coreference-level features**

11. agree\_comp compatibility in agreement between anaphor and antecedent
12. npform\_comp compatibility in NP form between anaphor and antecedent
13. wdist distance between anaphor and antecedent in words
14. mdist distance between anaphor and antecedent in markables
15. sdist distance between anaphor and antecedent in sentences
16. syn\_par anaphor and antecedent have the same grammatical function (yes, no)

---

**Features introduced for spoken dialogue**

17. ante\_exp\_type type of antecedent (NP, S, VP)
  18. ana\_np\_pref preference for NP arguments
  19. ana\_vp\_pref preference for VP arguments
  20. ana\_s\_pref preference for S arguments
  21. mdist\_3mf3p (see text)
  22. mdist\_3n (see text)
  23. ante\_tfidf (see text)
  24. ante\_ic (see text)
  25. wdist\_ic (see text)
-

# Results

	<b>correct found</b>	<b>total found</b>	<b>total correct</b>	<b>precision</b>	<b>recall</b>	<b>f-measure</b>
<b>baseline, features 1-16</b>	456	739	1250	61.71	36.48	45.85
<b>combined</b>	509	897	1250	56.74	40.72	47.42

- Refers to manually tune, domain specific implementation which has 51% f-measure
- Acknowledge “Major problem for a spoken dialog pronoun resolution algorithm is the abundance of pronouns without antecedents.”
- Tested on only 20 switchboard dialogues
- Features selected to improve performance on data, is it really portable? Or does take extensive work to go fine tune the performance?



# Incremental Reference Resolution

David Schlangen, Timo Baumann, Michaela Atterer

- Discuss the task of incremental reference resolution.
- Specify metrics for measuring the performance of dialogue system components tackling this task.
- Task is to identify the pieces of Pentomino game.
- Presents a Bayesian filtering model of IRR using words directly: it picks the right referent out of 12 for around 50 % of real- world dialogue utterances in test corpus.

# Incremental Reference Resolution

“The Red Cross”

If only one *red cross*, one *green circle*, and two *blue squares* are there, one can say that after “the red” the reference is “Red Cross”.

If there are two red crosses, need to look for further restricting information (e. g. “. . . on the left”).

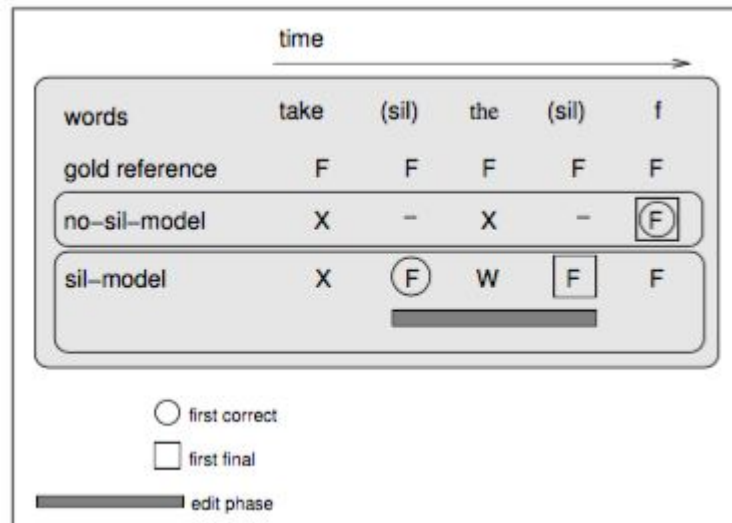
IRR words encountered that express features that reduce the size of the set of possible referents.

“Red”, “Cross”, “Left”...

At each step the expression is checked against the world model to see whether the reference has become unique.

# Evaluation Metrics

- Focuses on identification of an entity by an utterance.
- Assumption - there is one intention behind the referring utterances, and intention is there from the beginning of the utterance and stays constant.
- Positional Metric - measures when a certain event happens
- Edit metric - measures the “jumpiness” of the decision process (how often changes mind during an utterance)

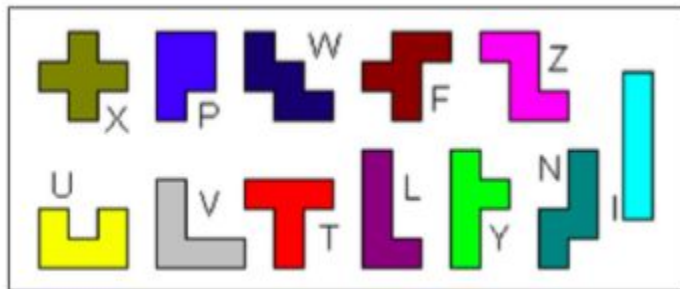


# Evaluation Metrics

- average first correct - how deep into the utterance do we make the first correct guess?
- average first final - how deep into the utterance do we make the correct guess and don't subsequently change our mind?
- ed-utt (mean edits per utterance) - may still change its mind even after it has already made a correct guess. This metric measures how often the module changes its mind before it comes back to the right guess.
- Correctness - how often the model guesses correctly

# Corpora

Instruction Giver (IG) instructs an Instruction Follower (IF) on which puzzle pieces to pick up



Intra-utterance silences (hesitations) could potentially be used as an information source in the corpus data.

# Belief Update Model

The authors use a Bayesian model which treats the intended referent as a latent variable generating a sequence of observations

$$P(r|w_{1:n}) = \alpha * P(w_n|r, w_{1:n-1}) * P(r|w_{1:n-1})$$

Before the first observation,  $P(r)$  is a distribution over all possible referrals.

E. g., an utterance like “take the long, narrow piece” will be processed one word at a time.

# Decision

In the arg max approach, at each state the referent with the highest posterior probability is chosen - can cause many edits.

In the adaptive threshold approach, start with a default decision -“undecided”.

New decision is only made if the maximal value at the current step is above a certain threshold, where this threshold is reset every time this condition is met. Favours strong convictions and reduces jitter.

# Machine Learning & Reference Resolution

- Both the papers focused on a very limited data
- The first paper attempted to provide techniques with 50% accuracy
- The second paper focused on Instructions giving and taking on Pentomino game.
- Are machine learning techniques better than handcrafted techniques for a specific domain?
- Are they better than Hobb's algorithm or Multi-sieve algorithms?
- Reference resolution is integral part of any dialog system which involve interaction with humans.



lopez380:

Would you the latest precision and recall values for anaphora resolution. I read the paper "A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue" and precision is around 79 %. The paper was written in 2003. I would be interested in knowing the latest upper limits possible,

Jeff Heath:

What exactly would the co-reference graph described in the paper look like? Can someone provide an example that could clarify its construction and use, especially demonstrating the adjustment of link weights when a link is grounded (when the hearer displays correct understanding) or rejected (when hearer understanding fails)?

It seems like over-specifying the characteristics when producing (generating) a reference gives a slight reduction in communication efficiency, but under-specifying would result in confusion and likely a very inefficient exchange to resolve the confusion. So wouldn't it be better to err on the side of over-specifying references when producing an utterance?

In a multi-party dialogue, each speaker must keep a co-reference graph for each of the other partners. How might one produce a reference when speaking in that scenario? Always speak at the level of the least informed of the hearers? Does that make sense from our experience?

carye:

The primary paper mentions that grounding is often implicit in this context: “...the hearer only provides evidence of the failure of a reference attempt.” But the author goes on to suggest the necessity for computational models to track participants’ understandings of common information during the dialog. How could we track successful comprehension in this case? Do we just assume the absence of certain speech cues means the reference was successful?

mnij525

Towards the end of the paper, the author discusses non-humanlike reference. They mention that humanlikeness may be unnecessary or maladaptive at times. Also, earlier in the paper, the author mentions that humans are subject to memory limitations which may prevent optimal referring expressions. Im curious about how observations like these will impact NLP. Thoughts?

jason:

At the end of the primary paper the author mentions non-collaborative dialog and lists some examples such as teaching a student, selling, a product, and hiding information. It also seems to define non-collaborative dialog as instances where dialog partners are not fully cooperative or fully task-focused. So, does the distinction mostly based around intention? That is to say, a teacher talking to a student is non-collaborative because of the distinct roles taken by the teacher and the student as opposed to the fact that one participant is talking at length and without interruption, barring an occasional question from a student which must be acknowledged by the teacher. On the other hand, a conversation between two people where one is very enthusiastic about a subject and the other is entirely disinterested would still collaborative even if, from an outside perspective, it's almost the same a teacher-student dialog. Or is that incorrect? If one dialog partner does little or no participation in a dialog is it no longer collaborative? Also, if you are looking at non-collaborative dialog then is the earlier speaker-focused model a viable option?

# Information Structure and Prosody in Dialog

Calhoun et al., 2005

*A Framework for Annotating Information Structure in Discourse*, in Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky (ACL 2005).

Hirschberg, 1990

*Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech* Proc. AAAI 90, pp. 952-957.

- **Text w/o Annotation**

But Yemen's president says the FBI has told him the explosive material could only have come from the U.S., Israel, or two arab countries. And to a former federal bomb investigator, that description suggests a powerful military-style plastic explosive C-4 that can be cut or molded into different shapes.



- **Text w/ Annotation**

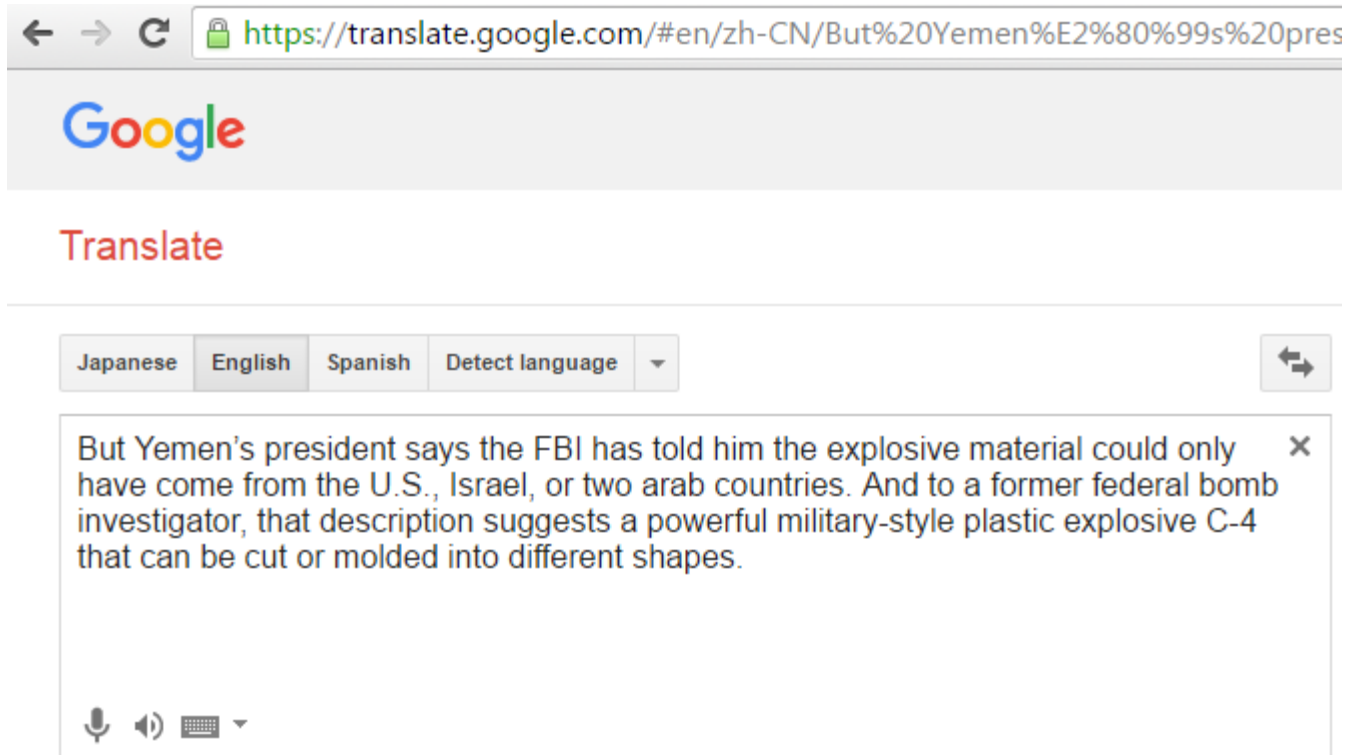
[But [[[Yemen's]<sub>med/general</sub> president]<sub>med/poss</sub> ]<sub>Contrastive</sub> says ]<sub>THEME</sub>  
 [[the FBI]<sub>old/identity</sub> has told [him]<sub>old/identity</sub> ]<sub>THEME</sub> [ [the explosive  
 material]<sub>med/set</sub> could only have come from [[[the U.S.]<sub>med/general</sub>, [Israel]<sub>med/  
 general</sub>, or [[two arab countries]<sub>med/set</sub>]<sub>med/aggregation</sub> ]<sub>Adverbial</sub> ]<sub>RHEME</sub> [And to  
 [[a former federal bomb investigator]<sub>new</sub> ]<sub>Contrastive</sub> ]<sub>THEME</sub> [[that  
 description]<sub>old/event</sub> suggests]<sub>THEME</sub> [[a powerful military-style plastic  
 explosive C-4]<sub>med/set</sub> ]<sub>Answer</sub> [[that]<sub>old/relative</sub> can be cut or molded into [different  
 shapes]<sub>new</sub> ]<sub>RHEME</sub>

# Applications

- Paraphrase analysis and generation;
- Topic detection;
- Information extraction;
- Speech synthesis in dialogue systems.



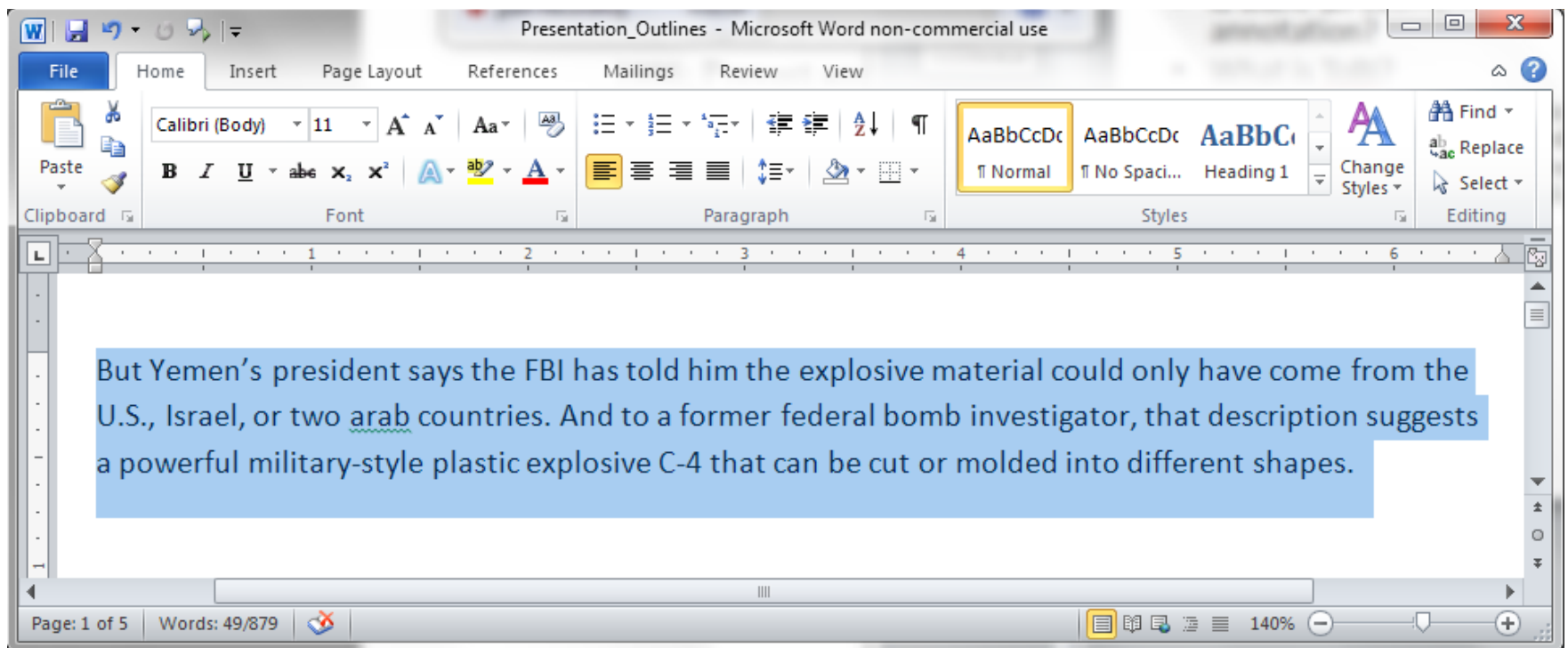
# A Google's TTS



The screenshot shows the Google Translate web interface. The browser's address bar contains the URL: <https://translate.google.com/#en/zh-CN/But%20Yemen%E2%80%99s%20pres>. The Google logo is visible at the top left. Below it, the word "Translate" is written in red. A language selection bar shows "Japanese", "English", "Spanish", and "Detect language" with a dropdown arrow. To the right of this bar is a bidirectional arrow icon. The main text input area contains the following text: "But Yemen's president says the FBI has told him the explosive material could only have come from the U.S., Israel, or two arab countries. And to a former federal bomb investigator, that description suggests a powerful military-style plastic explosive C-4 that can be cut or molded into different shapes." A small 'x' icon is in the top right corner of the text box. At the bottom left of the text box, there are icons for voice input (microphone), volume (speaker), and keyboard input (keyboard icon with a dropdown arrow).



# Microsoft TTS



# Capturing textual and prosodic characteristics

- Information Status

Expresses the *availability* of entities in discourse.

- Prosodic Structure

How *intonation* phrase is organized in the discourse model, and how *salient* (*i.e. noticeable*) the speaker wishes to make each entity, property or relation.

# Information Status

- **New:** not have been previously referred to; unknown to the hearer.  
*e.g.* [a former federal bomb investigator]<sub>*new*</sub>, [different shapes]<sub>*new*</sub>
- **Mediated:** newly mentioned but the hearer can infer from the prior context.  
*e.g.* [the U.S.]<sub>*med/general*</sub>, [the explosive material]<sub>*med/set*</sub>  
  
subtypes: *general, bound, part, situation, event, set, poss., func-value,*  
and *aggregation*
- **Old:** not new nor mediated  
*e.g.* [the FBI]<sub>*old/identity*</sub>, [that description]<sub>*old/event*</sub>  
  
subtypes: *identity, event, general, ident\_generic, relative.*

# Prosodic Structure

- **Theme/Rheme:** A prosodic is marked as *theme* if it only contains information which links the utterance to the *preceding context*; Otherwise, it is marked as *rheme*.

e.g. I lived over in England for four years.

Where I lived was a town called Newmarket.

*Theme*

*Rheme*

L+H\*    L+H\* -    -    H\*                    H\*    LL%    (pitch  
accent)    (Hirschberg 1990)

e.g. [[that description]<sub>old/event</sub> suggests]<sub>THEME</sub>    [[a powerful  
military-style plastic explosive C-4]<sub>med/set</sub> ]<sub>Answer</sub>    [[that]<sub>old/relative</sub>  
can be cut or molded into [different shapes]<sub>new</sub> ]<sub>RHEME</sub>

# Prosodic Structure (cont'd)

- Theme & Rheme Identification

Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. (1997)

*Semantic and Discourse Information for Text-to-Speech Intonation.* In Proceedings of Workshop on Concept-to-Speech Generation Systems.

# Prosodic Structure (cont'd)

- **Background/Kontrast:** Anything that cannot be marked as *kontrast* is marked as *background*; Kontrast categories:

**Correction:** (now are you sure they're **HYACINTHS**) (because that is a BULB)

**Contrastive:** (A) I live in *Garland*, and we're just beginning to build a real big recycling center... (B) (YEAH there's been) (NO emphasis on recycling at ALL) (in San ANTONIO)

**Subset:** (THIS woman owns *THREE day cares*) (**TWO** in Lewisville) (and **ONE** in Irving) ...

**Adverbial:** ... *only*...from [[the U.S.]<sub>med/general</sub>, [Israel]<sub>med/general</sub>, or [[two arab countries]<sub>med/set</sub>]<sub>med/aggregation</sub>]. **Adverbial**

**Answer:** suggest [[a powerful military-style plastic explosive C-4]<sub>med/set</sub>]. **Answer**

# Data and Tools Used

- **Source Data:**  
Switchboard Corpus (Godfrey et al., 1992)
- **Tool:**  
Nite XML Toolkit (NXT)  
([https://sourceforge.net/projects/nite/files/nite/nxt\\_1.4.4/](https://sourceforge.net/projects/nite/files/nite/nxt_1.4.4/))
- **Output Data:**  
Multi-layered XML-conformant schema



# Validation of the Scheme

- Rule:  $K \geq .80$  (Kappa statistics)

- Result:

2 Anotators, 1738 markables, 3 main categories (*old*, *mediated*, and *new*), and the *non-applicable* category.

$K = .845$  for the high-level categories, and

$K = .788$  when including subtypes.

- Conclusion

These results show that overall the annotation is reliable and that the scheme has good reproducibility.

# Q & A 1

[George Cooper]: For Calhoun et al. 2005, how different would you expect the annotation results to be if the annotators did not have *access to the audio files* when annotating information structure? Are there cases in which the *audio would be truly necessary* for distinguishing between different annotations?

A: I do not think it will make much difference. Audio files are primarily used for prosodic information collection. The theme and rheme can be very different for different audio even if the corresponding texts are the same.

## Q & A 2

[John T. McCranie] : 1) Is there anything like the Swithboard corpus for other languages?

2) Is ToBI for English only?

Would it just need to be tweaked a bit for other languages, are it is too tightly coupled to English prosody?

A: Yes. ToBI is English language specific, and tightly coupled to it. Different ToBI needs to be developed for different languages. ToBI systems have been defined for a number of other languages; for example, J-ToBI refers to the ToBI conventions for Tokyo Japanese.

## Q & A 3

- [laurenf7]: In *Section 4* of the primary paper the authors insist that anything annotated as "*theme*" *must sound acceptable* when spoken with a *highly marked tune*, even if this is not the tune the speaker used. This makes me wonder *how useful* examining prosody would even be in this case, as it's clear that the extra pitch accent is not necessarily required and it may be that the speaker chose purposely to leave it unaccented. Annotating a prosodic phrase based on its acceptability with a different pitch contour than that used seems to lose important information. Thoughts?

Laurie Hiyakumoto, Scott Prevost, and Justine Cassell. (1997)

Q: I know the SMART programmer wrote the SPEEDY algorithm,  
(But WHICH algorithm) (did the STUPID programmer write?)  
L+H\* L-H% H\* L-L%

A: (The 

STUPID L+H* <i>theme-focus</i>
--------------------------------------

 programmer wrote) (the 

SLOW H* <i>rheme-focus</i>
----------------------------------

 algorithm.)  
*Theme* *Rheme*  
L-H% L-L%

Figure 1: An Example of Information Structure

## Q & A 4

[spencedm]: What are relatively good/bad kappa scores for such inter-annotator agreement? Is comparing just two annotators for such a scheme pretty common?

A: .80. Kappa is for comparing two raters. You could have more than one, then you will have multiple pairs of raters and consequently multiple Kappa cores.

# Mixed Initiative

Maria Sumner

May 19, 2016

# What is initiative?

- “taking the conversational lead”
- “control”
- “Initiative is about leading the conversation toward the dialogue goal.”

Mixed initiative - the system or user being able to arbitrarily take or give up the initiative in various ways.

(Jurafsky & Martin)



# Highlights

- Chu-Carroll & Brown (1997)
- Strayer, Heeman & Yang (2003)
- English & Heeman (2005)
- Yang & Heeman (2007)
- Morbini et al (2012)

## Tracking Initiative in Collaborative Dialogue Interactions- Chu-Carroll & Brown (1997)

S: I want to take NLP to satisfy my course requirement.

S: Who is teaching NLP?

(a) A: Dr. Smith is teaching NLP.

(b) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP.

(c) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP. You should take distributed programming to satisfy your requirement, and sign up as a listener for NLP.

# Tracking Initiative in Collaborative Dialogue Interactions- Chu-Carroll & Brown (1997)

S: I want to take NLP to satisfy my course requirement.

S: Who is teaching NLP?

no change

(a) A: Dr. Smith is teaching NLP.

(b) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP.

(c) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP. You should take distributed programming to satisfy your requirement, and sign up as a listener for NLP.

# Tracking Initiative in Collaborative Dialogue Interactions- Chu-Carroll & Brown (1997)

S: I want to take NLP to satisfy my course requirement.

S: Who is teaching NLP?

(a) A: Dr. Smith is teaching NLP.

(b) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP.

(c) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP. You should take distributed programming to satisfy your requirement, and sign up as a listener for NLP.

no change

dialogue init

# Tracking Initiative in Collaborative Dialogue Interactions- Chu-Carroll & Brown (1997)

S: I want to take NLP to satisfy my course requirement.

S: Who is teaching NLP?

(a) A: Dr. Smith is teaching NLP.

(b) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP.

(c) A: You can't take NLP because you haven't taken AI, which is a prerequisite for NLP. You should take distributed programming to satisfy your requirement, and sign up as a listener for NLP.

no change

dialogue init

task & dialogue  
init

# Tracking Initiative in Collaborative Dialogue Interactions- Chu-Carroll & Brown (1997)

- Created a model for predicting dialogue initiative and task initiative
- Used evidence from cues (linguistic, domain knowledge)
- Predicted with 99.1%/87.8% accuracy and found improvements in other domains

# The good and the bad

- Identified the need to consider initiative as multi-threaded
- Improved understanding of shift cues
- Generalizable model

- Low kappa scores
- Affected  $\frac{1}{4}$  turns
- Improvements in the other domains were tested against a very simple baseline

## Reconciling Control and Discourse Structure- Stayer, Heeman, & Yang (2003)

- Found that control is subordinate to discourse structure
- Looked at task oriented dialogues (TRAINS)
- Control is with initiator of discourse segment (88%)
- Concluded that control does not need to be tracked, only intentional structure



# Learning Mixed Initiative Dialogue Strategies By Using Reinforcement Learning on Both Conversants- English & Heeman (2005)

- Dialog policy- an enumeration of all states a system can be in and corresponding action to take from those states
- Typical approaches: hand-crafting a policy, iterative Wizard-of-Oz, inducing from a human-human corpus
- Used reinforcement learning for both participants, furniture task, near hand-crafted systems levels
- Showed that you can use reinforcement learning to construct an effective dialog policy

# Design World Task (Walker 1995)

## The Task

2 agents arranging furniture

Furniture specified by type, color, value

Agents have preferences (ie If item X is in the room, item Y must also be in the room) and the preferences have values

Choose 5 furniture items to optimize score

Score = sum of furniture values - violated preferences

# Exploring Initiative Strategies Using Computer Simulation- Yang & Heeman (2007)

- Found support for empirical findings about initiative not bouncing back and forth
- Showed restrictive initiative was most time efficient, thus would be good for SDSs
- Using computer simulation to better understand human conventions

# A Mixed-Initiative Conversational Dialogue System for Healthcare- Morbini et al. (2012)

- Web application SimCoach designed primarily for mental health concerns for veterans
- Has to be able to take initiative and respond when the user does
- Information-state based dialogue system



# Example

Dialogue transcript	Notes
<p><b>BBQ Smalltalk</b></p> <p>Ask anybody about me, and they'll tell you that I love to BBQ</p> <p><i>Is this conversation secret?</i></p> <p><b>Confidentiality QA</b></p> <p>We don't share your info with anyone who can personally identify you. The techs can see what we say, but just to tell that the site is working. But they have no idea who said it, just what was said</p> <p>Did that help you?</p> <p><i>Yes it did.</i></p> <p>Great.</p>	<p>The character is equipped with a few operators for smalltalk about a few topics. BBQ is one of them.</p> <p>Here the system is interrupted by a user question and it decides that answering it is the best course of action.</p>

<p><b>BBQ Smalltalk</b></p> <p>Like I was saying, I love to BBQ</p> <p><i>What is PTSD?</i></p> <p><b>What is PTSD QA</b></p> <p>PTSD, or post-traumatic stress disorder is an anxiety condition associated with serious traumatic events. It can come with survivor guilt, reliving the trauma in dreams, numbness, and lack of involvement with reality.</p> <p><b>PTSD Topic Interest QA</b></p> <p>So, is PTSD something you're worried about. I only ask, because you've been asking about it. ●●●</p>	<p>After answering the question, the best course of action is to awaken the paused operator about the BBQ smalltalk.</p> <p>Again the BBQ smalltalk is interrupted by another question from the user.</p> <p>After answering the second question the system decides to ignore the paused operator and load a follow-up operator related to the important topic raised by the user's question. The selection is based on the expected reward that talking about PTSD can bring to the system.</p>
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

# Demo

- <https://www.youtube.com/watch?v=PGYUqTvE6Jo>

# GoPost Questions

- In the primary paper, the authors present a model that uses different counting methods that lead to different accuracy results on the prediction of initiative holders. Is there some insights why a 'constant-increment-with-counter' has the best performance than just looking at the empirical results?
- The primary paper makes the distinction about task and dialogue initiatives being different and useful to analyze separately- have other people taken this up?

Also in the primary paper- I'm kind of confused by the figure 2 graphs and why the const-inc accuracy dips so dramatically between 0.25-0.35 delta; was this explained?

- They kind of hand-waved about their cross-annotator agreement issues for dialog and task initiative labels and then they did not discuss it all for cue annotations. I'd be curious to see the per-cue-type break down of that agreement and see if it correlated with performance for that cue type in their tests.
- I'd like them to try something like MaxEnt to build their prediction models just to see how it performed relative to their approach.
- Given some of the prior topics surrounding using acoustic features to predict dialog elements, I'd wonder how acoustic features would aid in this prediction? My intuition is that they would help since reflection seems to change when one expects another to take up the conversation.



- My main issue was with the cross-annotator (dis)agreement as well -- they mention that their K scores were fairly low, even outside (or on the very, very low end of the spectrum) between  $0.67 < K < 0.8$  upon which "tentative conclusions" could be drawn. Despite their continuing argument that Kappa scores don't matter so much, wouldn't these scores suggest that any conclusions drawn in the paper have no legs to stand on?

I wonder if they were able to automate this system, perhaps by combining some kind of basic slot-filling model for certain (simpler) features with more machine-readable features (prosody, etc.), they could get some results that have a more solid, standardized foundation. If people have that much trouble tracking initiative with this model, it may not be a great model for the current state of NLP.

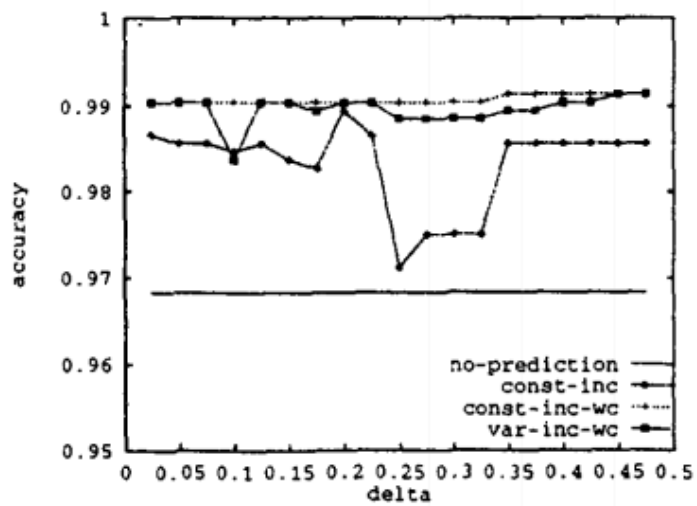
- I wish if they have presented some examples and analysis of why inter-annotator agreement was low. In particular what could happen is that dialog/task initiatives are ambiguous in certain cases. Further number of annotations seems to be relatively small: ~1000 turns so this could be only about 50-100 dialogs.

Analytical category of cues seems quite powerful for predicting switching task initiative to hearer. Still, it's looks as category of cues that would be hardest to extract automatically. Are, there some successfull attempts to do this?

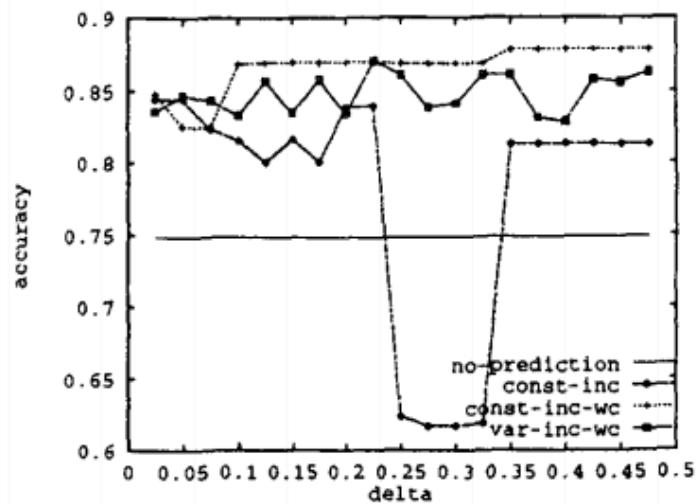
# Cues from Chu-Carrol & Brown

Class	Cue Type	Subtype	Effect	Initiative	Example
Explicit	Explicit requests	give up	both	hearer	"Any suggestions?" "Summarize the plan up to this point"
		take over	both	speaker	"Let me handle this one."
Discourse	End silence		both	hearer	
	No new info	repetitions	both	hearer	A: "Grab the tanker, pick up oranges, go to Elmira, make them into orange juice." B: "We go to Elmira, we make orange juice, okay."
		prompts	both	hearer	"Yeah", "Ok", "Right"
	Questions	domain	DI	speaker	"How far is it from Bath to Corning?"
		evaluation	DI	hearer	"Can we do the route the banana guy isn't doing?"
	Obligation fulfilled	task	both	hearer	A: "Any suggestions?" B: "Well, there's a boxcar at Dansville." "But you have to change your banana plan." A: "How long is it from Dansville to Corning?"
		discourse	DI	hearer	A: "Go ahead and fill up EI with bananas." B: "Well, we have to get a boxcar." A: "Right, okay. It's shorter to Bath from Avon."
Analytical	Invalidity	action	both	hearer	A: "Let's get the tanker car to Elmira and fill it with OJ." B: "You need to get oranges to the OJ factory."
		belief	DI	hearer	A: "It's shorter to Bath from Avon." B: "It's shorter to Dansville." "The map is slightly misleading."
	Suboptimality		both	hearer	A: "Using Saudi on Thursday the eleventh." B: "It's sold out." A: "Is Friday open?" B: "Economy on Pan Am is open on Thursday."
	Ambiguity	action	both	hearer	A: "Take one of the engines from Corning." B: "Let's say engine E2."
		belief	DI	hearer	A: "We would get back to Corning at 4." B: "4PM? 4AM?"

Table 2: Cues for Modeling Initiative



(a) Task Initiative Prediction



(b) Dialogue Initiative Prediction

# References

- J. Chu-Carroll and M. Brown. (1997) "Tracking Initiative in Collaborative Dialogue Interactions". Proceedings of ACL 1997 .
- M. English and P. Heeman. (2005) Learning mixed initiative dialog strategies by using reinforcement learning on both conversants. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, p. 1011-1018.
- Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David Traum and Albert Rizzo. A Mixed-Initiative Conversational Dialogue System for Healthcare. Demonstration in SIGdial 2012, the 13th Annual SIGdial meeting on Discourse and Dialogue, Seoul, South Korea, 2012.
- D. Novich and S. Sutton. What is mixed-initiative interaction? In *Proceedings of the AAIL Spring Symposium on Computational Models of Mixed Initiative Interaction*, 1997.
- S. Strayer, P. Heeman, and F. Yang. (2003) Reconciling control and discourse structure. In J. van Kuppevelt and R.W.Smith, editors, *Current and New Directions in Discourse and Dialogue*. Kluwer Academic Publishers, Dordrecht, Chapter 14, p. 305-323
- F. Yang and P. Heeman. Exploring Initiative Strategies Using Computer Simulation. In Proceedings of the 10th European Conference on Speech Communication and Technology, Antwerp Belgium, August 2007.