Computational Models of Reference, Anaphora and Coreference (CRAC)
2019-06-07, Minneapolis, MN

Minneapolis, MN
NAACL HLT 2019
June 2-7, 2019

Coreference and Coherence Revisited

Amir Zeldes

Georgetown University

amir.zeldes@georgetown.edu

Corpling@GU

Plan

1. Coherence (and problems) in coreference resolution
2. Predicting subsequent mention
3. Discourse graphs as heat maps
4. Conclusion and future work

Why coherence and coreference?

- Recent years have seen substantial gains in f-scores on coref in OntoNotes
- But there is a lingering sense of dissatisfaction:
  - Scores in 70s do not lead to trustworthy results
  - System errors sometimes bizarre
  - Out of domain performance often worse than older systems

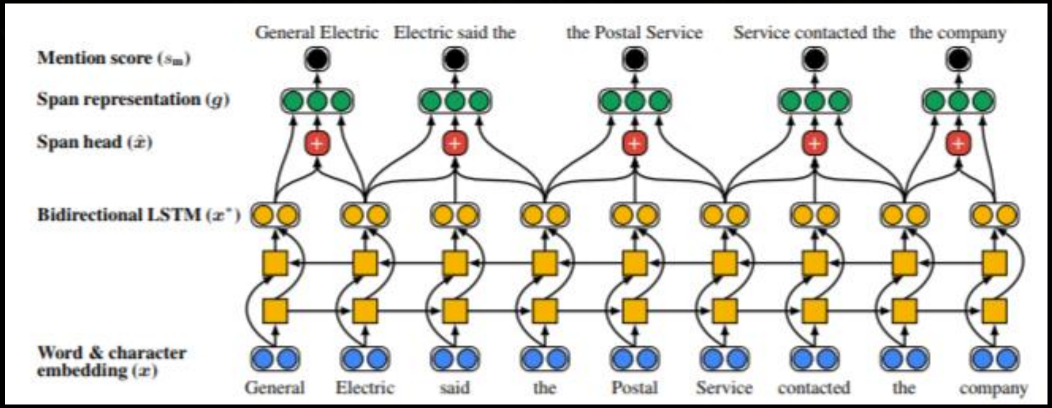
- Are current systems ignoring some important things?
- In this talk: looking back to **discourse**

Cohesion and coherence – J. Renkema

- ***Cohesion*** *is the connection which results when the interpretation of a textual element is dependent on another element in the text* (coreference, bridging, connectives…)
- ***Coherence*** *is the connection which is brought about by something outside the text* (e.g. world knowledge)

Where did discourse in coref go?

- Early work relating discourse to coreference showed problems with hard constraints
(Cristea et al. 1998, Poesio et al. 2002, Tetrault & Allen 2003…)
- Current coreference resolution systems model discourse implicitly
(Lee et al. 2017, Swayamdipta et al. 2018, Liu et al. 2019)



The good

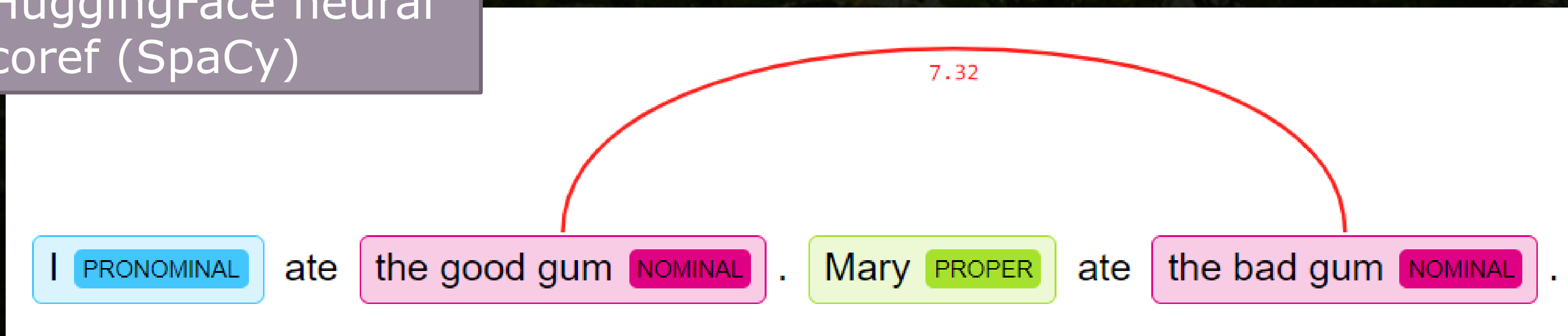
- (Contextual) embeddings allow relating OOV items to training data
- No need to curate KBs – just plug in a training corpus
- End to end architecture (e.g. Lee et al. 2017)
  - Avoids parsing error propagation
  - Recognizes quirky mention spans

The bad

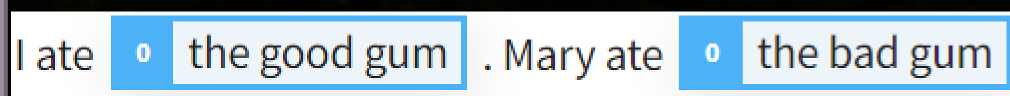
- No explicit semantic modeling
  - Synoymy/antonymy, cardinality
  - Models of entities in discourse, entity types
  - Overfit lexical features in data (Moosavi & Strube 2017, Webster et al. 2018)
- Rely heavily on pre-trained LM
  - Do not account for distributions in **current** text
  - Sensitive to changes in genre/domain
- No model of position in discourse structure
- Not viable for low resource languages (in this case: almost all languages…)

Antonymy

HuggingFace neural coref (SpaCy)



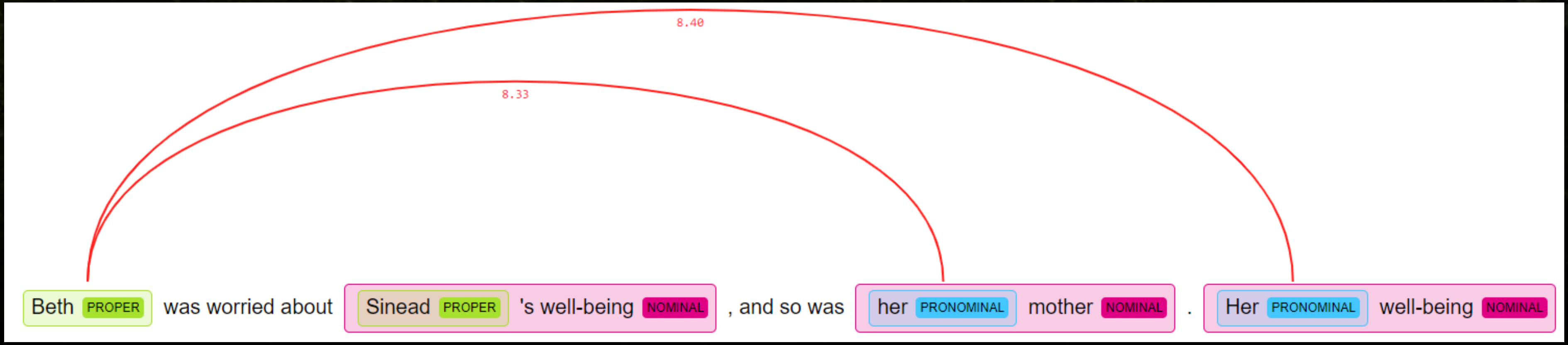
AllenNLP (e2e)



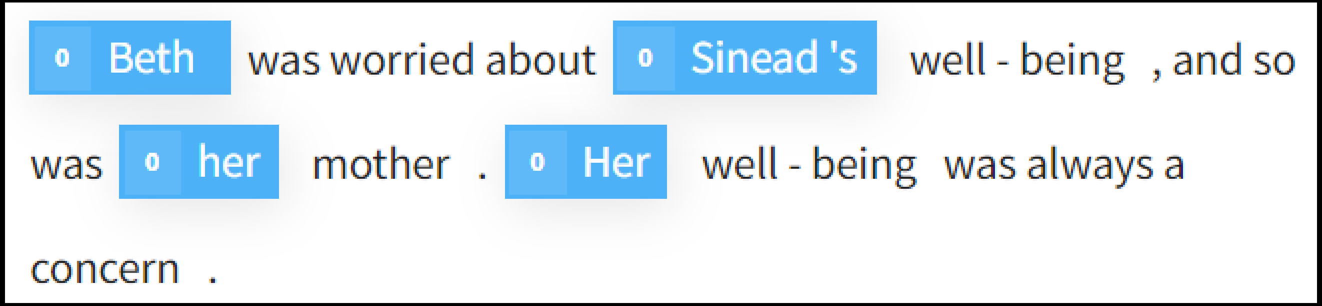
- Do we need more training data?
- Explicit lexicon/antonym feature?
- Discourse relations? (-->contrast)

Ad-hoc semantic models

- Cohesion can emerge at test time
- Pre-trained LMs can make wrong decisions:

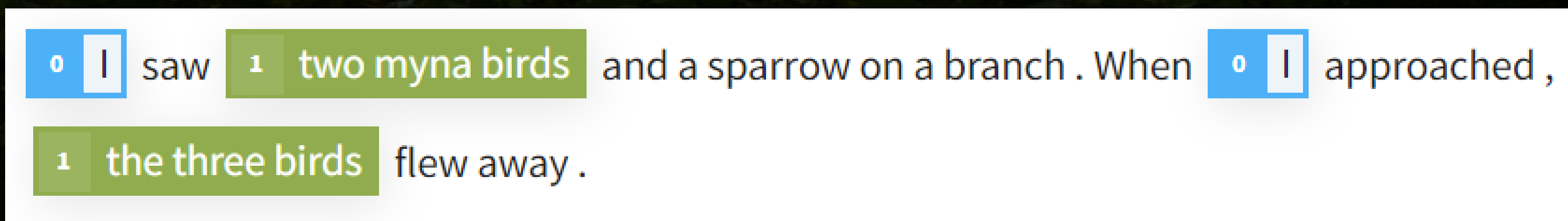


- Or just be wacky:



Cardinality - AllenNLP

- Minimal examples offer little ‘context’:



- Is it better in domain?

Cardinality - AllenNLP

- Not necessarily – from OntoNotes **train**:

0 The U.S. , claiming some success in 0 its trade diplomacy , removed 1 South Korea , Taiwan and Saudi Arabia from a list of countries 0 it is closely watching for allegedly failing to honor U.S. patents , copyrights and other intellectual - property rights . However , five other countries -- China , Thailand , India , Brazil and Mexico -- will remain on that so - called priority watch list as a result of an interim review , U.S. Trade Representative Carla Hills announced . Under the new U.S. trade law , 1 those countries could face accelerated unfair - trade investigations and stiff trade sanctions if 1 they do n't improve 1 their protection of intellectual property by next spring .

Part II

Mention ranking – do we need a crystal ball?

Centering Theory (Grosz et al. 1995)

- A theory about mentions in consecutive utterances
- Each utterance U_t has
  - Cf – forward looking centers – ordered list of mentioned entities by likelihood of next mention
  - Cb – a single entity linking back to the previous utterance
  - Cp – preferred center – rank 1 in Cf, most likely to be referred back to at U_{t+1}
- Ideally, Cb in U_{t+1} is Cf and Cb in U_t:
  - Continuation – Cb remains Cb and Cf
  - Retain – previous Cb is mentioned again but not longer Cp
  - Shift – current Cb is not previous Cb

Centering Theory (Grosz et al. 1995)

- Main claims:
  - *Constraint 1: All utterances of a segment except for the first have exactly one* ***Cb***
  - *Rule 1: If any* ***Cf*** *is pronominalized, the* ***Cb*** *is*
  - *Rule 2: (Sequences of) continuations are preferred over (sequences of) retains, which are preferred over (sequences of) shifts*

➢ How are Cf and Cb determined?

Ranking function

- Grosz et al.:
  - subj > obj > other
  - Other grammatical function hierarchies?
- Rambow (1993):
  - linear order (early -> salient)
- Strube & Hahn (1999):
  - Given > accessible > new
- Sidner (1979), Pearson et al. (2001):
  - Animate > inanimate (or other hierarchy…)
- Stevenson et al. (2000), Kehler & Rohde (2013):
  - Discourse function, connectives

Evaluation in previous work

- Poesio et al. (2004) survey a range of operationalizations of Centering
- Main conclusions:
  - *"Versions of Rule 1 make **very weak claims** about pronominalization"*
  - *"Strong C1 **does not hold**"* [modulo bridging]
  - *"Weak C1 .. **says nothing about entity coherence's** being what ensures local coherence"* [discourse relations are suggested instead]

Is Centering a good model of entity ranking?

- Why do we have the intuitions behind Centering if it's wrong?
  - Why is it actually wrong in the wild?
  - Can we reformulate it as a quantitative model?
  - Do we need discourse information?

- We need annotated data!

Data


THE WALL STREET JOURNAL.

What's News

Terror Strikes Europe's Heart

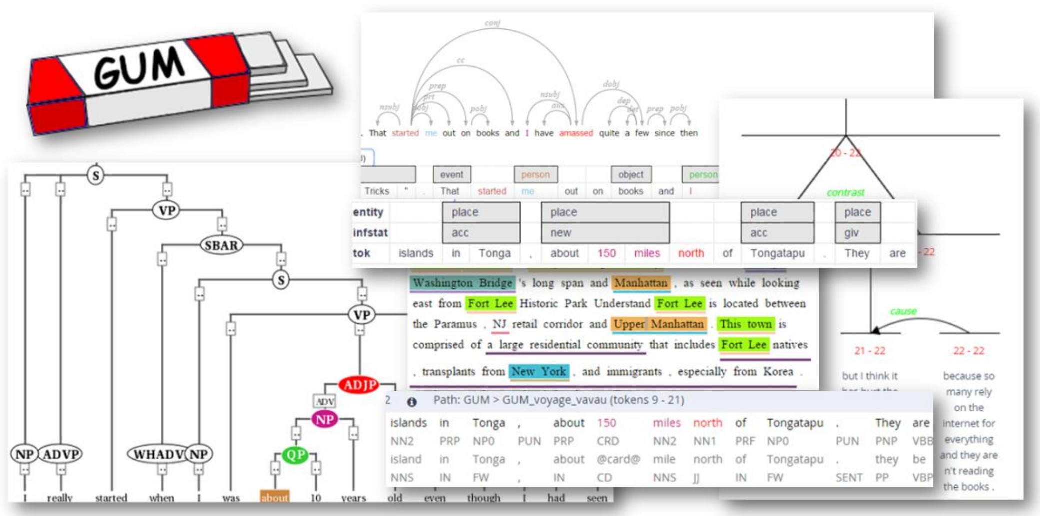
- Not many discourse + coref annotated corpora
- Use RST-DT ~ OntoNotes? (Carlson et al. 2003 + Hovy et al. 2006 – 182 documents overlap)
- But:
  - Only subset of anaphora reliably annotated
  - No singleton entity mentions for ranking
  - No bridging
  - Many other phenomena omitted (see Zeldes & Zhang 2016)

Phenomena not in OntoNotes

- **Indefinites/generics:** [*Program trading*] *is "a racket,"… [program trading] creates … swings*
- **Modifier nouns:** *small investors seem to be adapting to greater [stock market] volatility … Glenn Britta … is "factoring" [the market's] volatility "into investment decisions."*
- **Metonymy:** *a strict interpretation … requires [the U.S.] to notify foreign dictators of certain coup plots … [Washington] rejected the bid …*
- **Nesting:** *He has in tow [his prescient girlfriend, whose sassy retorts mark [her] …]*
- **Bridging:** *Mexico's President Salinas said [the country]'s recession had ended and [the economy] was growing again.*

The **G**eorgetown **U**niversity **M**ultilayer corpus

- POS tagging (PTB, CLAWS, TT, UPOS)
- Sentence type (SPAAC++)
- Document structure (TEI)
- Date/time expressions (ISO)
- Syntax trees (PTB + Stanford + UD)
- Information status (SFB632)
- Entity types (OntoNotes subset)
- **Coreference**
- **Bridging**
- **Rhetorical Structure Theory (RST)**



http://corpling.uis.georgetown.edu/gum/

- 8 genres (news, interview, forum, bio, fiction, how-to, travel, academic)
- 126 documents
- 109K tokens
- Freely available and growing!

Is Centering a good model of entity ranking?

- Data set:
  - 29K entity mentions from G
  - Full coref annotation (definite/indefinite, verbal, bridging…)
  - Rich annotations:
- Task:
  1. For each mentio ned again in next dis
  2. Exhaustively **rank** all mentions for subsequent mention likelihood (=fill out Cf)

NB: This is totally unreasonable!

[a] cloze task is a measure of ***prescience*** *— whether […] model can predict events based on those that co-occurred with it*

(Simonson 2018)

* Span scoring in neural coref systems does something like this!

Let's try it!

- Can you guess/rank which entities will be mentioned in the next sentence?
  - *[One indication of [the importance of [replication]]] is found in [the 50 or more calls] for [[replication] research] in [the field of [[second language (L2)] research] alone*

(see [references for [50 calls] and [commentaries] in [Appendix S1] in [the Supporting Information] online)

A linear model?

- All suggested ranking factors definitely significant:

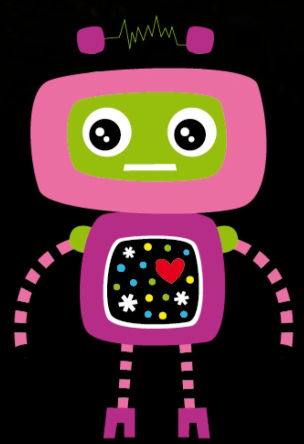
```
            Df Sum Sq Mean Sq F value    Pr(>F)
gram_func    6    193   32.23 228.109  < 2e-16 ***
infstat      3    261   87.02 615.794  < 2e-16 ***
animate      1     94   93.94 664.789  < 2e-16 ***
sent_posit   1     10   10.33  73.069  < 2e-16 ***
disc_func   21     14    0.65   4.603 1.18e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction accuracy

- As additive effects, very negligible improvement over majority baseline
  - Majority baseline: 79.82
  - Linear model: 80.27

Remember baseline = never refer back = "deranged chatbot"

- The Lakers won again.
- Eclairs are delicious!
- She's your movie.
- Did I call him?
- …



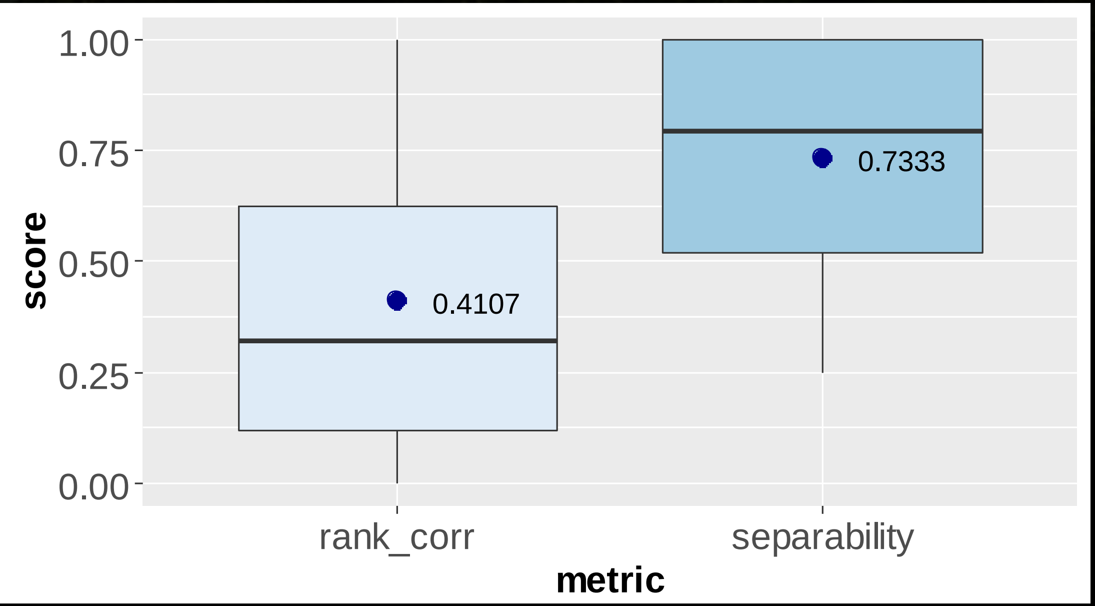
Is this something that humans can do?

- Experiment:
  - 6 human raters for 46 entities in 10 sentences
  - Give complete ranking within each sentence
  - Alternative scenario: binary classification – will/won't be mentioned in next sentence
- Prediction accuracy:
  - Binary yes/no 67.39%
  - Rank 1=yes 73.91%
  - Mean rank correlation r=.4108

Many thanks to:
Corpling@GU

Why are humans bad at this?

- Disagree on arbitrary bad candidate order
- Tendency to ask "could I imagine…?"
- More lenient % separability metric still 73.3



Can an RNN get this from text?

- RNN with concatenated pretrained:
  - Fixed word embeddings (GloVe, Pennington et al. 2014)
  - Contextual embeddings (Flair, Akbik et al. 2018)
  - Character embeddings (AllenNLP)
- Fine tuning
- Concatenate sentence and mention representations -> encoder + binary clf

➢ Prediction accuracy: 82.22

➢ No improvement from adding previous sentence context

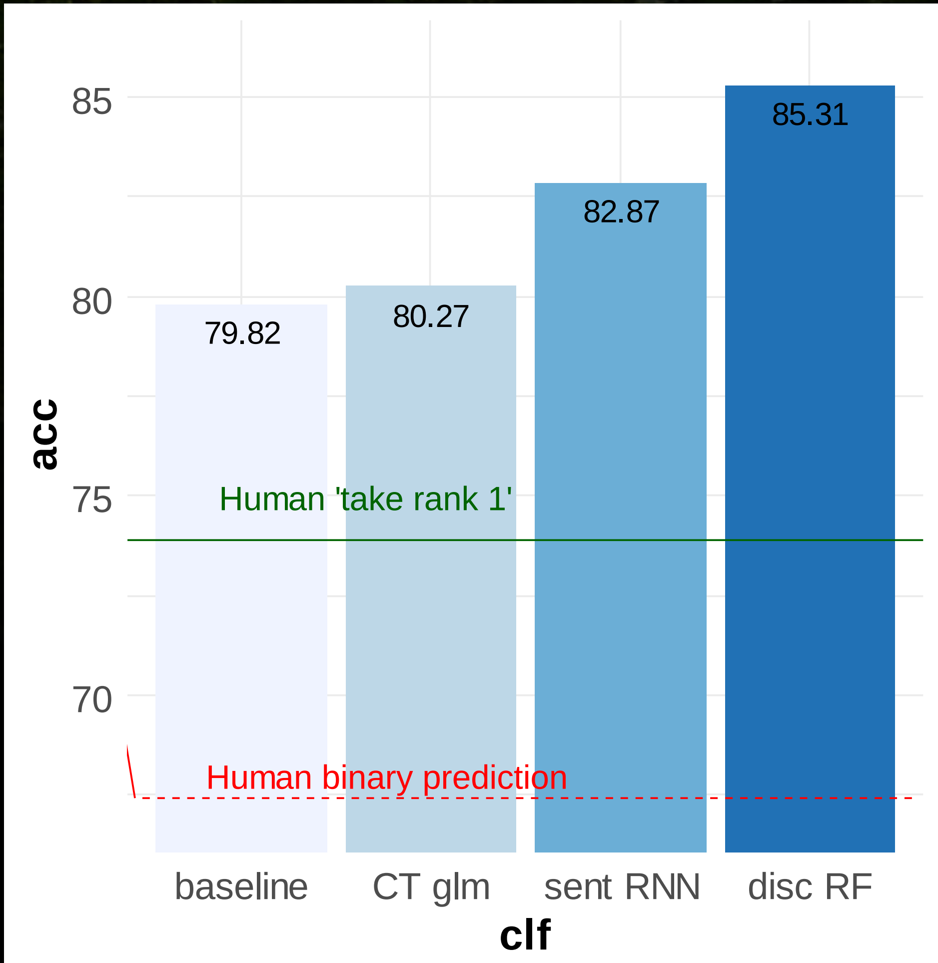
Local cues for Cf are weak!

What about non-local features?

- Silly confounds
  - Next sentence length! (U_{t+1})
- Entity features
  - Salient entities typically **discussed previously**
  - Typically mentioned **recently**
- Discourse features
  - Genre
  - Position in document
  - Discourse tree (RST)
    - Labels for U_t, U_{t+1}
    - Distance to parents

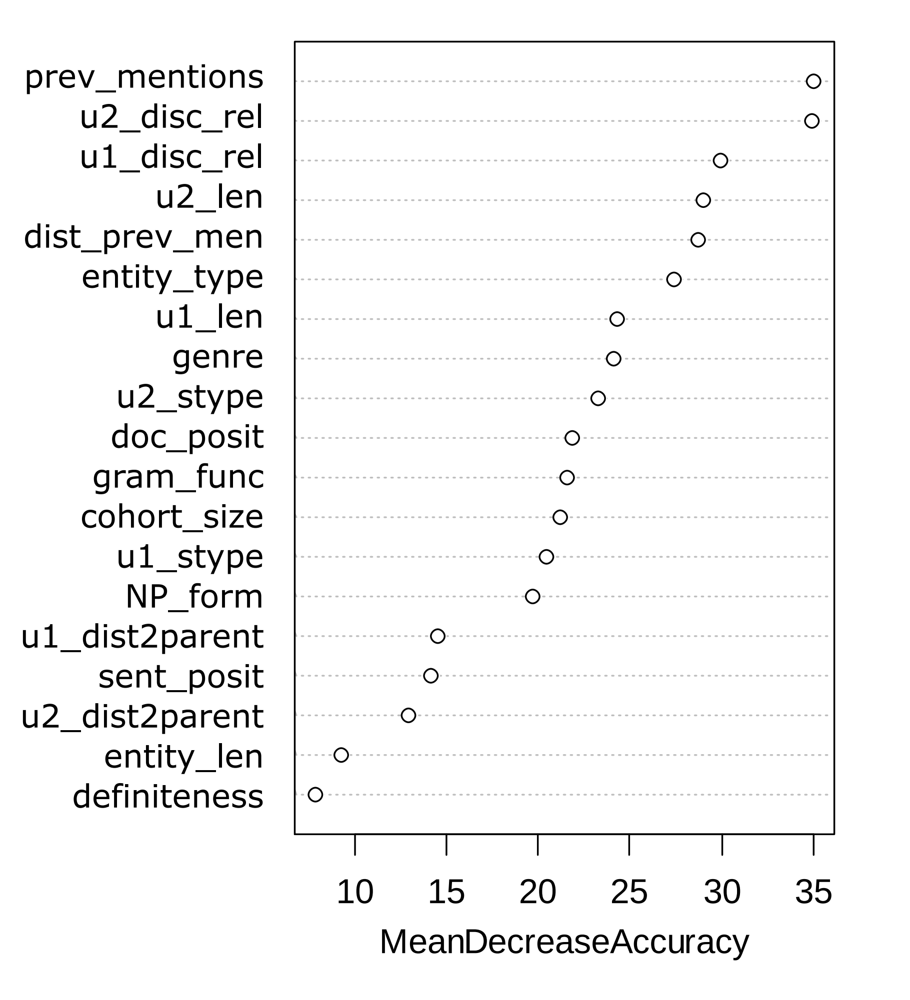
Results

- Feed features to Random Forest classifier
- Non-local model performs better
  - Does data from discourse help?
  - Or is the RNN just overfitting?
- Need to look at feature contributions

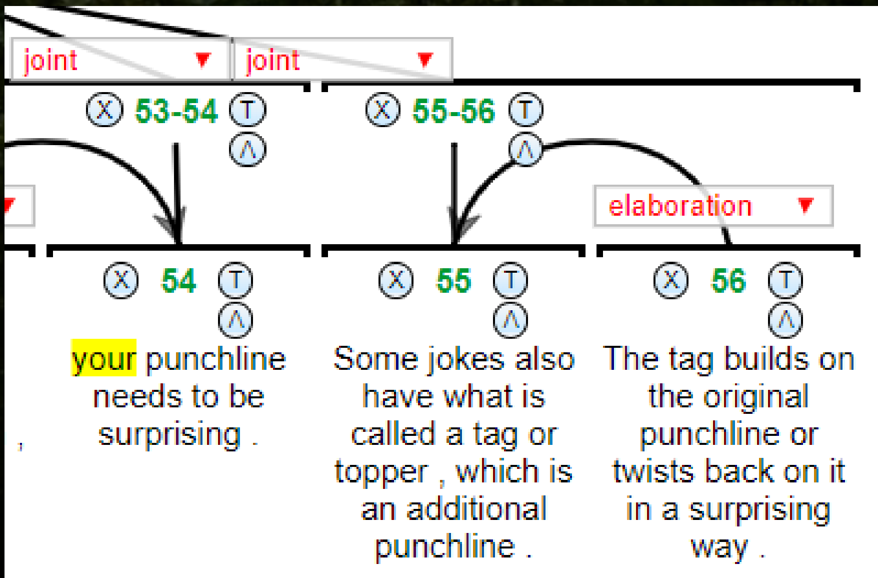


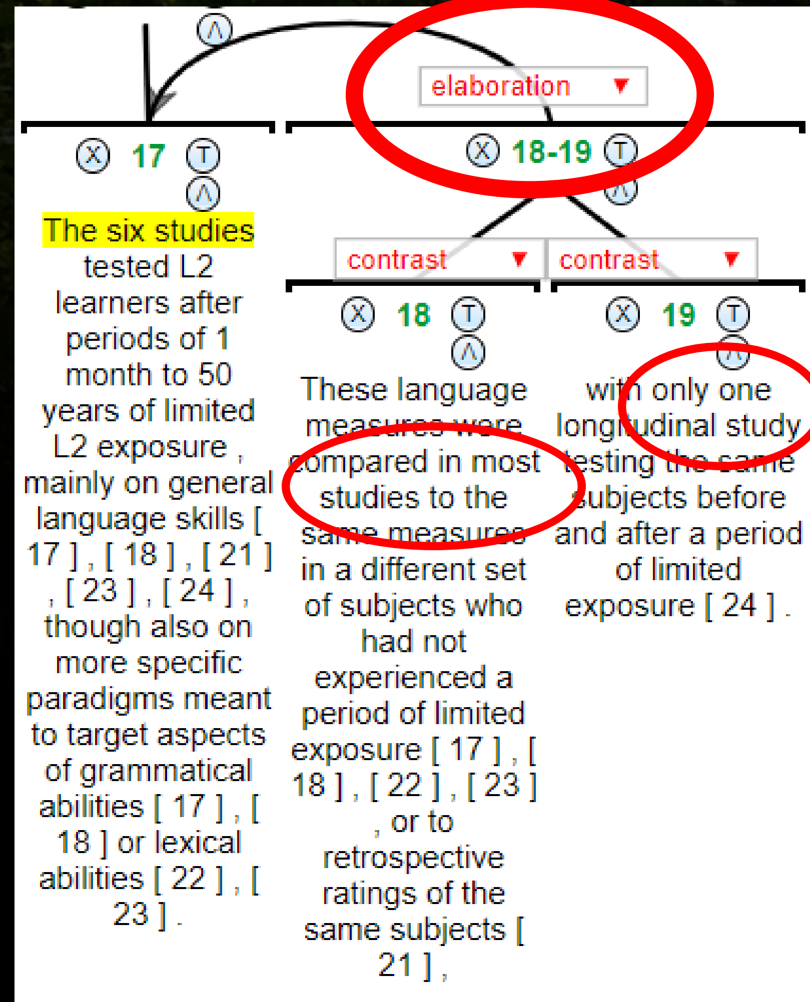
Feature permutation importances

- Top 5 are all non-local features!
- Next unit length only 4th place…
- Relation types outrank all but prev. mentions
  - NB MDA under-rates mutual redundancy!
  - But relations are irreplaceable, not redundant with other discourse features
  - Confirms discourse constraints on coreference



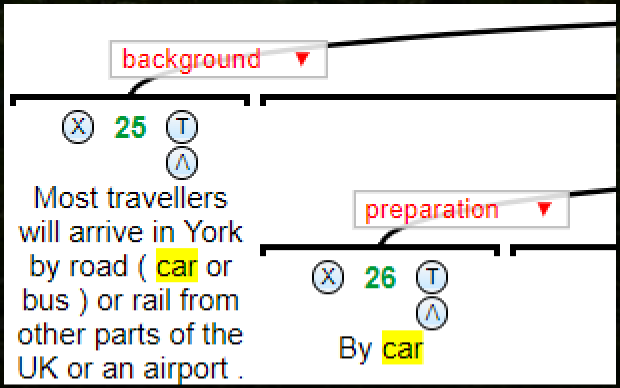
Error analysis – false positives

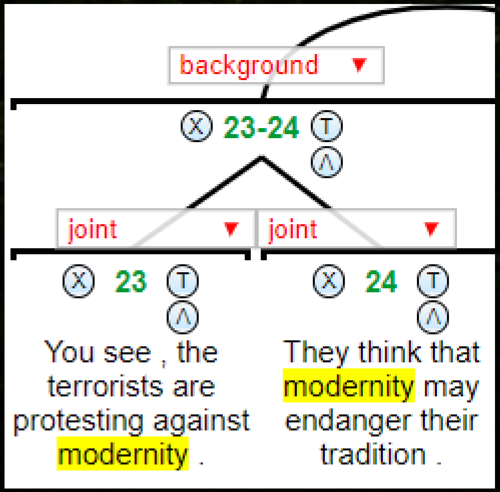




- elaboration, joint
- pronominal/definite
- prev. mentions
- early in sentence
- subject

Error analysis – false negatives



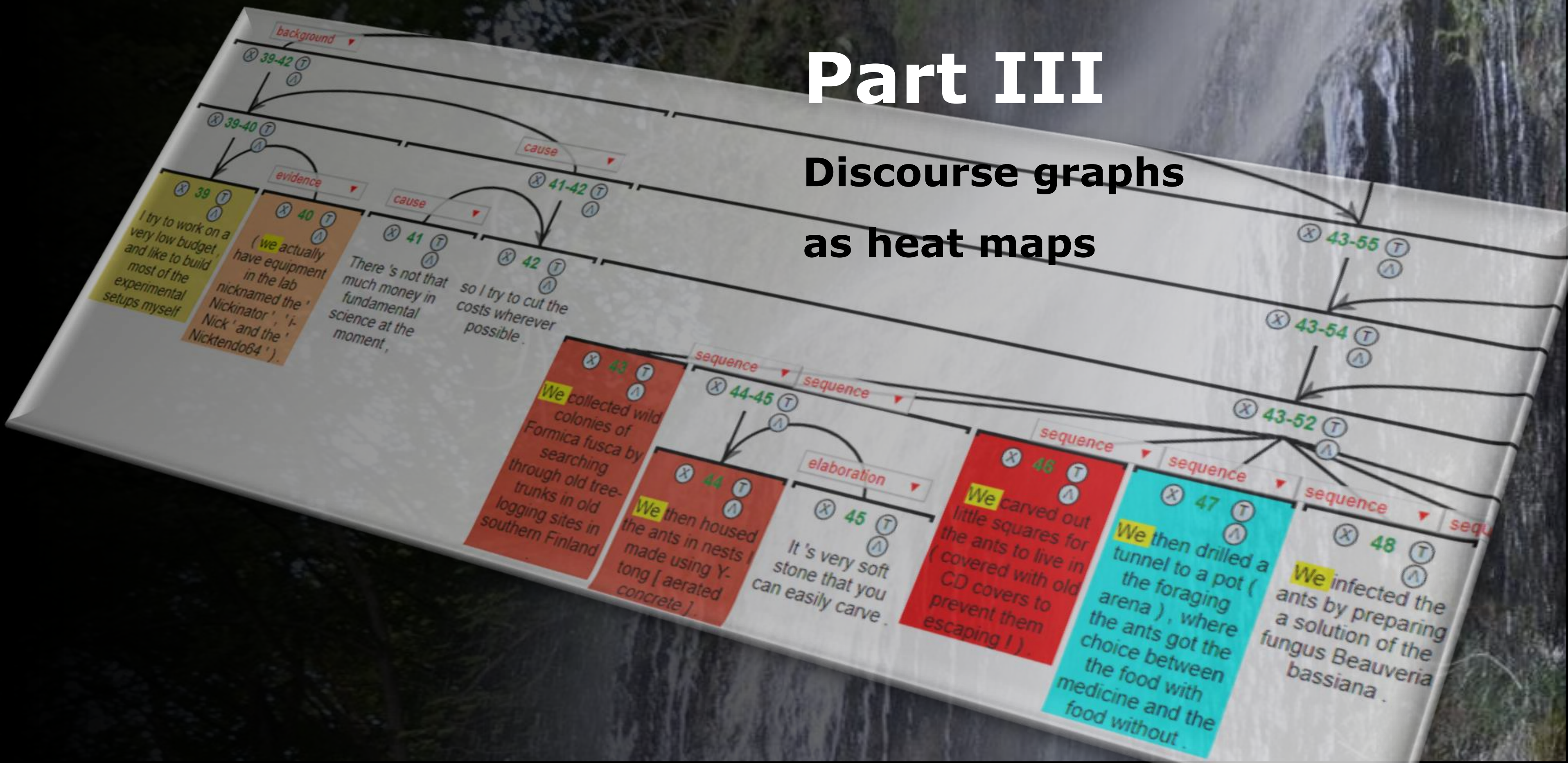


Can e2e sent + span models know this?

- not previously mentioned
- discourse-disjoint
- background, prep
- non-subj
- inanimate
- late in sent
- indefinite
- short, common
- non-continuing structures (cf. VT)
- …

Part III

Discourse graphs as heat maps

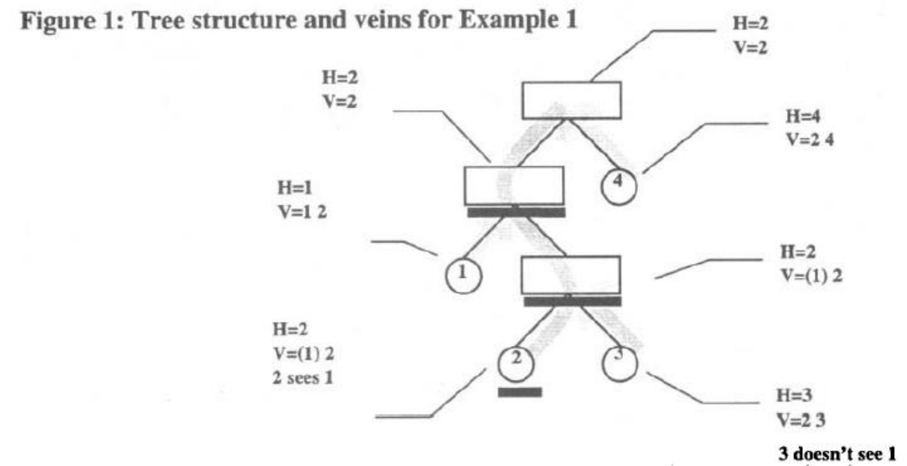


Categorical discourse constraints

- Early computational work suggested forms of "discourse encapsulation":
  - Stack models (Polanyi 1988)
  - Veins Theory (Cristea et al. 1998)
  - Right Frontier Constraint (Asher & Lascarides 2003)

Categorical discourse constraints

- Right frontier constraint (SDRT) – narration blocks back reference
  - *John ate salmon. Then he won a dance competition. #**It** was a beautiful pink*
    (cf. Asher & Vieu 2005)
- VT postulates **Domains of Referential Accessibility (DRAs)**
  - discourse units can ‘see’ their modifiers
  - Modifiers can only access their parents



Categorical discourse constraints

- Problematic in practice:
  - Tetreault & Allen (2003:7) on Veins Theory:

    Our results indicate that ***incorporating discourse structure*** *does not improve performance, and in most cases can actually* ***hurt performance****.*

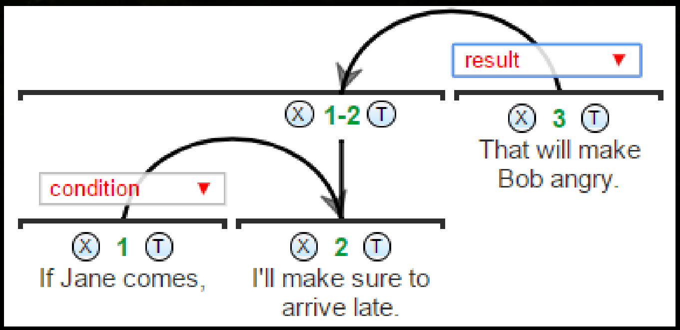
Some research questions

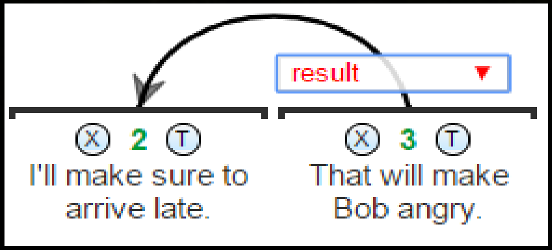
- Are discourse constraints on coref 'wrong'?
  - If so why the intuitions?
  - If not, what's the problem?

- I suggest at least two kinds of problems: (Zeldes 2017)
  - Confounds
  - Need for quantitative interpretation

RST and Rhetorical Distance (RD)

- We want a quantitative notion of 'veins'
- Distance between Elementary Discourse Units (**EDUs**)
- Using non-terminal spans is problematic:



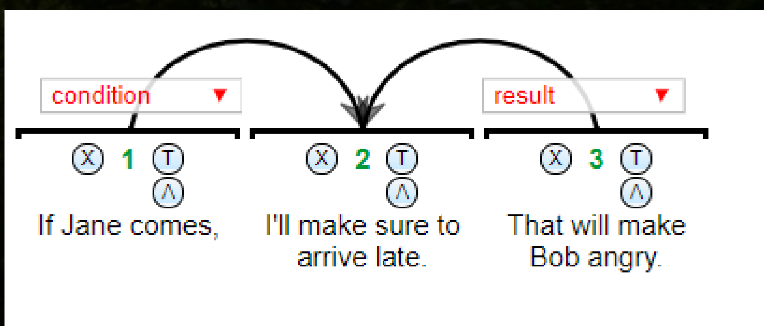


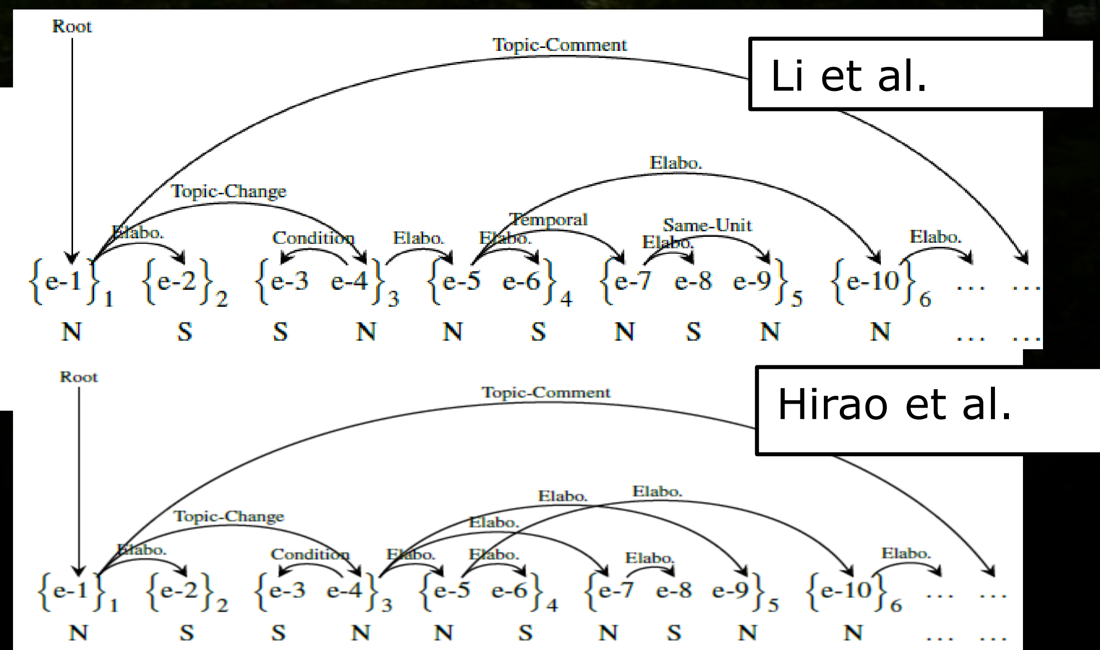
RD(2,3) = 2

RD(2,3) = 1

Switch to Dependency Representation

- Following Hayashi et al. (2016), use Li et al.'s (2014) dependency interpretation*

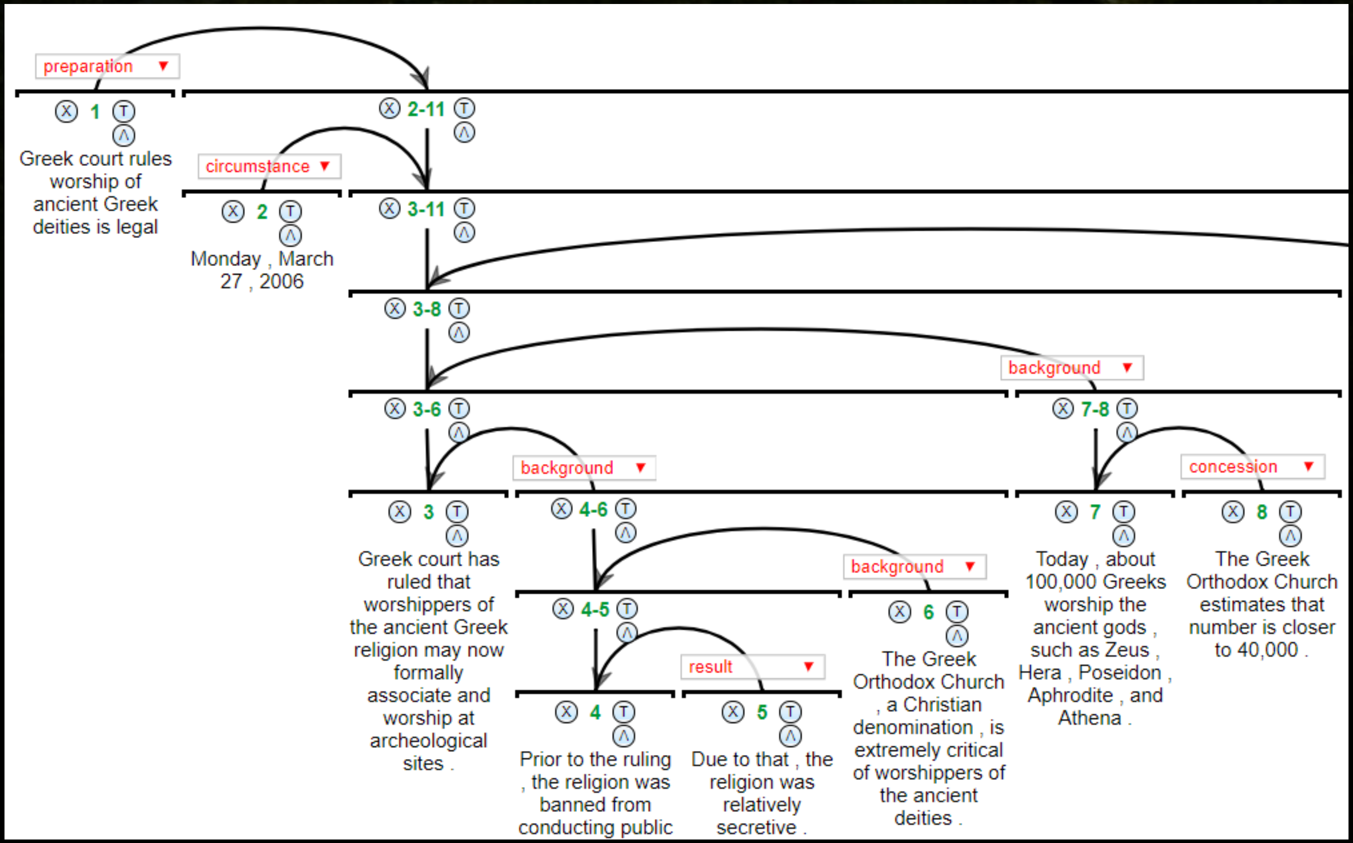




* conversion code from .rs3 available at: https://github.com/amir-zeldes/rst2dep

Operationalizing the parent vein

- Ancestry: Is one EDU a direct ancestor of the other in the dependency tree?



Target variable

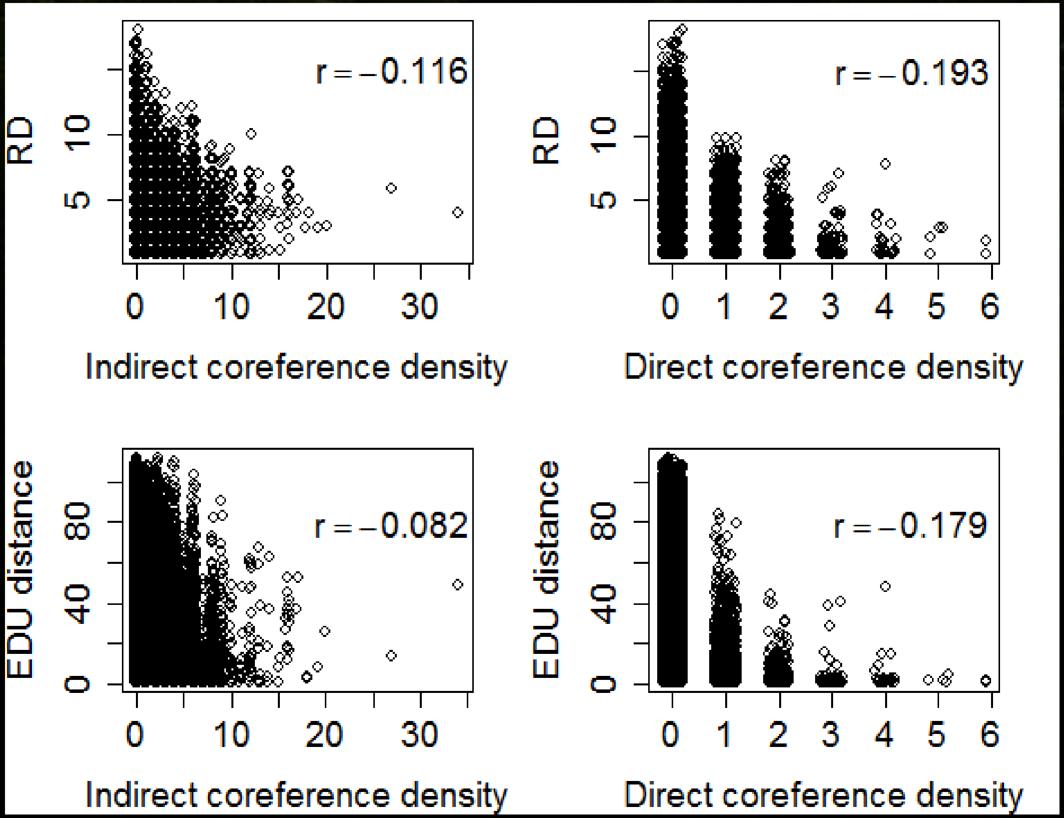
- What are we trying to predict?
  - Binary domains:
    - Can there be coreference between two EDUs?
    - Explore for each coreference type
  - Coreference **density**:
    - How much coreferentiality exists between two EDUs? (# coreferent pairs)
  - Direct and indirect antecedents:
    - Check if the **immediate antecedent** of entity in EDU2 is in EDU1 (NB: makes surface distance very important!)
    - Alternatively, just check for coreference

What's more important?

- As a first objective we can check the relative importance of:
  - Surface distance
  - Rhetorical distance
  - Direct ancestry
- ~170K possible EDU pairs grouped by document
- 10% data held out for testing, stratified by coreference density

Only weak correlations…

- For all EDU pairs:
  - Most have 0 coreference
  - Especially direct antecedents have very low distance
  - Not much predictability (cf. Tetreault & Allen) ✓ ☹



Why is RD weak despite intuition?

- Again, lots of **confounds**!!
  - **Length:** what if the main RST trunk nucleus is really short? -> Unlikely to contain coreferent mentions
  - **Relations:** not all satellites are equal -> ***Purpose*** rarely exhibits coreference; ***Cause*** often does!
  - **Sentence type:** imperatives and fragments have fewer entities than declaratives and questions
  - … + tense, genre, syntactic function, POS, document position, …

Is RD significant? Gaussian mixed model

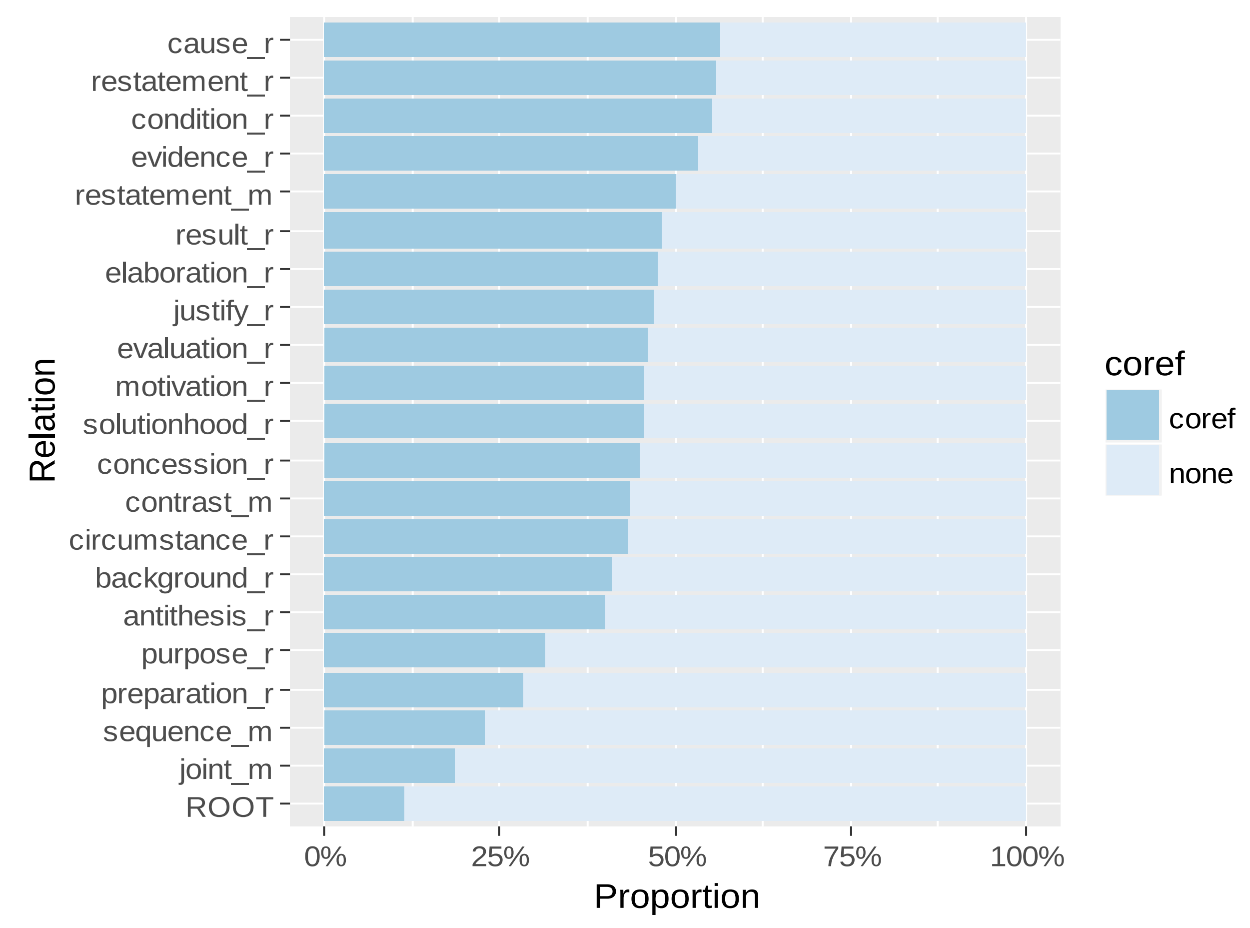
- Yes, and so is surface distance!
- But not as important as length

```
Random effects:
 Groups   Name        Variance Std.Dev.
 doc      (Intercept) 0.09789  0.3129
 Residual             0.82965  0.9109
Number of obs: 172150, groups:  doc, 76

Fixed effects:
              Estimate Std. Error t value
(Intercept)  0.2695836  0.0723038    3.73
scale(len1)  0.2043943  0.0023432   87.23
scale(len2)  0.1833124  0.0023811   76.99
rsd_dist    -0.0511588  0.0014351  -35.65
edu_dist    -0.0015377  0.0001168  -13.17
genrenews   -0.0348780  0.0997936   -0.35
genrevoyage -0.2161897  0.1047555   -2.06
genrewhow    0.0969725  0.1016942    0.95
directTrue   0.2280120  0.0091334   24.96
```

Which discourse relations favor coreference?

- Unsurprisingly:
  - ↑Restatement, Cause
  - …
  - ↓ Joint, Sequence



Putting it all together

- Taken in isolation we can’t interactions between factors:
  - Restatements favor coref … unless short?
  - Can direct ancestry overturn high RD?
  - Questions are high-density while shorter than declaratives…?
- A model knowing all of this together can make better decisions than linear regression
- Back to a tree ensemble

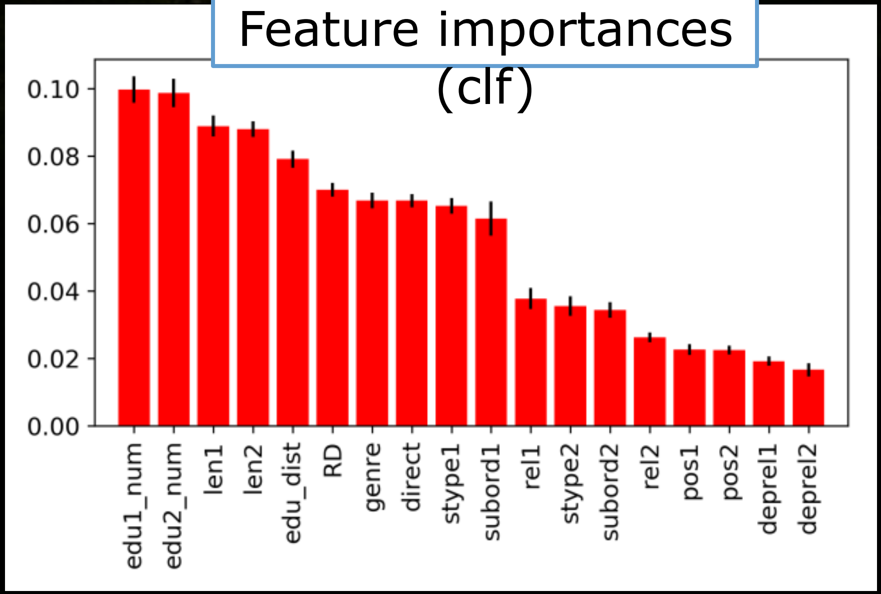
Results

- Two settings: classification (coref yes/no), regression (predict density)

Performance

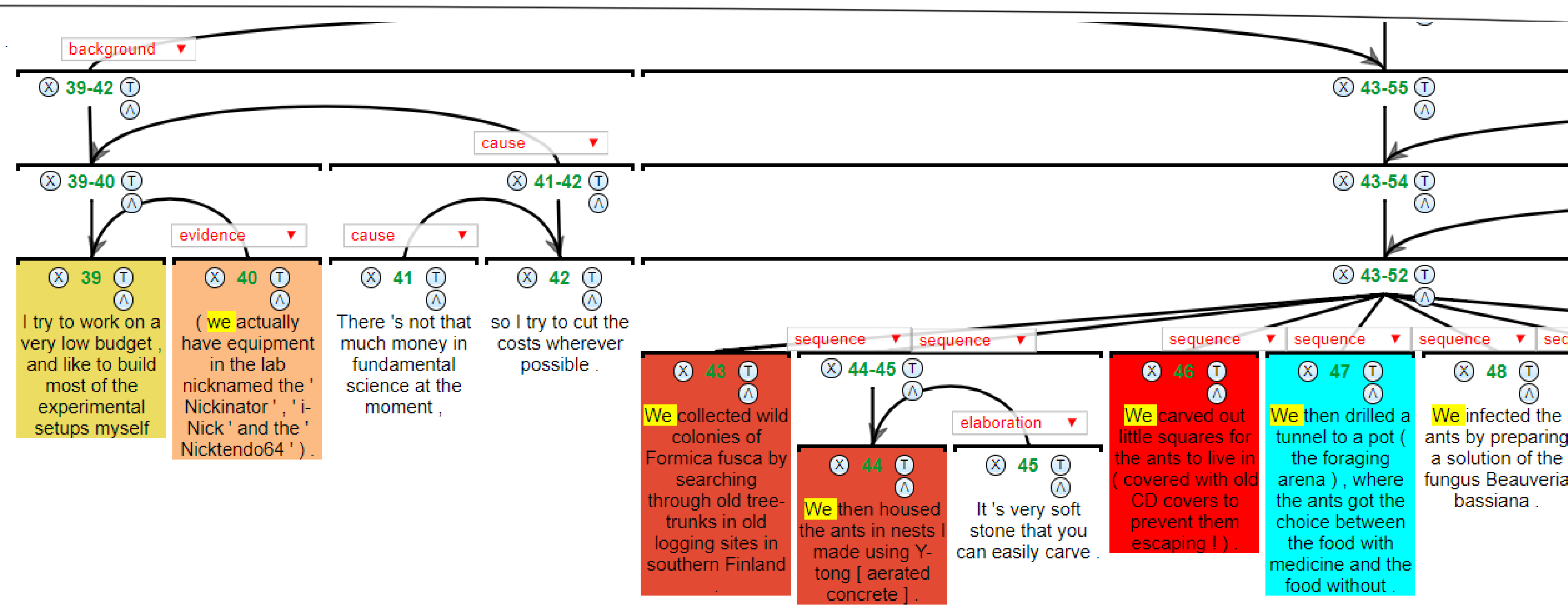
features	RMSE (reg)	accuracy (clf)
EDU	0.9501	78.36%
RD	0.9453	78.79%
all	0.7107	86.83%

Feature importances (clf)

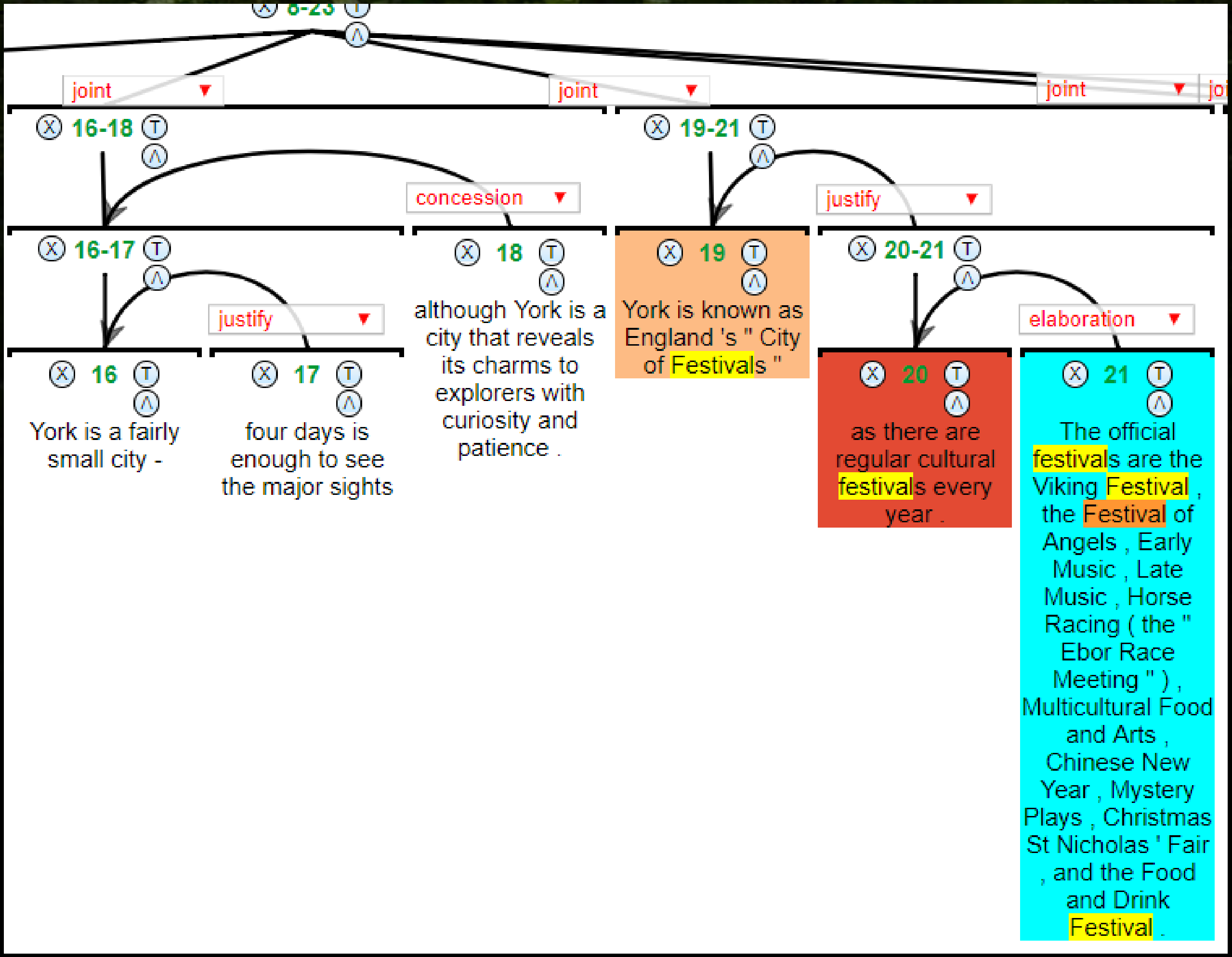


What do the predictions look like?

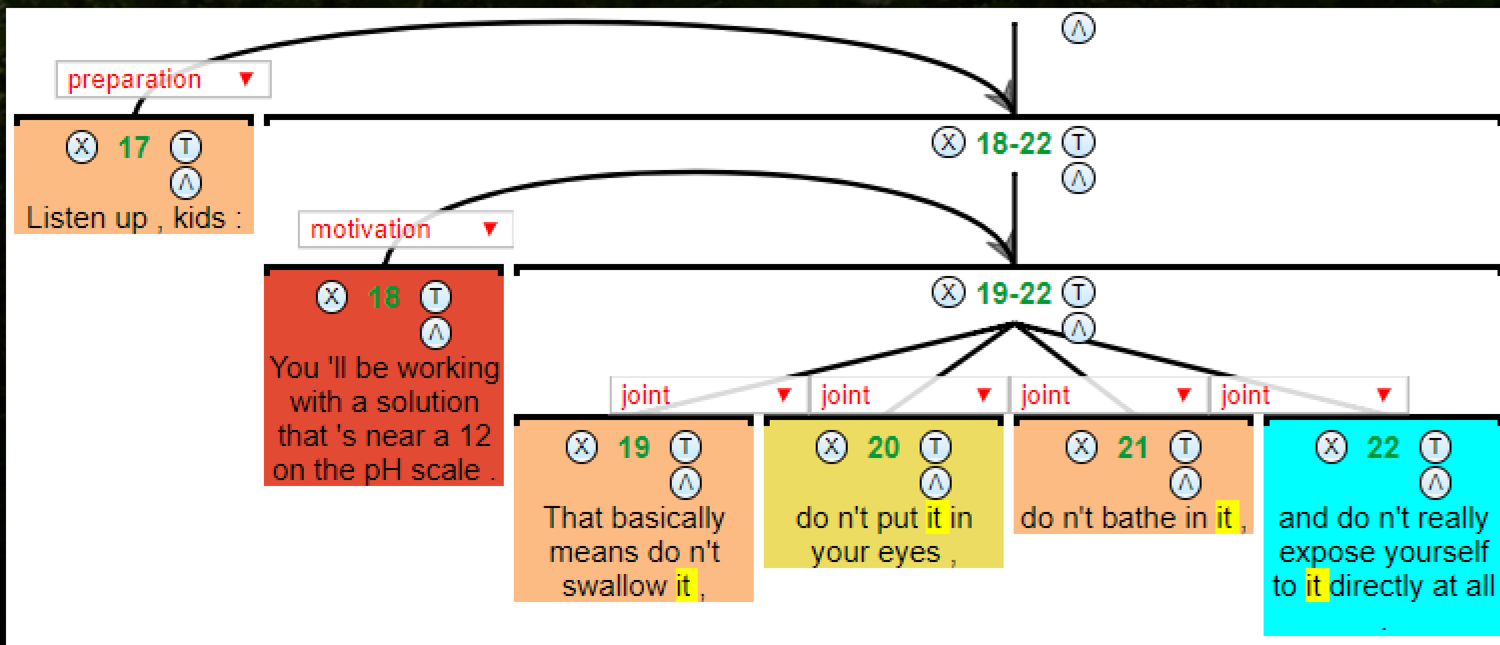
- We can visualize predictions as a heat map:



What do the predictions look like?



What do the predictions look like?



Conclusion

- There are good reasons to think coherence and coreference are related
- We do not have good ways of representing discourse effects in whole paragraph/document
  - Not enough training data in OntoNotes to use much larger contexts
  - Pairwise comparisons become expensive
  - Other methods using paragraph/document vectors?
  - Categorical/numerical feature representation?
  - Use predicted discourse parses? (getting much better, see Braud et al. 2017, Lin et al. 2019!)

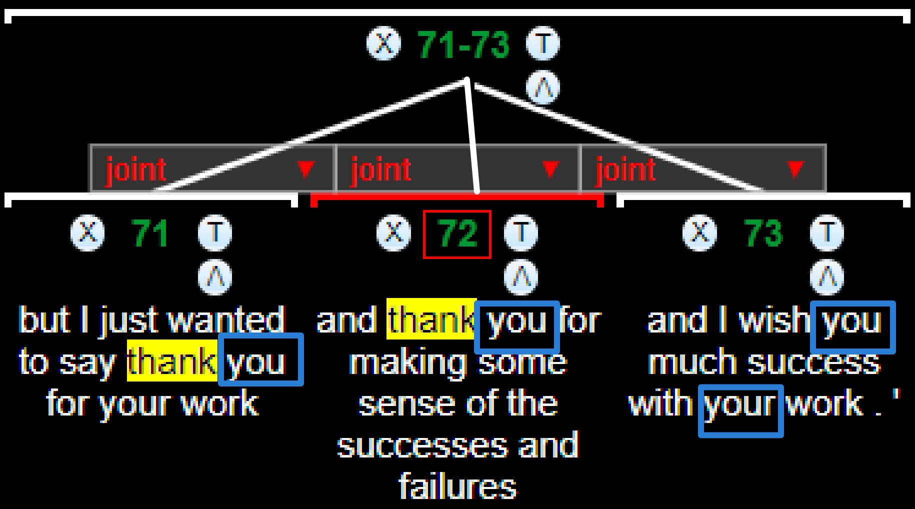
Future work

- We are looking at **signals** of discourse relations at all levels (Liu & Zeldes 2019)
- Coreference is one of the cues that models for relation extraction can learn to attend to:

Microsoft has launched an aggressive campaign to persuade users to stop using IE6 <--ELAB-- **Its** goal is to decrease IE6 users to less than one percent .

- Hopefully more soon!

Thanks!



--GUM_interview_messina