Deciding entailment and contradiction with stochastic and edit distance-based alignment

Marie-Catherine de Marneffe,

Sebastian Pado, Bill MacCartney, Anna N. Rafferty, Eric Yeh and Christopher D. Manning

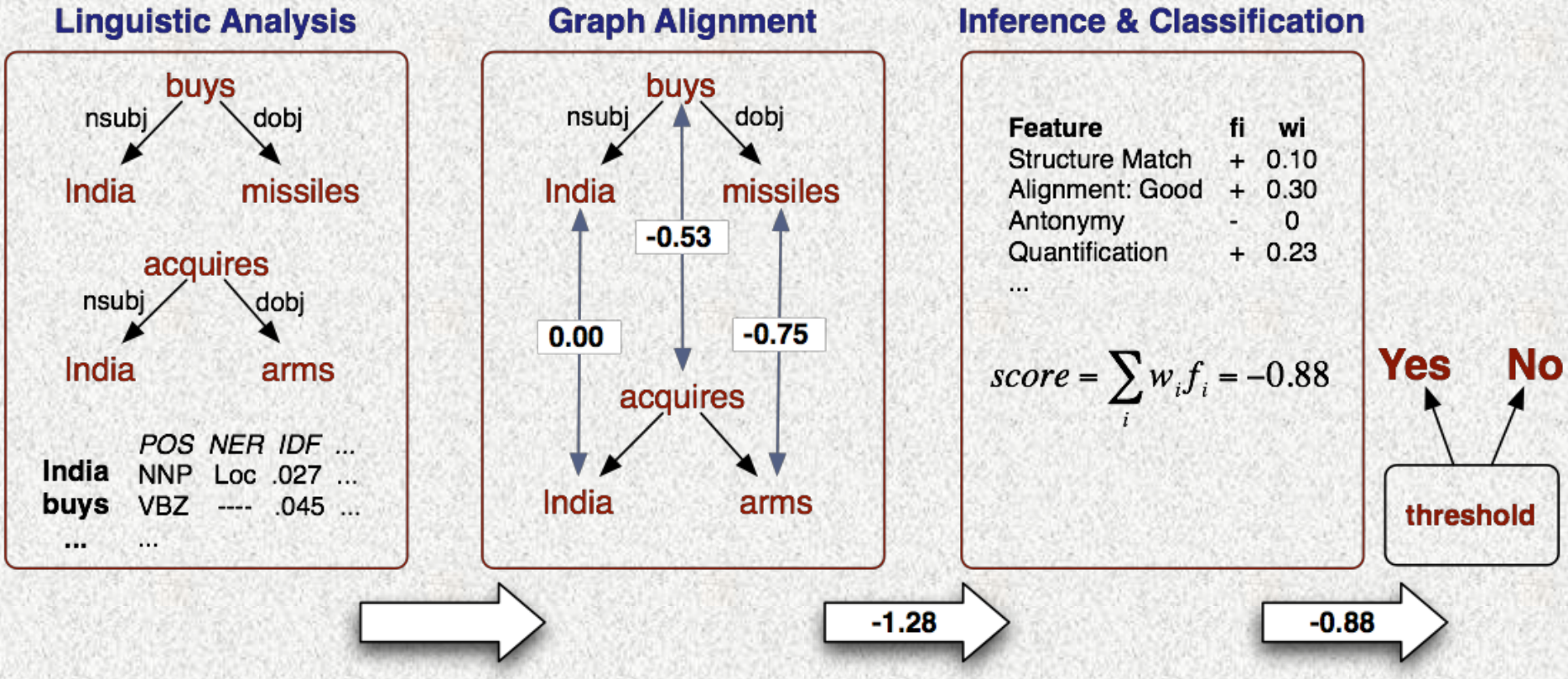
NLP Group

Stanford University

Three-stage architecture

[MacCartney et al. NAACL 06]

T: India buys missiles.
H: India acquires arms.



Attempts to improve the different stages

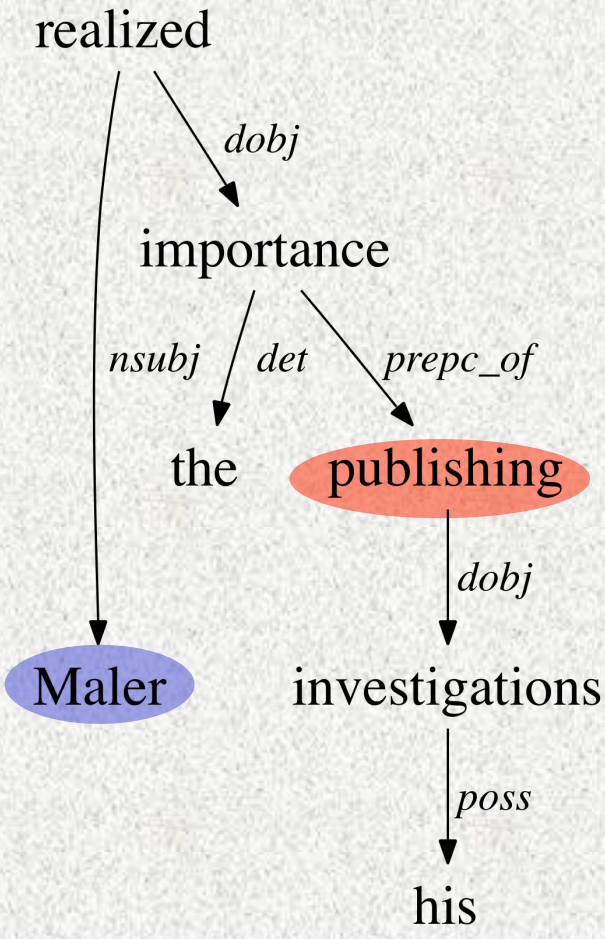
1) Linguistic analysis:
   improving dependency graphs
   improving coreference

2) New alignment:
   edit distance-based alignment

3) Inference:
   entailment and contradiction

Stage 1 – Capturing long dependencies

Maler realized the importance of publishing his investigations



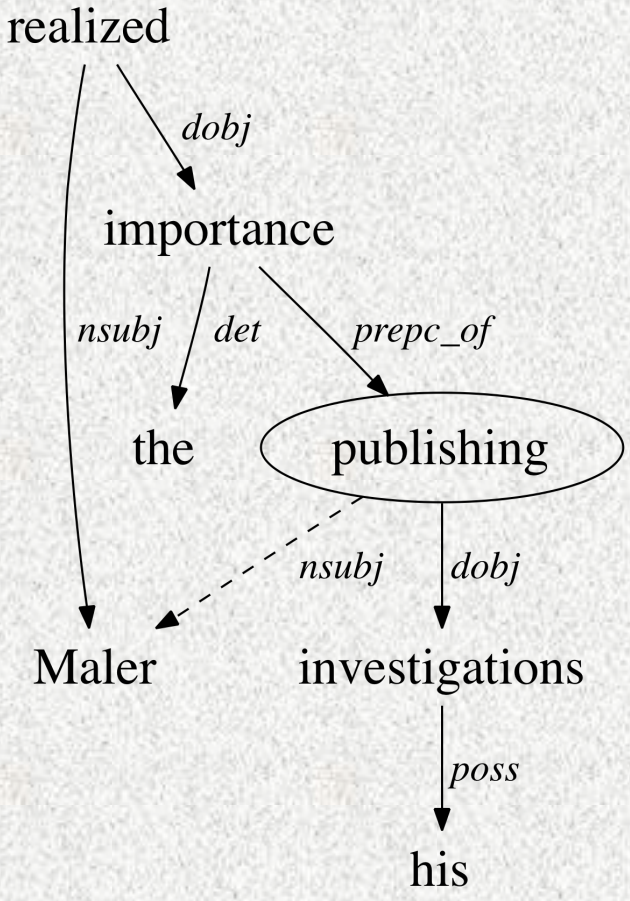
Recovering long dependencies

- Training on dependency annotations in the WSJ segment of the Penn Treebank
- 3 MaxEnt classifiers:

1) Identify governor nodes that are likely to have a missing relationship

2) Identify the type of GR

3) Find the likeliest dependent (given GR and governor)



Some impact on RTE

- Cannot handle conjoined dependents:

  Pierre Curie and his wife realized the importance of advertising their discovery

- RTE results:

	Accuracy	With recovery
RTE2 test	61.25	63.38
RTE3 test	65.25	66.50
RTE4	62.60	62.70

Coreference with ILP

[Finkel and Manning ACL 08]

- Train pairwise classifier to make coreference decisions over pairs of mentions
- Use integer linear programming (ILP) to find best global solution
  - Normally pairwise classifiers enforce transitivity in an ad-hoc manner
  - ILP enforces transitivity by construction
- Candidates:

  all based-NP in the text and the hypothesis
- No difference in results compared to the OpenNLP coreference system

Stage II – Previous stochastic aligner

$$\text{score}(a) = \sum_{i \in h} \text{score}_w(h_i, a(h_i)) + \sum_{(i,j) \in e(h)} \text{score}_e((h_i, h_j), (a(h_i), a(h_j)))$$

Word alignment scores: semantic similarity

Edge alignment scores: structural similarity

- Linear model form: $s_w(h_i, t_j) = \theta_w \cdot \text{f}(h_i, t_j)$

  $s_e((h_i, h_j), (t_k, t_l)) = \theta_e \cdot \text{f}((h_i, h_j), (t_k, t_l))$
- Perceptron learning of weights

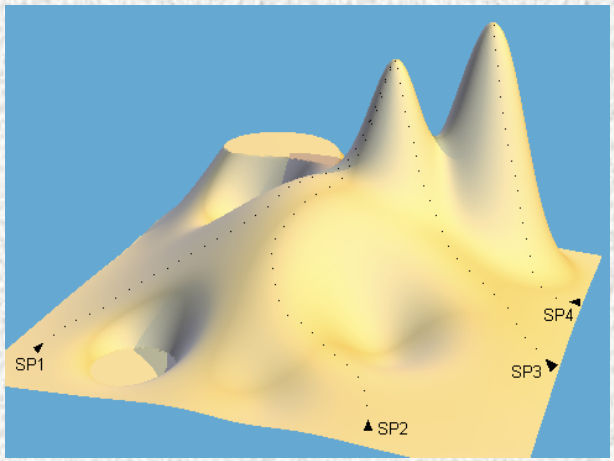
Stochastic local search for alignments

Complete state formulation

Start with a (possibly bad) complete solution, and try to improve it

At each step, select hypothesis word and generate all possible alignments

Sample successor alignment from normalized distribution, and repeat



New aligner: MANLI

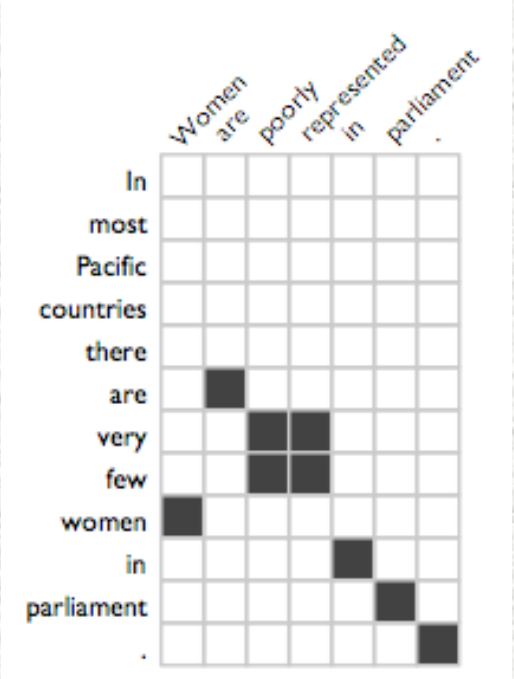
[MacCartney et al. EMNLP 08]

4 components:

1. Phrase-based representation
2. Feature-based scoring function
3. Decoding using simulated annealing
4. Perceptron learning on MSR RTE2 alignment data

Phrase-based alignment representation

An alignment is a sequence of phrase edits: EQ, SUB, DEL, INS



DEL(In_1)
…
DEL($there_5$)
EQ(are_6, are_2)
SUB($very_7$ few_8, $poorly_3$ $represented_4$)
EQ($women_9$, $women_1$)
EQ(in_{10}, in_5)
EQ($parliament_{11}$, $parliament_6$)

- 1-to-1 at phrase level but many-to-many at token level:
  avoids arbitrary alignment choices
  can use phrase-based resources

A feature-based scoring function

- Score edits as linear combination of features, then sum:

$$s(E) = \sum_{e \in E} s(e) = \sum_{e \in E} \sum_{i} w_i \cdot \phi_i(e)$$

- Edit type features:

  EQ, SUB, DEL, INS

- Phrase features:

  phrase sizes, non-constituents

- Lexical similarity feature (max over similarity scores)

  WordNet, distributional similarity, string/lemma similarity

- Contextual features:

  distortion, matching neighbors

RTE4 results

	2-way	3-way	Av. P
stochastic	61.4	55.3	44.2
MANLI	57.0	50.1	54.3

Error analysis

- MANLI alignments are sparse
  - sure/possible alignments in MSR data
  - need more paraphrase information

- Difference between previous RTE data and RTE4:
  length ratio between text and hypothesis

	RTE1	RTE3	RTE4
T/H	2:1	3:1	4:1

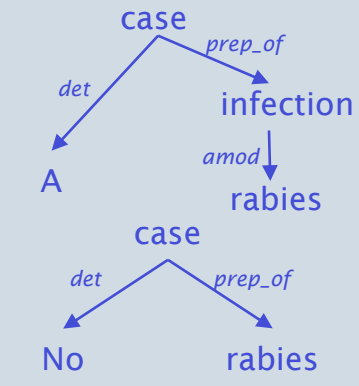
- All else being equal, a longer text makes it likelier that a hypothesis can get over the threshold

Stage III – Contradiction detection

[de Marneffe et al. ACL 08]

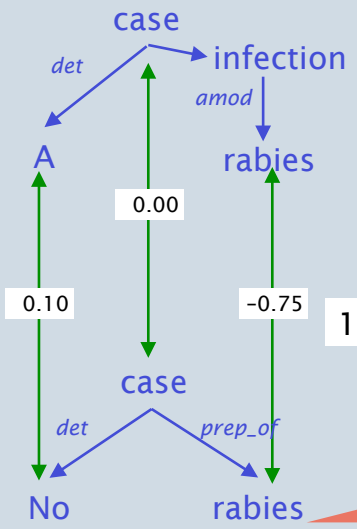
T: A case of indigenously acquired rabies infection has been confirmed.
H: No case of rabies was confirmed.

1. Linguistic analysis



rabies	POS NER IDF	NNS -- 0.027
…	…	…

2. Graph alignment



1.84

3. Contradiction features & classification

Feature	f_i	w_i
Polarity difference	-	-2.00

score = $\sum_i w_i f_i$ = -2.00

contradicts

tuned threshold

doesn't contradict

Event coreference

Event coreference is necessary for contradiction detection

- The contradiction features look for mismatching information between the text and hypothesis
- Problematic if the two sentences do not describe the same event

T: More than 2,000 people lost their lives in the devastating Johnstown Flood.

H: 100 or more people lost their lives in a ferry sinking.

Mismatching information:

more than 2,000 != 100 or more

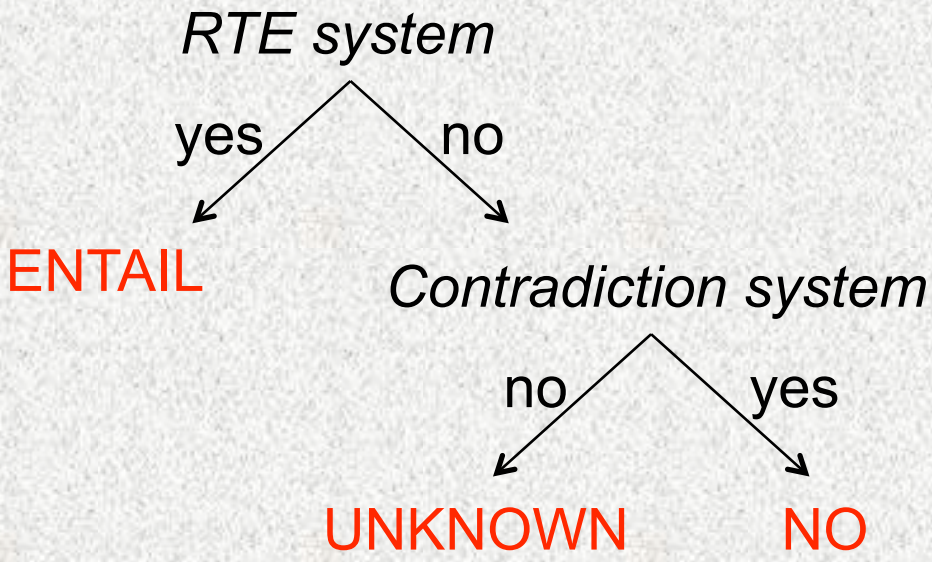
Contradiction features

RTE	Contradiction
Polarity	Polarity
Number, date and time	Number, date and time
Antonymy	Antonymy
Structure	Structure
Factivity	Factivity
Modality	Modality
Relations	Relations
Alignment	
AdjectiveGradation, Hypernymy	
Adjunct	

more precisely defined

Contradiction & Entailment

- Both systems are run independently
- Trust entailment system more



Contradiction results

		precision	recall
submission:	**alone**	26.3	10.0
	combined	28.6	8.0
post hoc:	**with filter**	27.54	12.67
	without filter	30.14	14.67

- Low recall
  - 47 contradictions filtered out by the “event” filter
  - 3 contradictions tagged as entailment
  - contradictions requiring deep lexical knowledge

Deep lexical knowledge

T: … Power shortages are a thing of the past.

H: Nigeria power shortage is to persist.

T: … No children were among the victims.

H: A French train crash killed children.

T: … The report of a crash was a false alarm.

H: A plane crashes in Italy.

T: … The current food crisis was ignored.

H: UN summit targets global food crisis.

Precision errors

- Hard to find contradiction features that reach high accuracy

	% error
Bad alignment	23
Coreference	6
Structure	40
Antonymy	10
Negation	10
Relations	6
Numeric	3

More knowledge is necessary

T: The company affected by this ban, Flour Mills of Fiji, exports nearly US$900,000 worth of biscuits to Vanuatu yearly.

H: Vanuatu imports biscuits from Fiji.

T: The Concord crashed […], killing all 109 people on board and four workers on the ground.

H: The crash killed 113 people.

Conclusion

- Linguistic analysis:

  some gain when improving dependency graphs

- Alignment:

  potential in phrase-based representation not yet proven: need better phrase-based lexical resources

- Inference:

  can detect some contradictions, but need to improve precision & add knowledge for higher recall