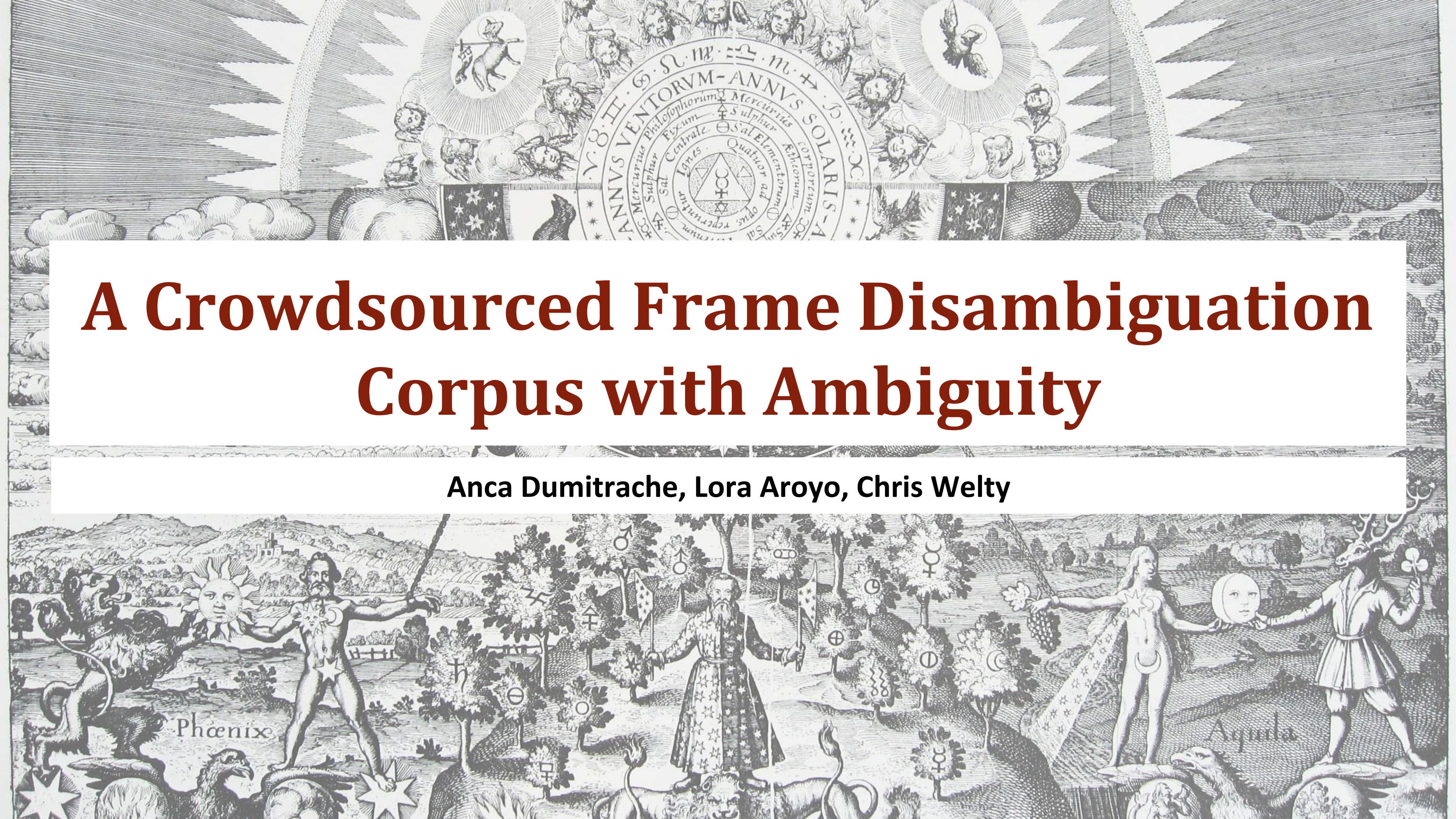
A Crowdsourced Frame Disambiguation Corpus with Ambiguity

Anca Dumitrache, Lora Aroyo, Chris Welty



TYPICAL EXPERT ANNOTATION TASK

Does the sentence express **TREATS**?

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades. ✓

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported. ✓

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood. ✗

BUT WHEN YOU ENCOURAGE DISAGREEMENT

Does the sentence express **TREATS**?

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

… AND ASK THE CROWD …

Does the sentence express **TREATS**?

Rheumatoid arthritis and **MALARIA** have been treated with **CHLOROQUINE** for decades.

95%

There's a difference between these two

BETTER

For prevention of malaria, use only in individuals traveling to malarious areas where **CHLOROQUINE** resistant P. falciparum **MALARIA** has not been reported.

75%

WORSE

Among 56 subjects reporting to a clinic with symptoms of **MALARIA** 53 (95%) had ordinarily effective levels of **CHLOROQUINE** in blood.

50%

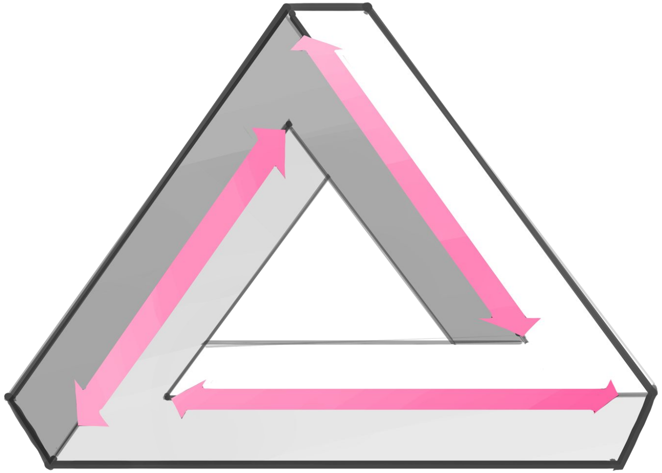
This one isn't utterly wrong

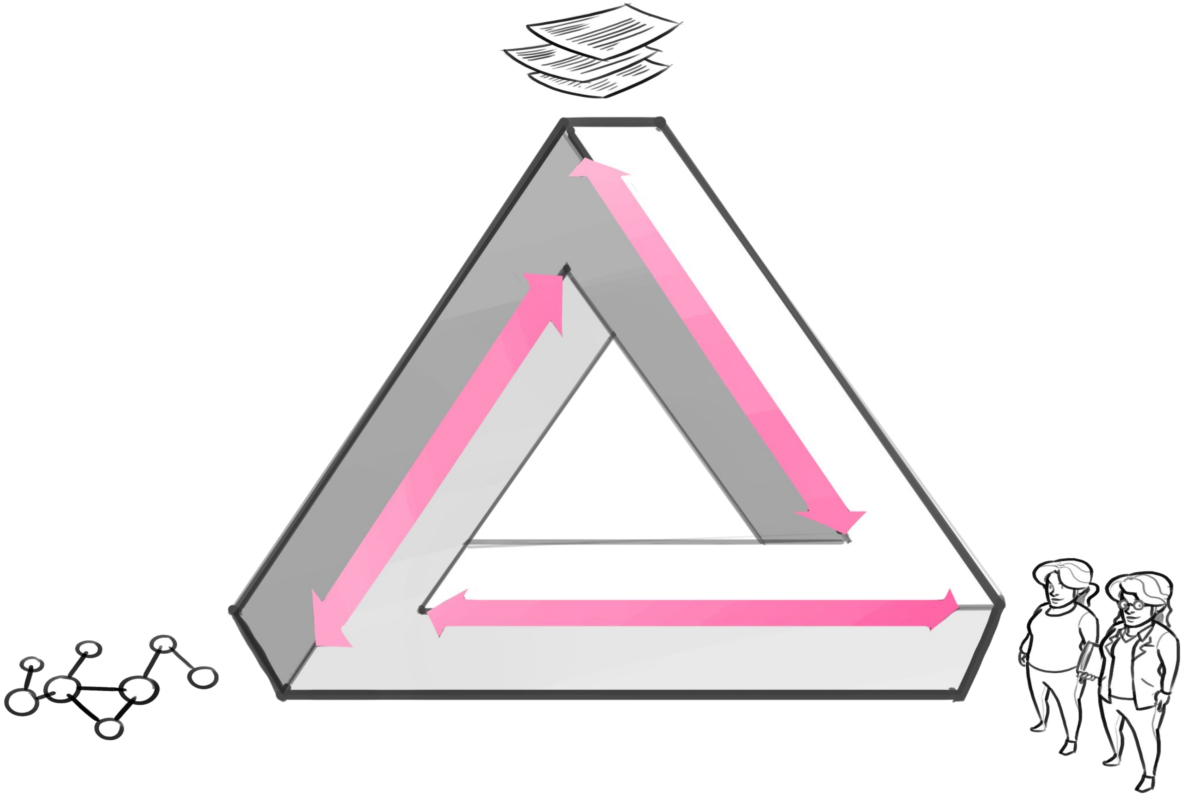
What causes disagreement?

- Workers
  - spam, lazy, unskilled
- Sentences
  - missing context
  - tokenization, span detection, etc.
  - doesn’t quite fit the task
  - poorly written, vague, ambiguous
- Target Semantics
  - unclear, confusing relations or types
  - granularity issues
  - limits of inference

What causes disagreement?

- **Workers**
  - spam, lazy, unskilled
- **Sentences**
  - missing context
  - tokenization, span detection, etc.
  - doesn’t quite fit the task
  - poorly written, vague, ambiguous
- **Target Semantics**
  - unclear, confusing relations or types
  - granularity issues
  - limits of inference





CROWDTRUTH

"Three Sides of CrowdTruth", *Human Computation 2014*, L. Aroyo, C. Welty

CrowdTruth Methodology

Annotator disagreement is **signal, not noise**

It is indicative of the **variation in human semantic interpretation**

It can indicate **ambiguity**, **vagueness**, **similarity**, over-generality, as well as **quality**

CrowdTruth.org

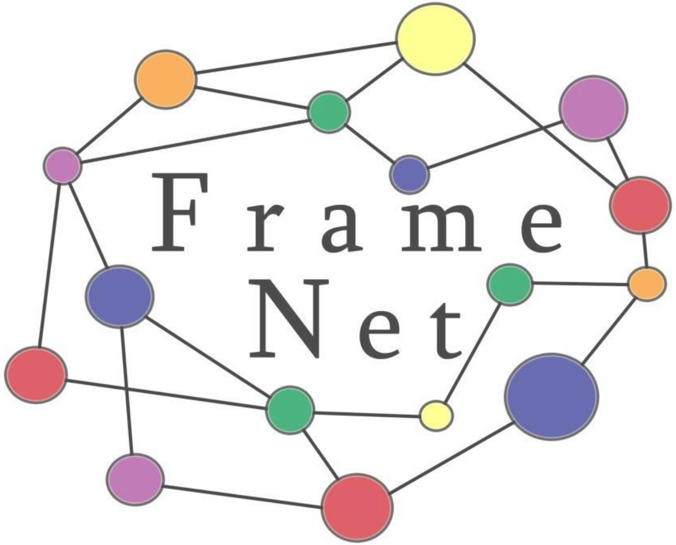
What is FrameNet?

FrameNet: computational linguistics resource based on the frame semantics theory (Baker, Fillmore, Lowe, 1998)

- collection of **semantic frames**
- **documents** annotated with these frames

semantic frame: abstract representation of a word sense, describing a type of *entity*, *relation*, or *event* grounded in *roles* implied by the frame

e.g. *from* & *to* are roles in a *movement* frame



Frame Disambiguation

= task of selecting the best frame for a word phrase

Illegal ***skimming*** of profits is rampant.

A. removing
B. theft
C. commiting crime
D. cause change

Frame Disambiguation

= task of selecting the best frame for a word phrase

Illegal ***skimming*** of profits is rampant.

A. removing (*)
B. theft
C. commiting crime
D. cause change

The frame picked by the expert is marked with (*).

Frame Net

What does the crowd think?

Frame Disambiguation

= task of selecting the best frame for a word phrase

Illegal ***skimming*** of profits is rampant.

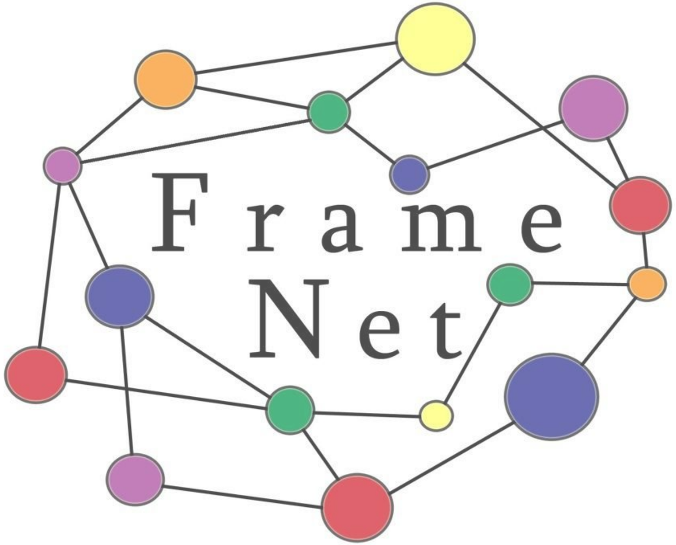
A. removing (*) → 7 votes

B. theft → 6 votes

C. commiting crime → 6 votes

D. cause change → 4 votes

The frame picked by the expert is marked with (*).



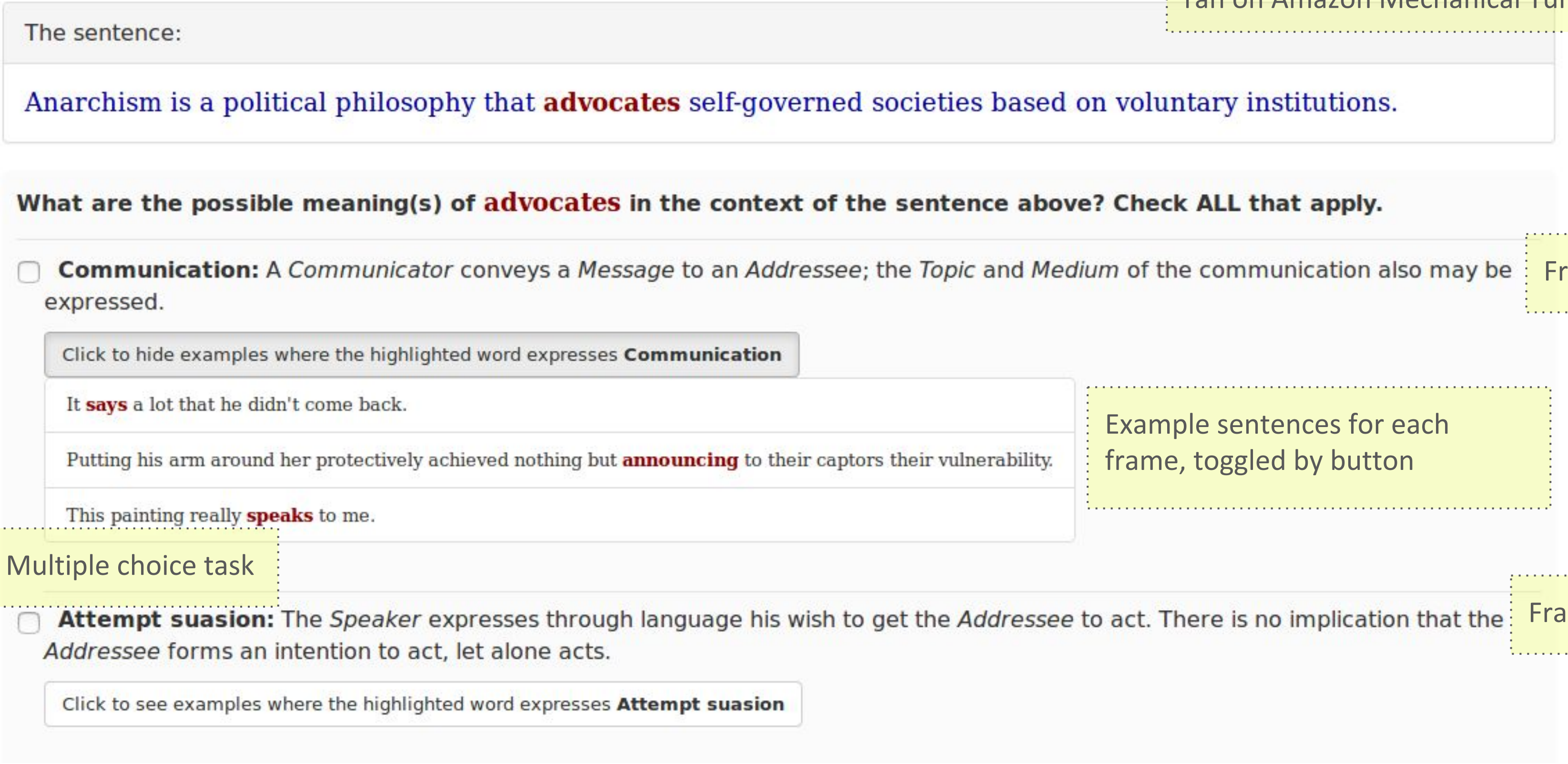
Dataset

- **9000 sentence-word pairs** from Wikipedia
  - <= 25 candidate frames per word
  - POS: verb, noun
  - in 1000 pairs from this set, the word (i.e. Lexical Unit) is not in FrameNet

- Pre-processing to find **candidate frames for each word**:
  - match word to *synonym sets* in WordNet corpus (Miller, 1995)
  - match synonym set to FrameNet frame using *Framester* corpus (Gangemi et al., 2016)

Crowdsourcing task

15 workers / sentence
$0.06 per judgment
ran on Amazon Mechanical Turk



The sentence:

Anarchism is a political philosophy that **advocates** self-governed societies based on voluntary institutions.

What are the possible meaning(s) of advocates in the context of the sentence above? Check ALL that apply.

□ **Communication:** A *Communicator* conveys a *Message* to an *Addressee*; the *Topic* and *Medium* of the communication also may be expressed.

Click to hide examples where the highlighted word expresses **Communication**

It **says** a lot that he didn't come back.

Putting his arm around her protectively achieved nothing but **announcing** to their captors their vulnerability.

This painting really **speaks** to me.

□ **Attempt suasion:** The *Speaker* expresses through language his wish to get the *Addressee* to act. There is no implication that the *Addressee* forms an intention to act, let alone acts.

Click to see examples where the highlighted word expresses **Attempt suasion**

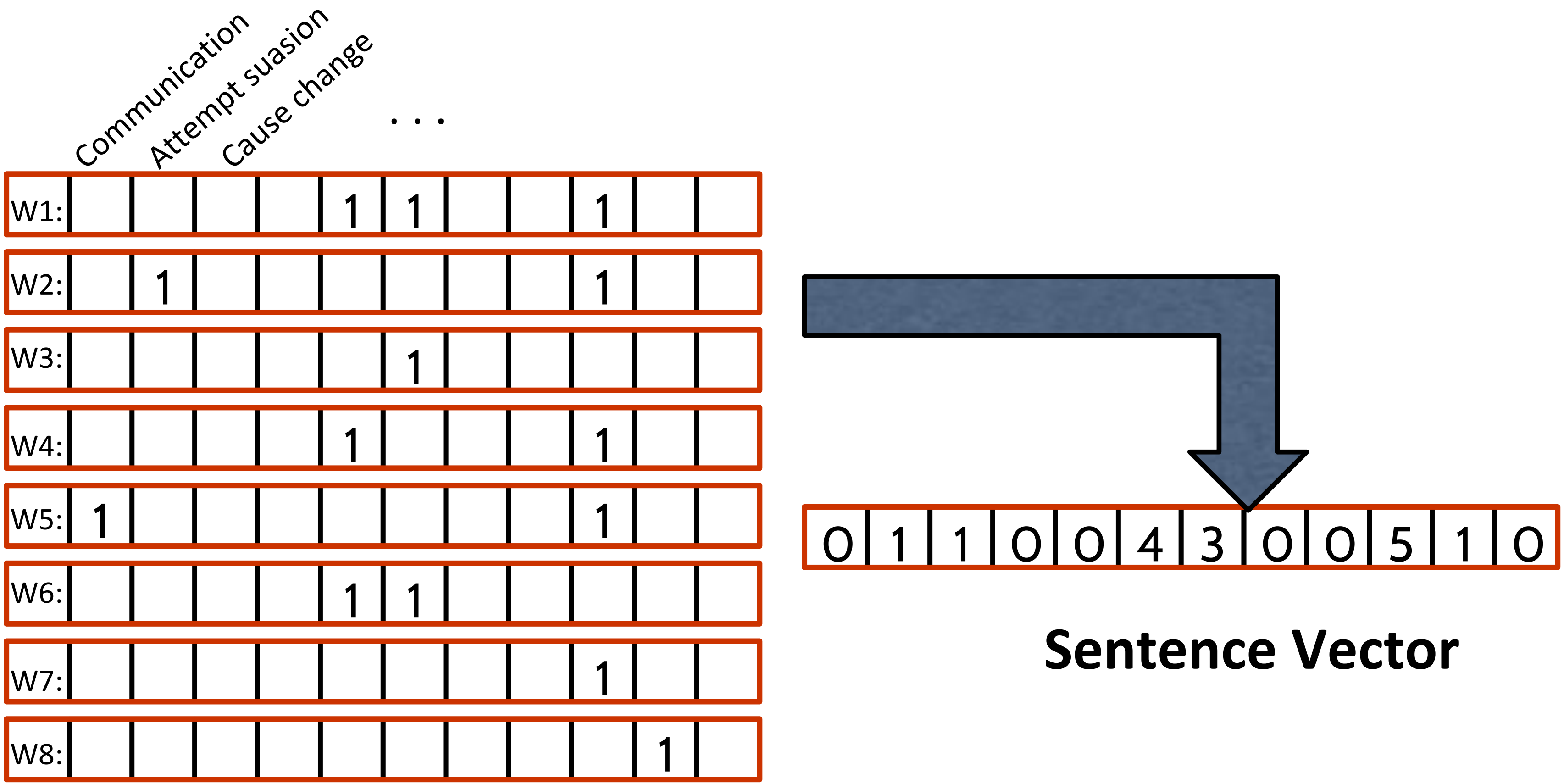
Frame definition

Example sentences for each frame, toggled by button

Multiple choice task

Frame definition

Worker Vectors



	Communication	Attempt suasion	Cause change			. . .					
W1:					1	1			1		
W2:		1							1		
W3:						1					
W4:					1				1		
W5:	1								1		
W6:					1	1					
W7:									1		
W8:										1	

0	1	1	0	0	4	3	0	0	5	1	0

Sentence Vector

CrowdTruth metrics

Frame-Sentence Score (FSS): the degree with which a particular frame matches the sense of the word in the sentence

$$FSS(s,f) = \frac{\sum \text{workers that picked } f \text{ in } s \cdot \text{worker quality}}{\sum \text{workers for } s \cdot \text{worker quality}}$$

Sentence Quality Score (SQS): overall worker agreement over one sentence, measured with cosine similarity

$$SQS(s) = \frac{\sum_{w1,w2\in \text{ workers for } s} \text{weighted cos sim}(\vec{w1}, \vec{w2}, FQS) \cdot \ \text{worker quality}(w1,w1)}{\sum_{w1,w2\in \text{workers for } s} \text{worker quality}(w1,w2)}$$

Frame Quality Score (FQS): agreement over a frame in all sentences where the frame was picked at least once

$$FQS(f) = \frac{\sum_{FSS(s,f)>0} FSS(s,f)SQS(s)}{\sum_{FSS(s,f)>0} SQS(s)}$$

Frame-Sentence Score (FSS):
how clearly the frame is expressed in the sentence

Example sentences with ***removing*** frame:

Egypt has provided no evidence demonstrating the *elimination* of its biological weapons.

***removing* - FSS = 0.938**

cause change - FSS = 0.175

Frame-Sentence Score (FSS): how clearly the frame is expressed in the sentence

Example sentences with ***removing*** frame:

Egypt has provided no evidence demonstrating the *elimination* of its biological weapons.

***removing* - FSS = 0.938**
cause change - FSS = 0.175

The Syrian Mujahiddin asked Hussein to *overthrow* the regime of Hafiz Al Assad.

change of leadership - FSS = 0.847
***removing* - FSS = 0.539**

Frame-Sentence Score (FSS): how clearly the frame is expressed in the sentence

Example sentences with ***removing*** frame:

Egypt has provided no evidence demonstrating the *elimination* of its biological weapons.

***removing* - FSS = 0.938**
cause change - FSS = 0.175

The Syrian Mujahiddin asked Hussein to *overthrow* the regime of Hafiz Al Assad.

change of leadership - FSS = 0.847
***removing* - FSS = 0.539**

Illegal *skimming* of profits is rampant.

***removing* - FSS = 0.532**
theft - FSS = 0.494
commiting crime - FSS = 0.459
misdeed - FSS = 0.431
cause change - FSS = 0.273

Sentence Quality Score (SQS): how ambiguous the sentence is

Example sentences with ***removing*** frame:

Egypt has provided no evidence demonstrating the *elimination* of its biological weapons.

removing - FSS = 0.938
cause change - FSS = 0.175

SQS = 0.841

The Syrian Mujahiddin asked Hussein to *overthrow* the regime of Hafiz Al Assad.

change of leadership - FSS = 0.847
removing - FSS = 0.539

SQS = 0.669

Illegal *skimming* of profits is rampant.

removing - FSS = 0.532
theft - FSS = 0.494
commiting crime - FSS = 0.459
misdeed - FSS = 0.431
cause change - FSS = 0.273

SQS = 0.366

Frame Quality Score (FQS): how ambiguous the frame is

Concrete frames have high FQS.

e.g. *removing*

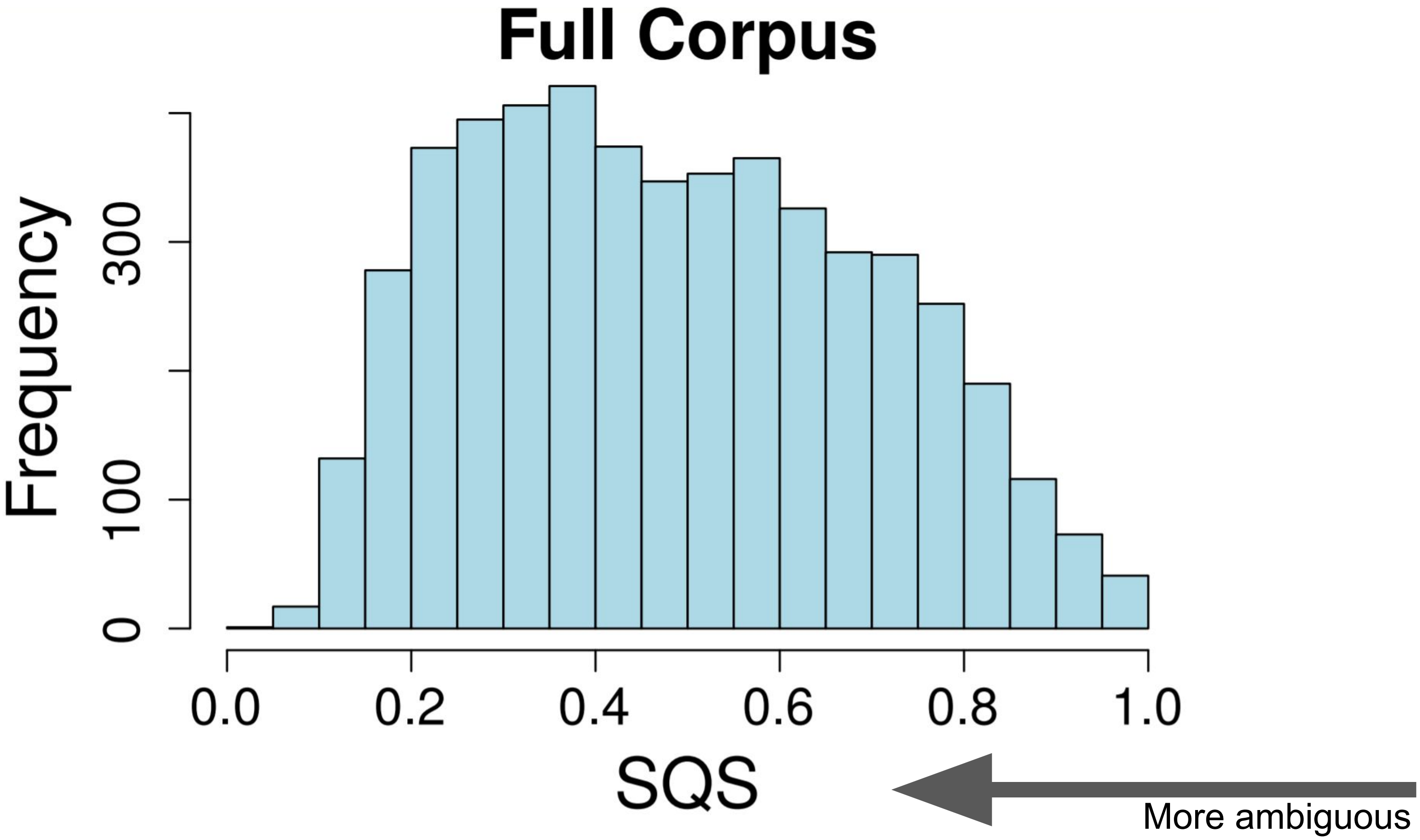
Abstract frames have low FQS.

e.g. *cause change*

Frames with **overlapping definitions** have low FQS.

e.g. *objective influence* & *subjective influence*

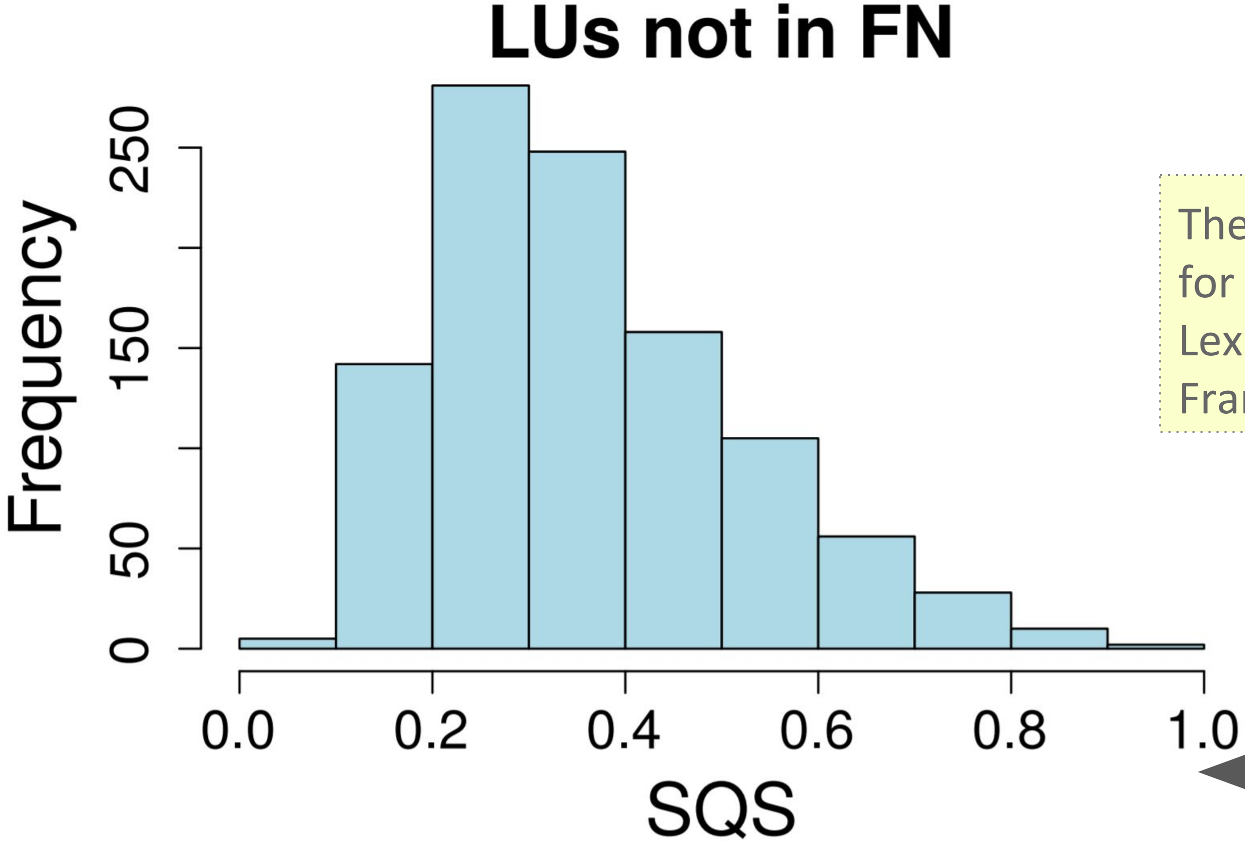
Ambiguity in the corpus



Full Corpus

More ambiguous

Ambiguity in the corpus



There is more ambiguity for sentences where the Lexical Unit is not part of FrameNet.

More ambiguous

Why does ambiguity happen?

These Articles *continue* to direct the ethos of the Communion.

activity ongoing - FSS = 0.862
process continue - FSS = 0.86

SQS = 0.795

parent-child relation between frames

Some aikido organizations use belts to *distinguish* practitioners' grades

differentiation - FSS = 0.867
distinctiveness - FSS = 0.703

SQS = 0.68

overlapping frame definitions

Cornwallis prematurely abandoned his outer position, *hastening* his subsequent defeat.

speed description - FSS = 0.39
assistance - FSS = 0.209
self motion - FSS = 0.165
travel - FSS = 0.16
causation - FSS = 0.124

SQS = 0.134

meaning of the word is a composition of frames

Evaluation with CrowdTruth data

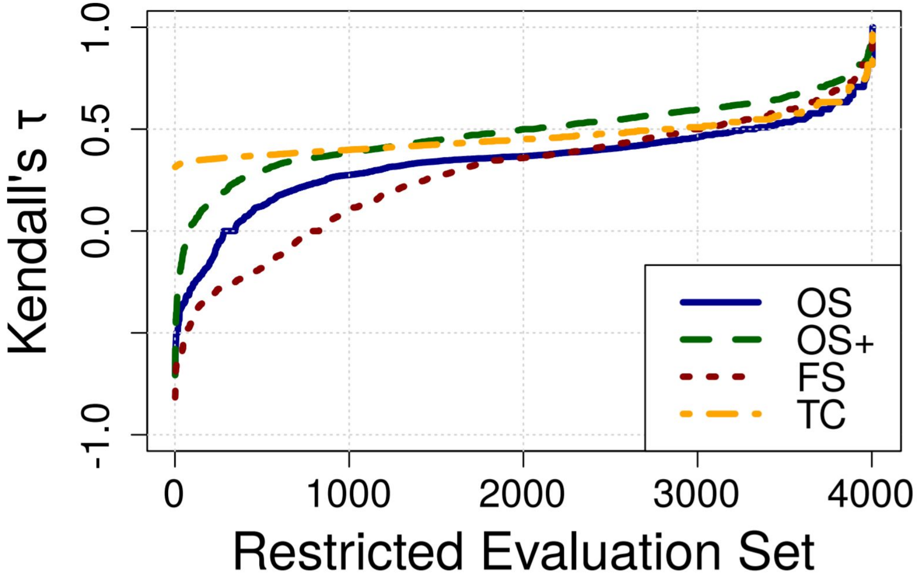
Models:

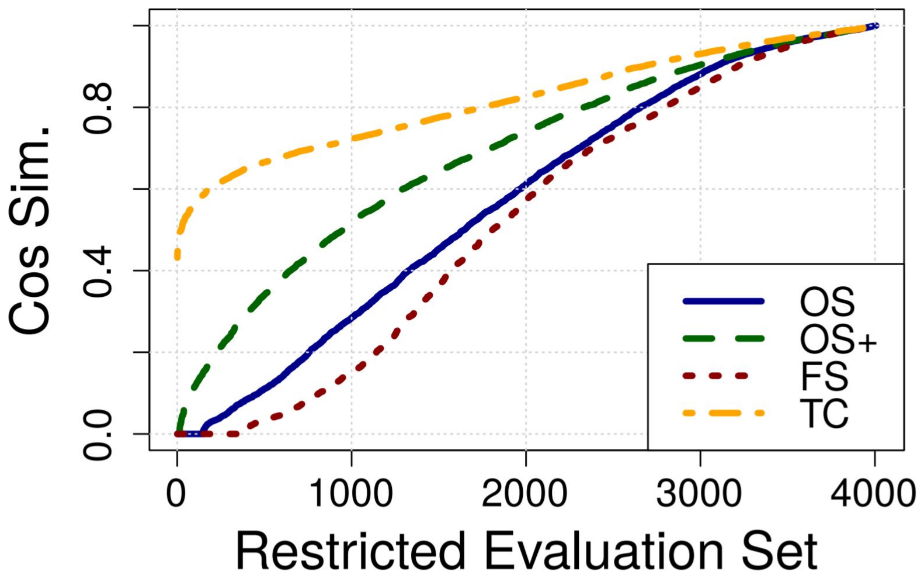
- **OS:** OpenSesame frame disambiguation classifier (Swayamdipta et al., 2017), results in 1 frame per sentence, cannot classify Lexical Units not in FrameNet
- **OS+:** OpenSesame modified to perform multi-label classification, cannot classify Lexical Units not in FrameNet
- **Framester:** rule-based multi-class multi-label classification; works on an older version of FrameNet
- **TC:** top frame picked by the crowd

Evaluation metrics:

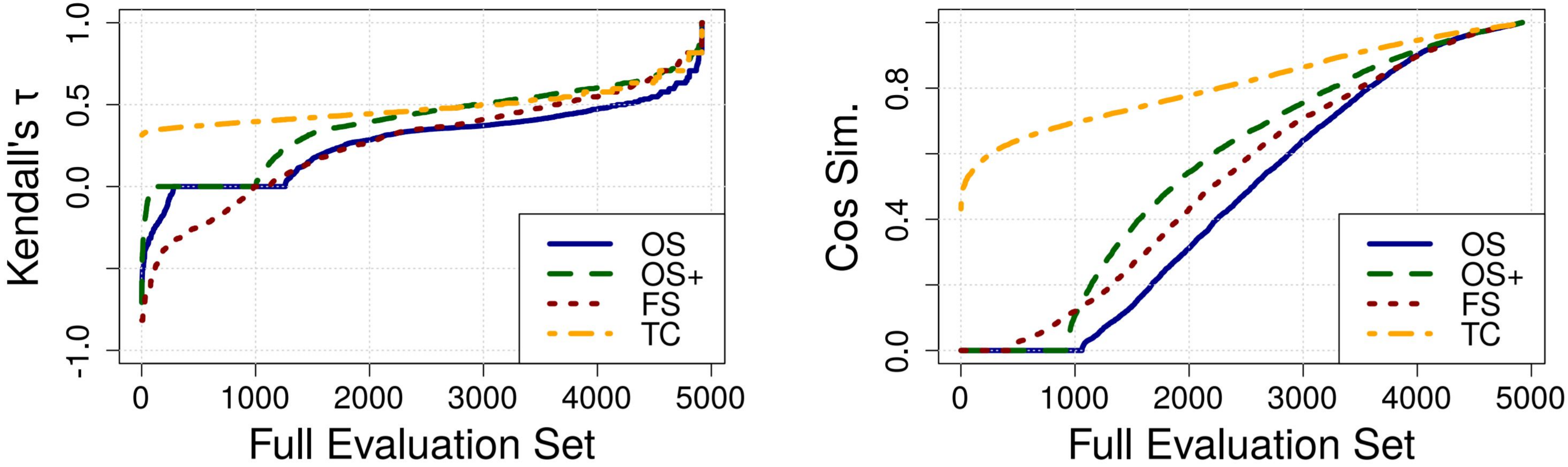
- **Kendall’s τ:** list ranking coefficient
- **cosine similarity:** distance between FSS-labeled crowd frames & frames predicted by the models

Restricted Set = sentences where all the Lexical Units are in FrameNet (i.e. less ambiguous)





OS+ does better than TC for Kendall's τ.
Correctly ranking multiple frames per sentence is more important than finding the single best frame.



OS+ performance drops, since it can't classify Lexical Units not in FrameNet.
FS performance is low because of missing frames in the older version of FrameNet it uses.

Conclusion

CrowdTruth

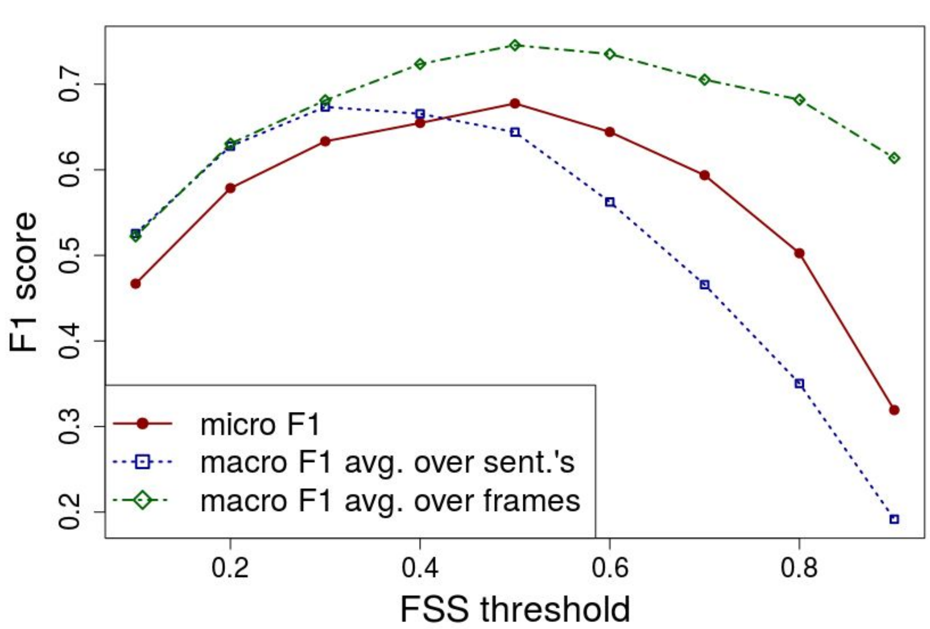
Results:

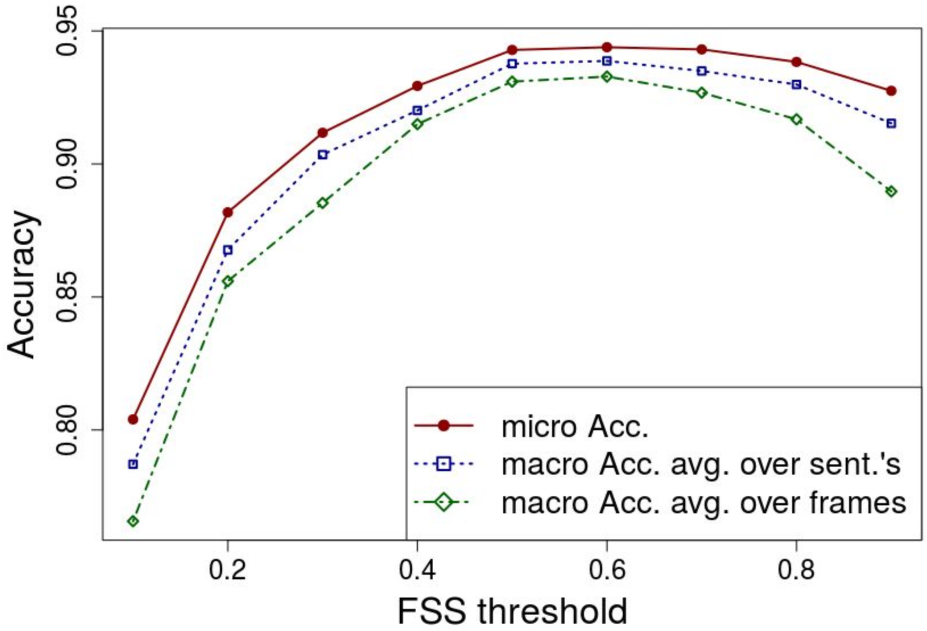
- 9000 sentences from FrameNet annotated with CrowdTruth
- There’s not *only one right answer* for each example, tolerate multiple outcomes
- Don’t assume lexical resources are perfect
- Disagreement is a good indicator of ambiguity in sentences & frames.

Resources:

- Dataset: https://github.com/CrowdTruth/FrameDisambiguation
- CrowdTruth metrics: https://github.com/CrowdTruth/CrowdTruth-core
- CrowdTruth metrics Python package: https://pypi.org/project/CrowdTruth/

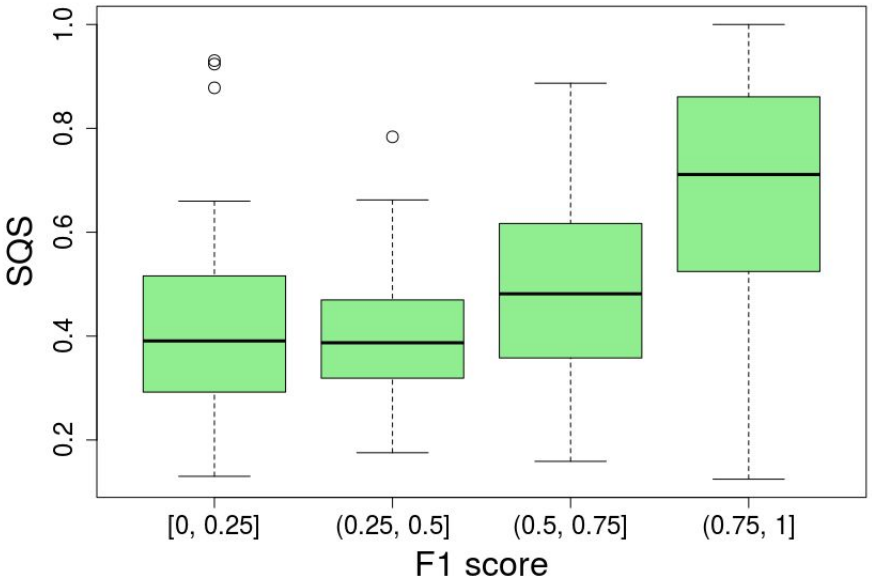
Crowd vs. FrameNet experts ground truth

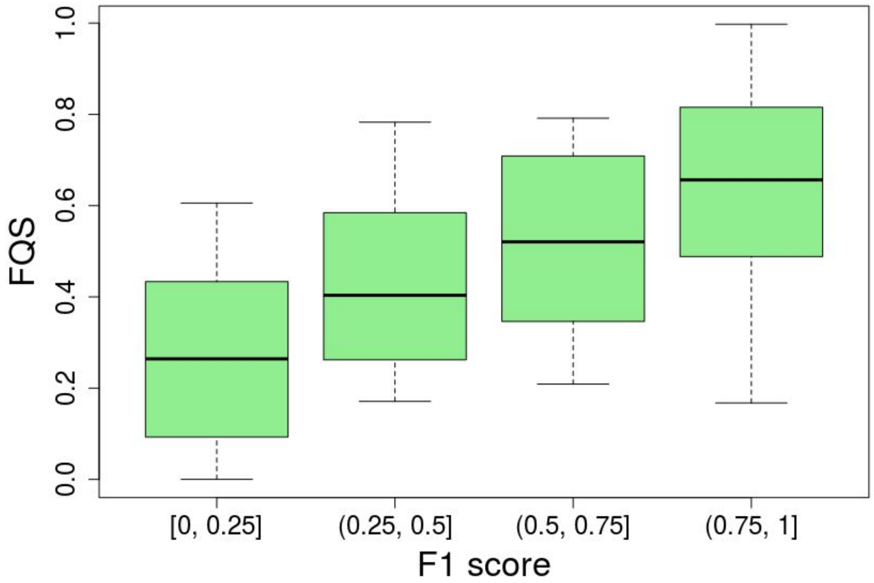




Crowd performance is comparable to the experts.

SQS and FQS vs. Expert ground truth





When the crowd workers agree with each other, they also agree with the expert.

But disagreement can have a good reason!

When crowd & expert disagree

Crowd **misunderstood** the frame definition.

Information in the sentence is **incomplete**.

The *investigation* has been stymied, stopped, obstructions thrown every step of the way.

Crowd: criminal investigation (FSS = 0.804)

Expert: scrutiny (FSS = 0.305)

Crowd is **correct**.

Does *supersizing* cause obesity?

Crowd: cause to start (FSS = 0.804)

Expert: causation (FSS = 0.608)

Crowd still picked the expert frame, but with lower FSS.