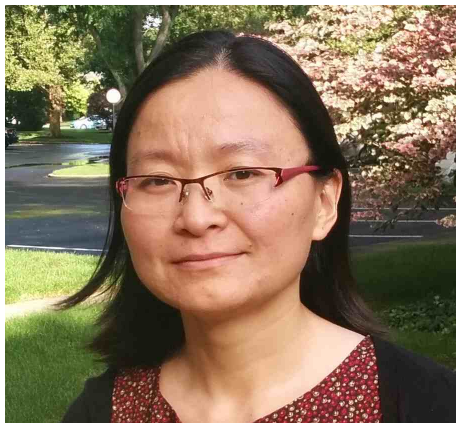
Generating Disambiguating Paraphrases for Structurally Ambiguous Sentences

Manjuan Duan, Ethan Hill, Michael White

August 11-12, 2016, LAW-X

The Ohio State University
Department of Linguistics

Joint work with



Manjuan Duan

Ethan Hill

Introduction

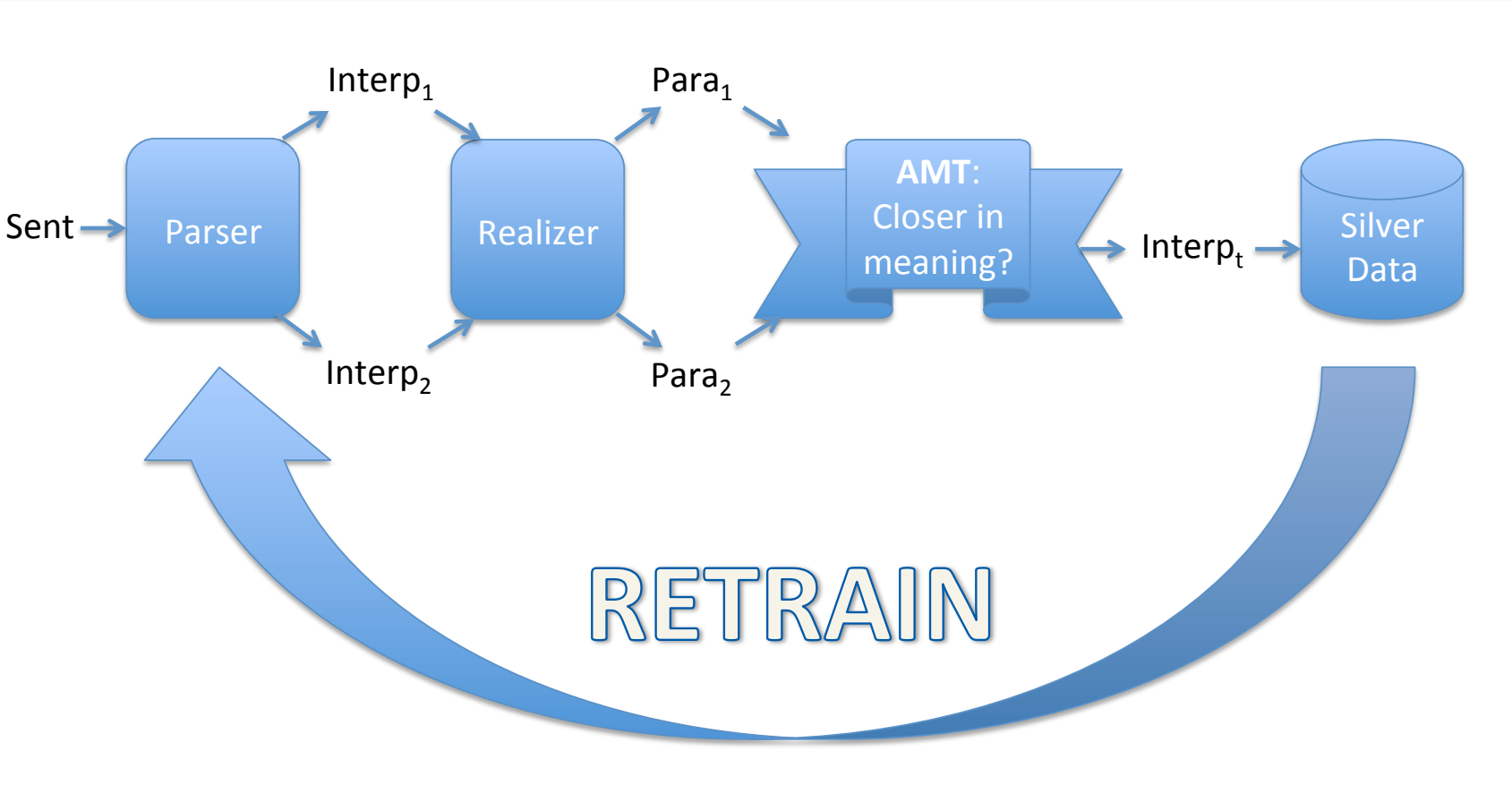
How can we crowd-source data for adapting parsers to new domains?

- To some extent, **MTurk workers** can perform meaning- and form-oriented tasks such as annotating PP-attachment points, with some training (Snow et al., 2008; Jha et al., 2010)
- Gerdes (2013) and Zeldes (2016) also found that it was possible to obtain fairly high quality **class-sourced** annotations, where students only received a modest amount of training

How can we crowd-source data for adapting parsers to new domains?

- To some extent, **MTurk workers** can perform meaning- and form-oriented tasks such as annotating PP-attachment points, with some training (Snow et al., 2008; Jha et al., 2010)
- Gerdes (2013) and Zeldes (2016) also found that it was possible to obtain fairly high quality **class-sourced** annotations, where students only received a modest amount of training
- In the current study, rather than annotating syntax, we use natural language clarification questions, simply asking Mturk workers to select the right paraphrase of a structurally ambiguous sentence

Big picture: Just ask people what ambiguous sentences mean

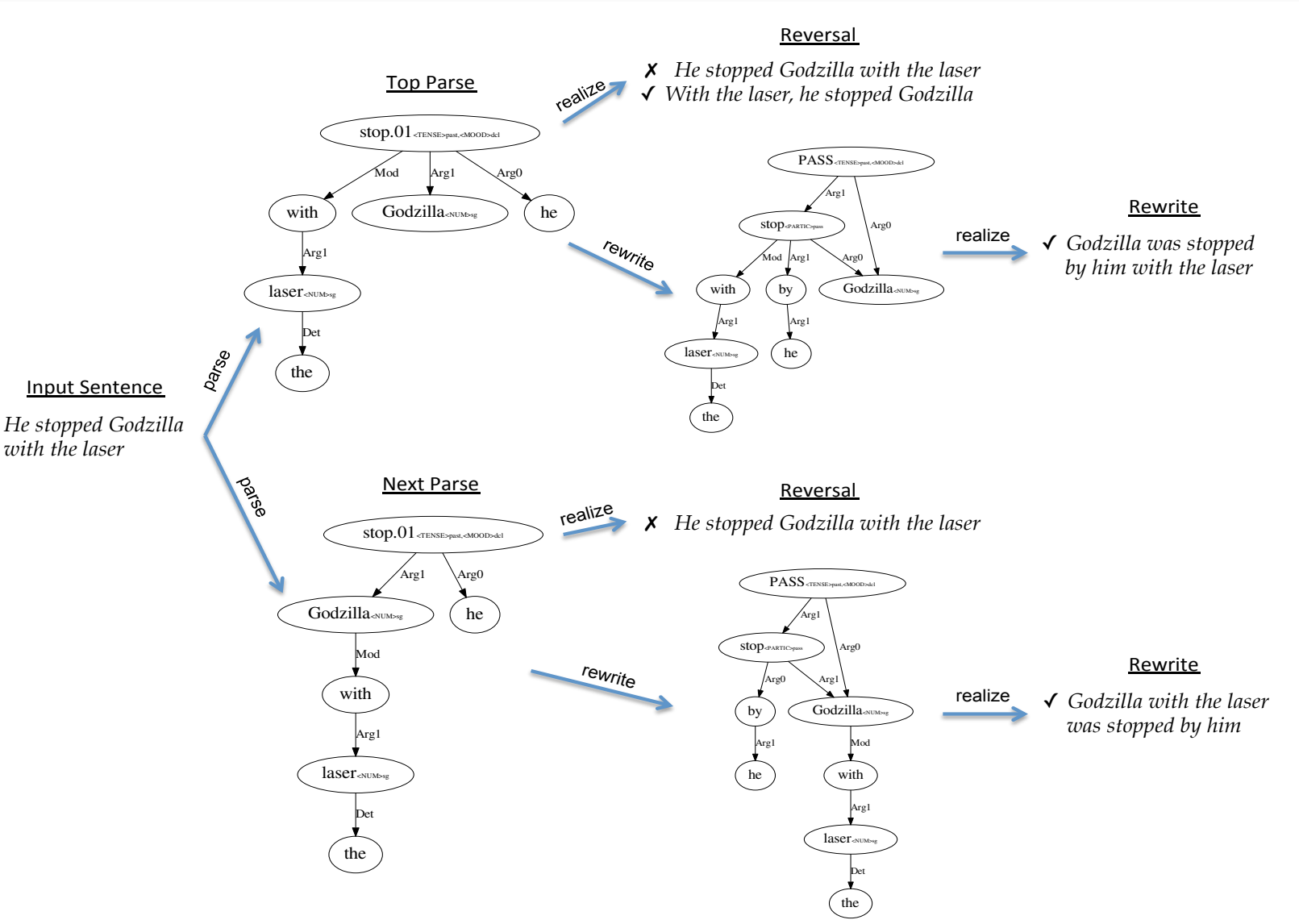


Difference from previous studies

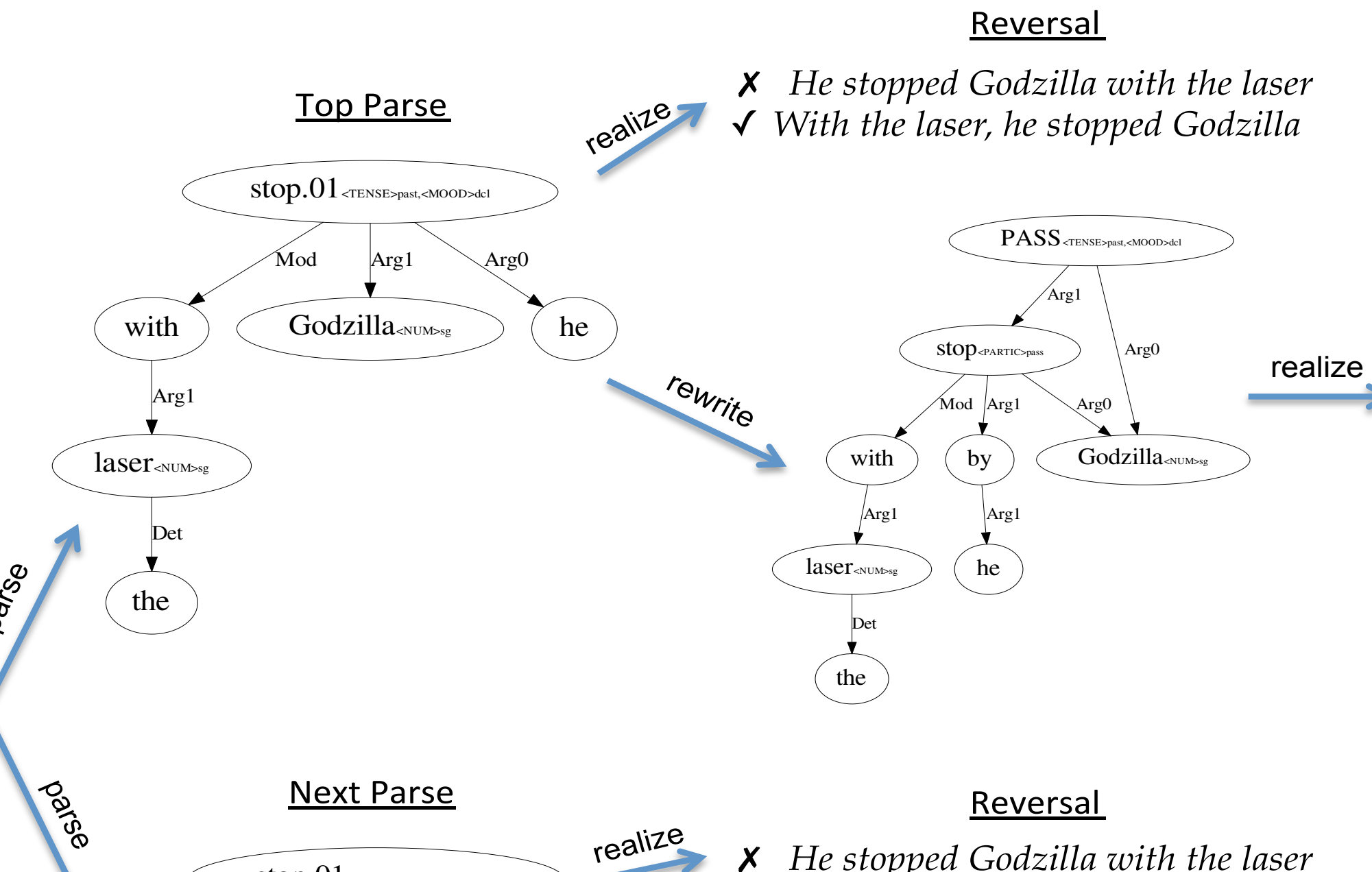
- Aiming (ultimately) for all structural ambiguities identifiable by an automatic parser, not confined to some specific constructions (Jha et al., 2010)
- AMT workers are making choices among paraphrases, not annotations, and no specific tutorial is needed

Methods

Generating disambiguating paraphrases: An illustration



Generating disambiguating paraphrases: An illustration

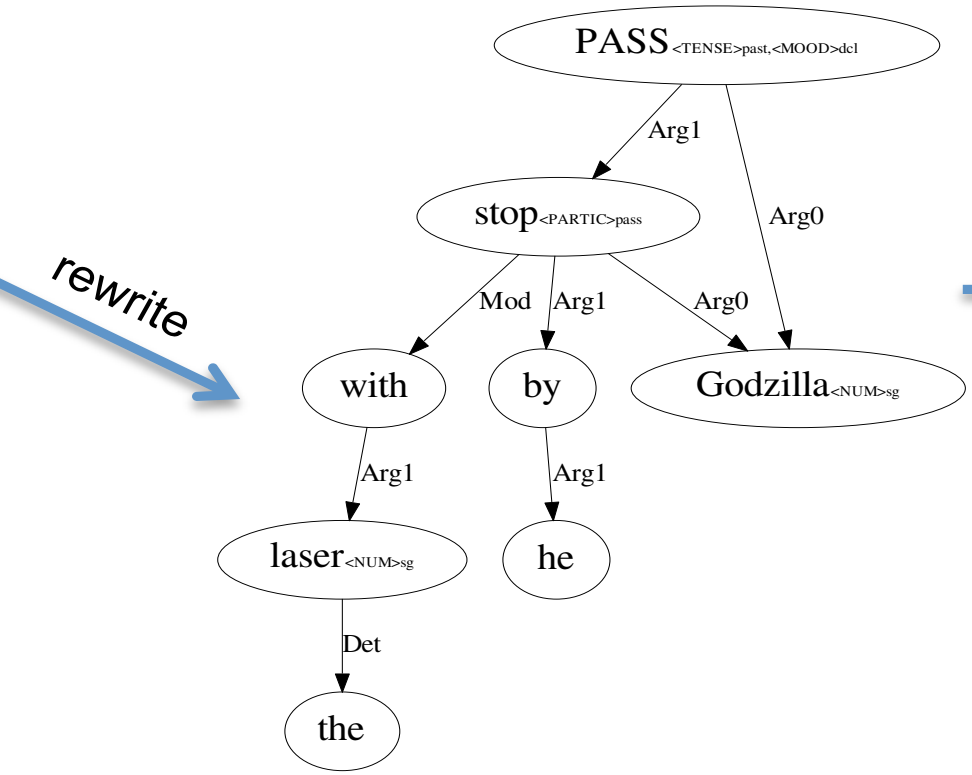


Generating disambiguating paraphrases: An illustration

Reversal

✗ *He stopped Godzilla with the laser*

✓ *With the laser, he stopped Godzilla*



rewrite

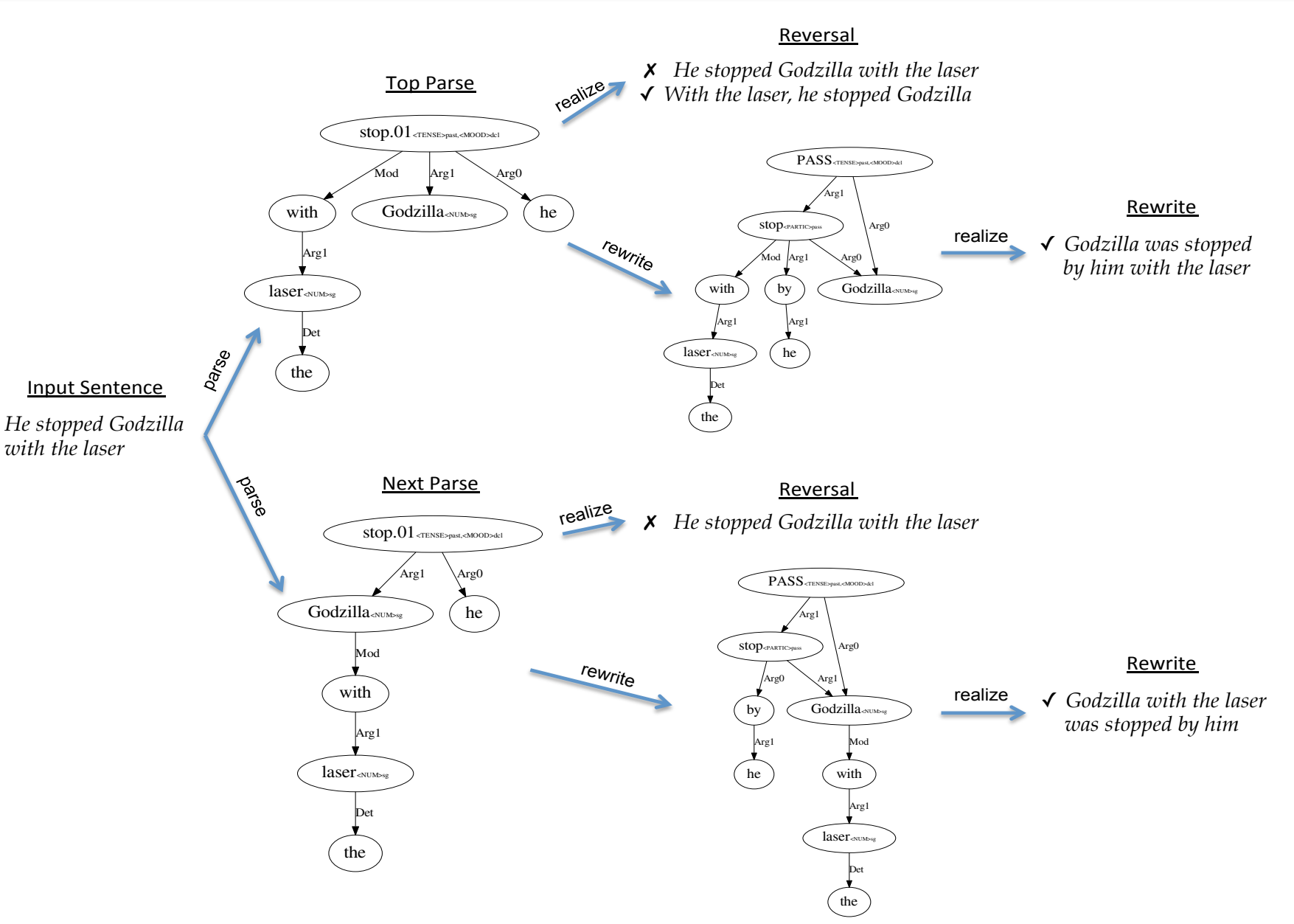
realize

Rewrite

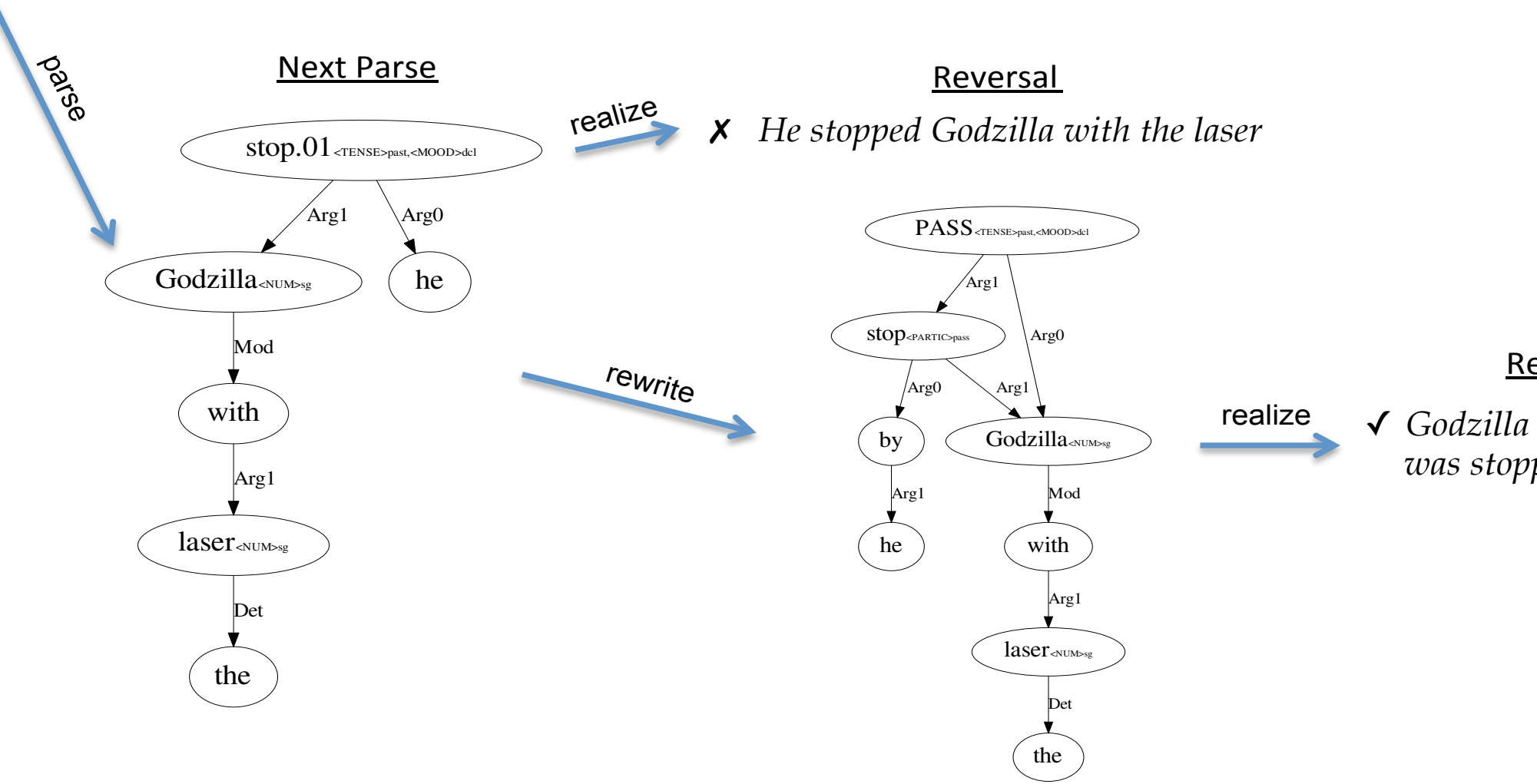
✓ *Godzilla was stopped by him with the laser*

Reversal

Generating disambiguating paraphrases: An illustration



Generating disambiguating paraphrases: An illustration



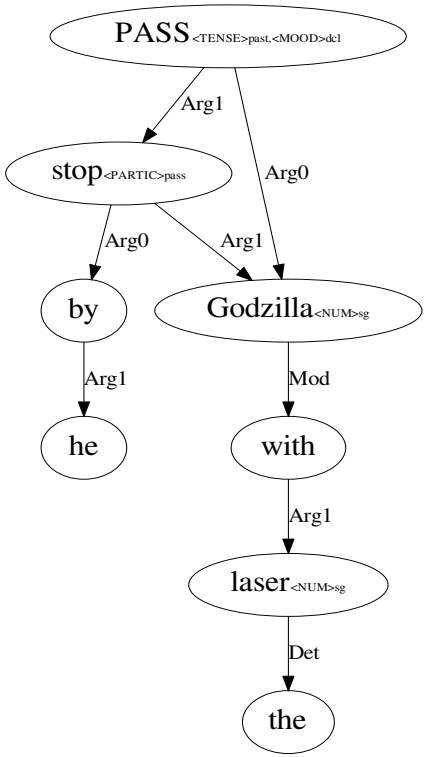
Generating disambiguating paraphrases: An illustration

lize

Reversal

✗ *He stopped Godzilla with the laser*

rewrite



realize

Rewrite

✓ *Godzilla with the laser was stopped by him*

Obtaining meaningfully distinct parses

1. Parse the input sentence with the OpenCCG parser to obtain its top 25 parses
2. Find a parse from the *n*-best parse list which is **meaningfully distinct** from the top parse:

Obtaining meaningfully distinct parses

1. Parse the input sentence with the OpenCCG parser to obtain its top 25 parses
2. Find a parse from the *n*-best parse list which is **meaningfully distinct** from the top parse:
    - Only compare the unlabeled and unordered dependencies from the two parses
    - The symmetric difference cannot be empty, with neither set of dependencies a superset of the other

Obtaining meaningfully distinct parses

1. Parse the input sentence with the OpenCCG parser to obtain its top 25 parses
2. Find a parse from the *n*-best parse list which is **meaningfully distinct** from the top parse:
    - Only compare the unlabeled and unordered dependencies from the two parses
    - The symmetric difference cannot be empty, with neither set of dependencies a superset of the other
    - Ambiguities involving only POS, named entity or word sense differences are disregarded

Obtaining meaningfully distinct parses

1. Parse the input sentence with the OpenCCG parser to obtain its top 25 parses
2. Find a parse from the *n*-best parse list which is **meaningfully distinct** from the top parse:
    - Only compare the unlabeled and unordered dependencies from the two parses
    - The symmetric difference cannot be empty, with neither set of dependencies a superset of the other
    - Ambiguities involving only POS, named entity or word sense differences are disregarded
3. If successful, this phase yields a *top* and *next* parse — the ones reflecting the **greatest uncertainty**

Two ways to obtain paraphrases

- **Paraphrases obtained from reverse realization** (*reversals*)
  - Able to generate paraphrases for ambiguities involving various constructions identifiable by an auto parser
- **Paraphrases obtained from logical form rewriting** (*rewrites*)
  - Triggered by specific syntactic constructions such as PP-attachment ambiguity and modifier scope ambiguity in coordination

Validating reverse realizations

Need to ensure paraphrases **actually disambiguate** intended meanings

Validating reverse realizations

Need to ensure paraphrases **actually disambiguate** intended meanings

1. **Realize** the *top* and *next* parse into a *n*-best realization list (*n*=25), using OpenCCG
2. Traverse the list to find a qualifying paraphrase, which has to
   - **be different** from the original sentence
   - **have different relative distance** among the words involving the ambiguity from the original sentence

Validating reverse realizations

Need to ensure paraphrases **actually disambiguate** intended meanings

1. **Realize** the *top* and *next* parse into a *n*-best realization list (*n*=25), using OpenCCG
2. Traverse the list to find a qualifying paraphrase, which has to
   - **be different** from the original sentence
   - **have different relative distance** among the words involving the ambiguity from the original sentence
3. **Parse** each candidate paraphrase to make sure the most likely interpretation includes the dependencies from which it was generated

Two-sided paraphrases and one-sided paraphrases

- *Two-sided paraphrases*: Two paraphrases are obtained for the original sentence, one generated from the *top* parse, and one from the *next*
- *One-sided paraphrases*: Only one paraphrase is obtained for the original sentence

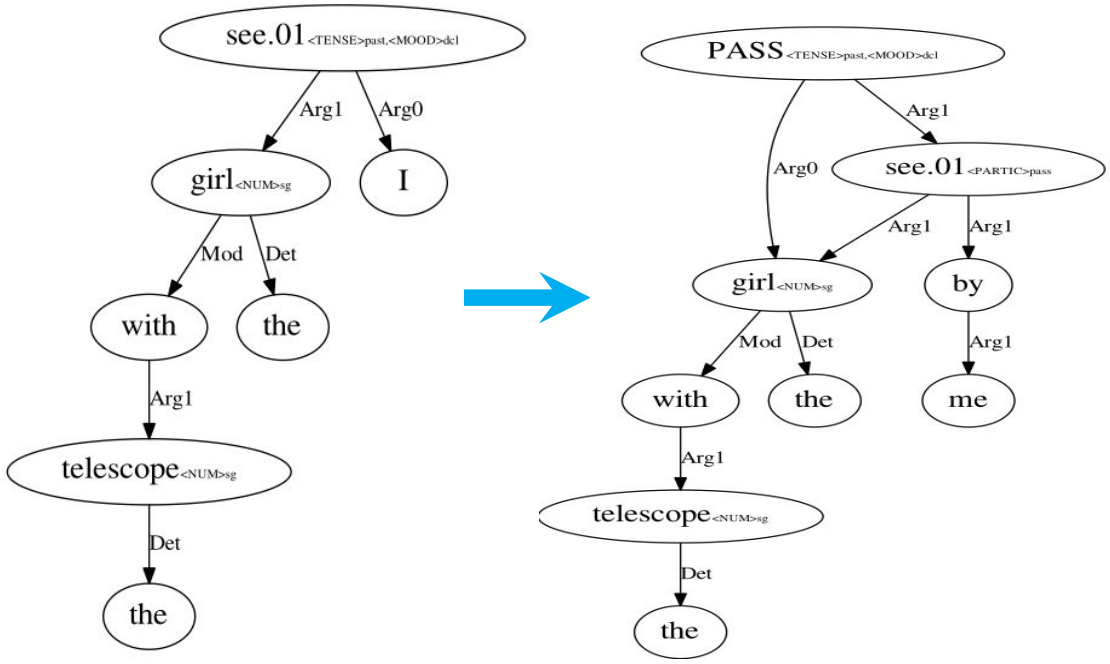
Logical form rewriting

Rewritten logical forms are realized to obtain paraphrases which highlight the ambiguous part

- Passive and cleft rewrites for PP-attachment ambiguities
- Coordination rewrites for ambiguities in the scope of modiers with coordinated phrases

Passive rewrites: An example

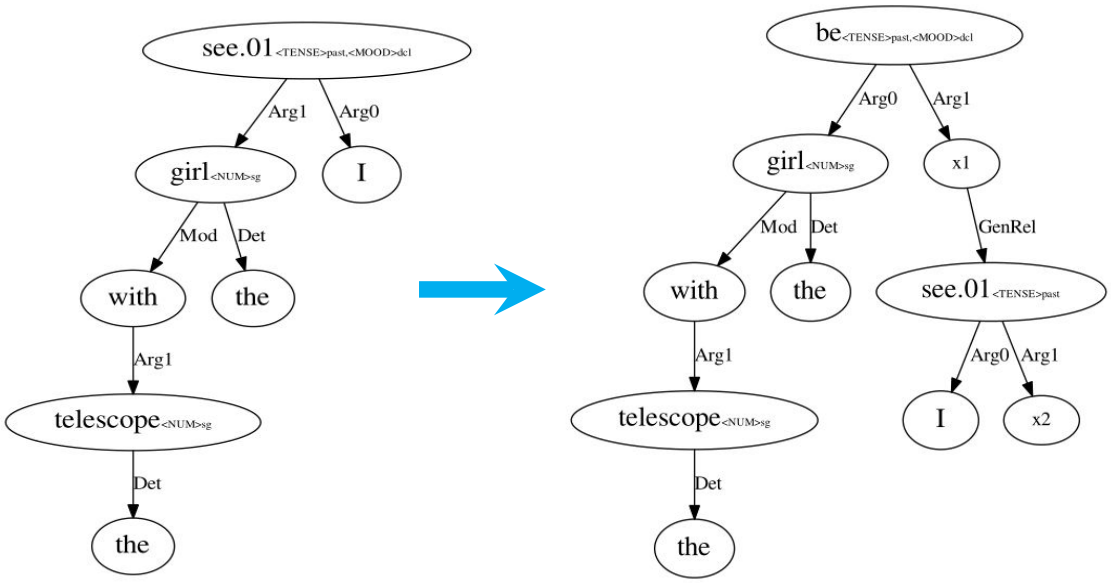
I saw the girl with the telescope.



⇒ *The girl with the telescope was seen by me.*

Cleft rewrites: An example

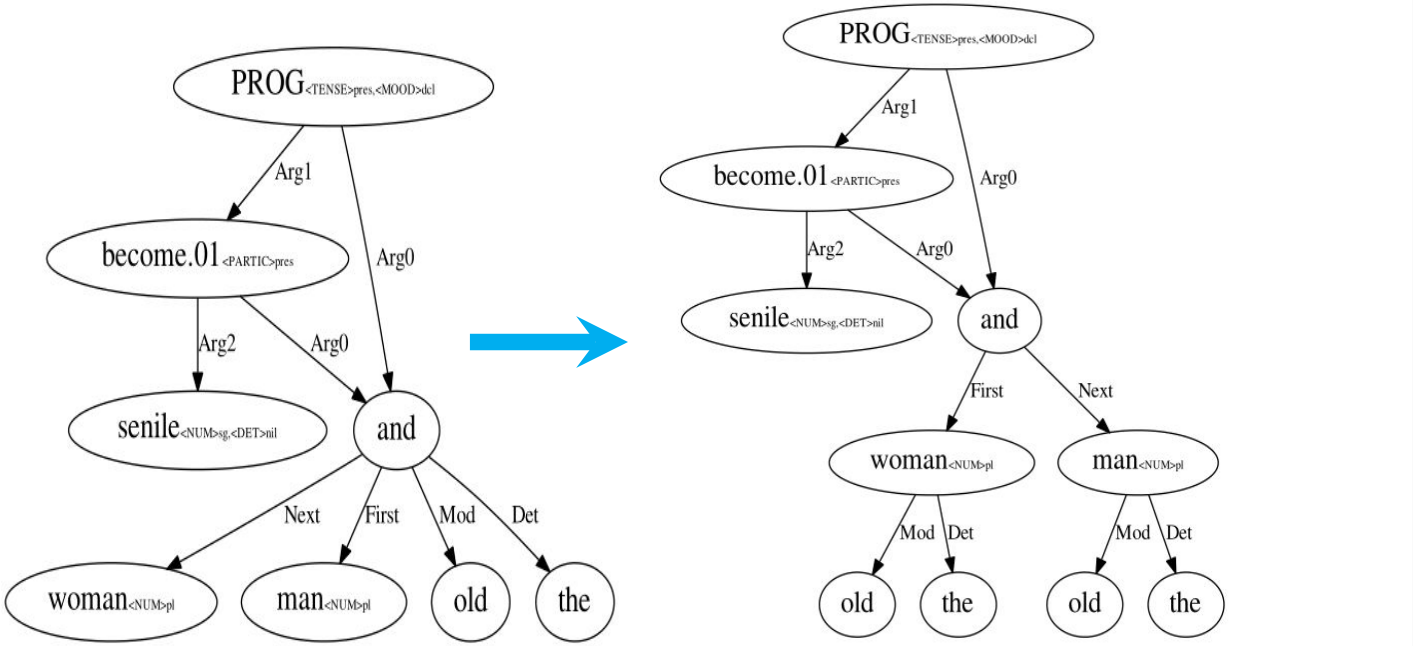
I saw the girl with the telescope.



⇒ *The girl with the telescope was what I saw.*

Coordination rewrites: An example (1)

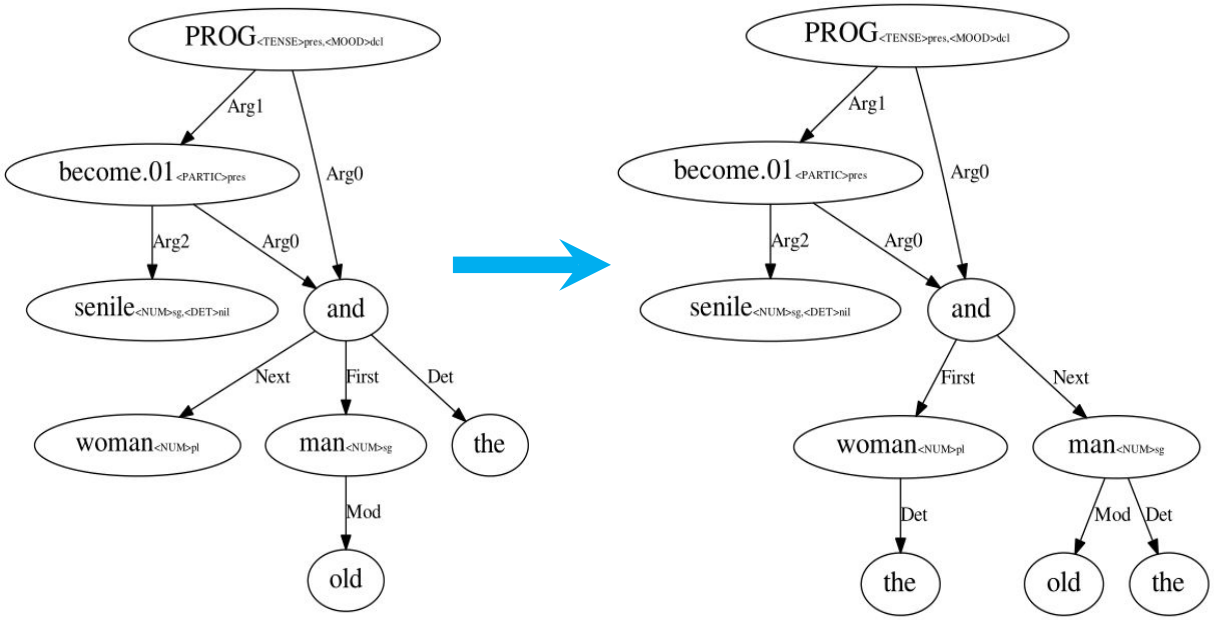
The old men and women are becoming senile.



$\Rightarrow$ *The old women and the old men are becoming senile*

Coordination rewrites: An example (2)

The old men and women are becoming senile.



$\Rightarrow$ *The women and the old men are becoming senile*

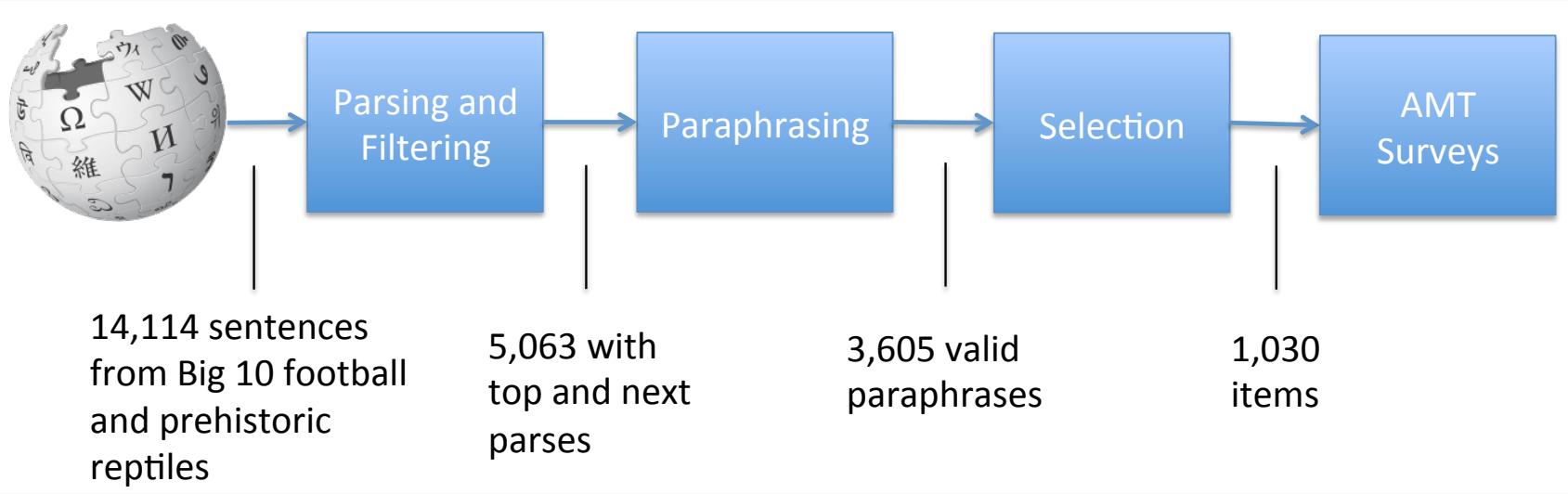
Experiment

Validation experiment

Aim: Examine the **quality of the crowd-sourced annotations** through disambiguating paraphrases

- Used AMT workers as our naive annotators
- For comparison, **hand annotated** 1,030 sentences as the optimal ('gold') annotations to measure the accuracy of the crowd-sourced annotations

Data preparation



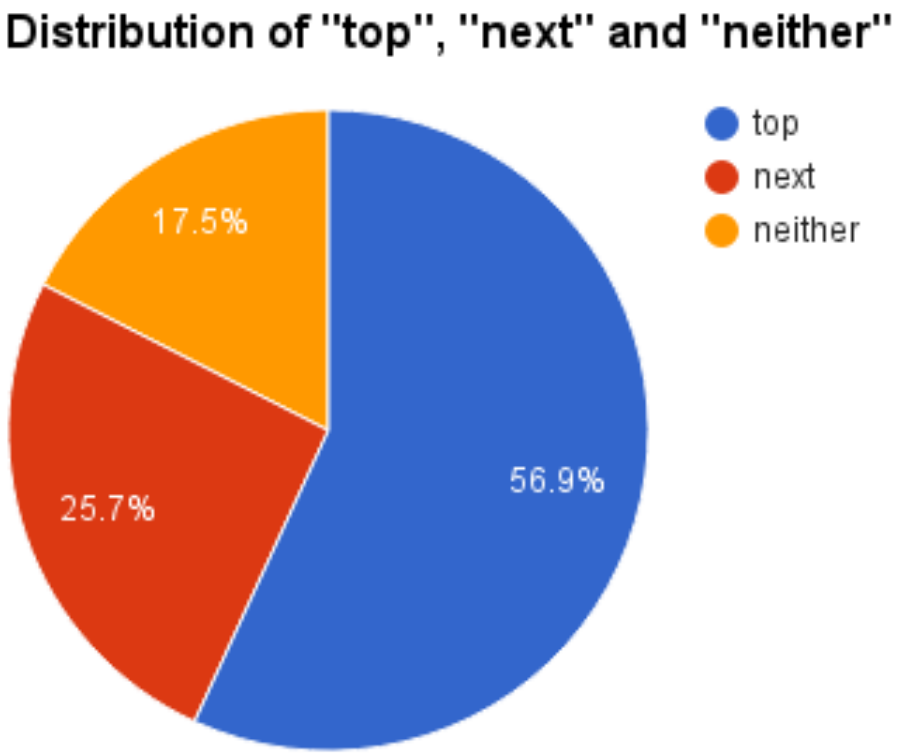
Working assumption: Unannotated data available in large quantities, so can focus on most informative ambiguities

Gold annotations

We selected the correct parse of the sentence by examining the dependency graphs of the input sentence:

- Annotated ‘top’ if the *top* parse was correct
- Annotated ‘next’ if the *next* parse was correct
- Annotated ‘neither’ if neither of them was more correct than the other one

Distribution of test data

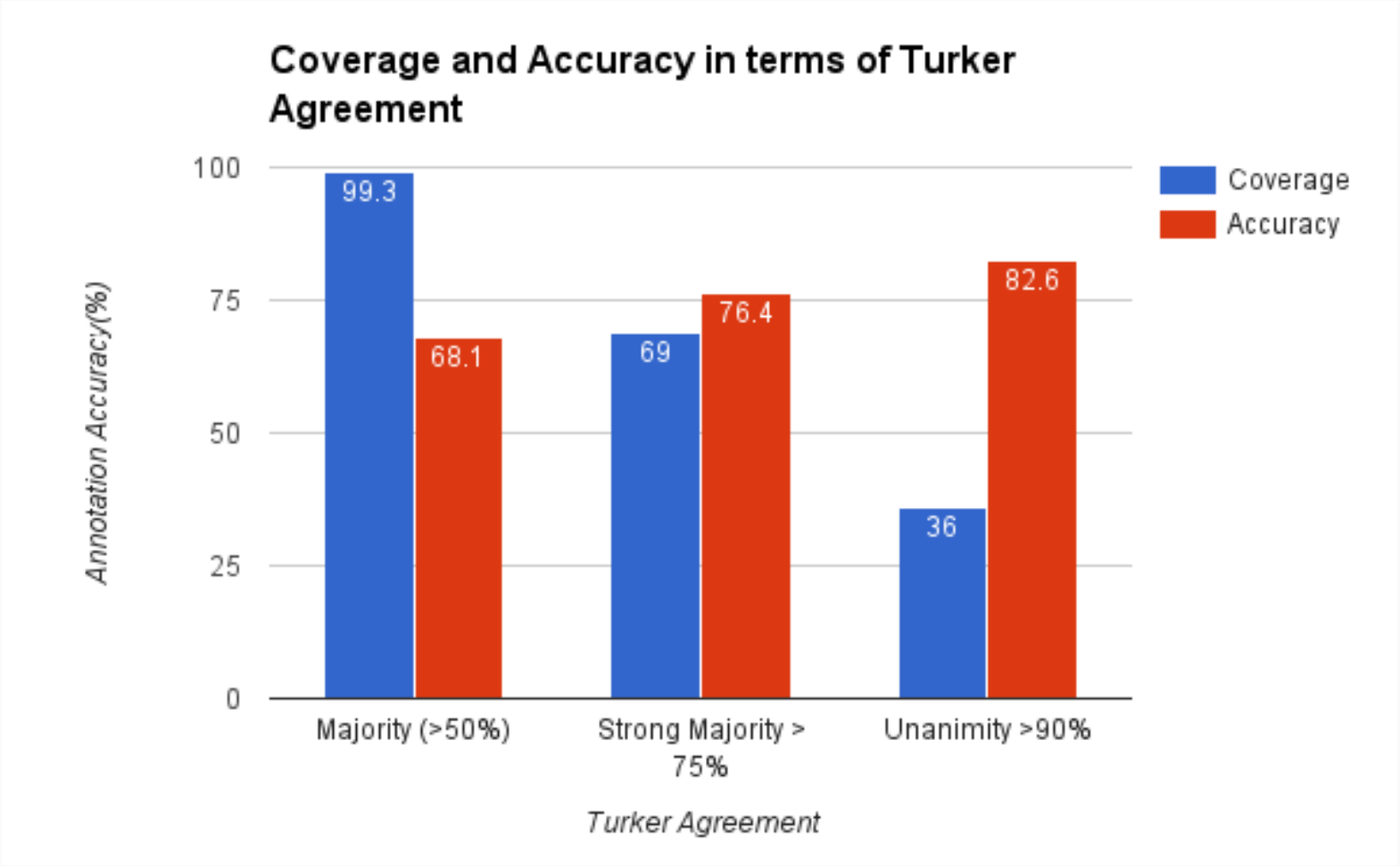


Collecting human judgments

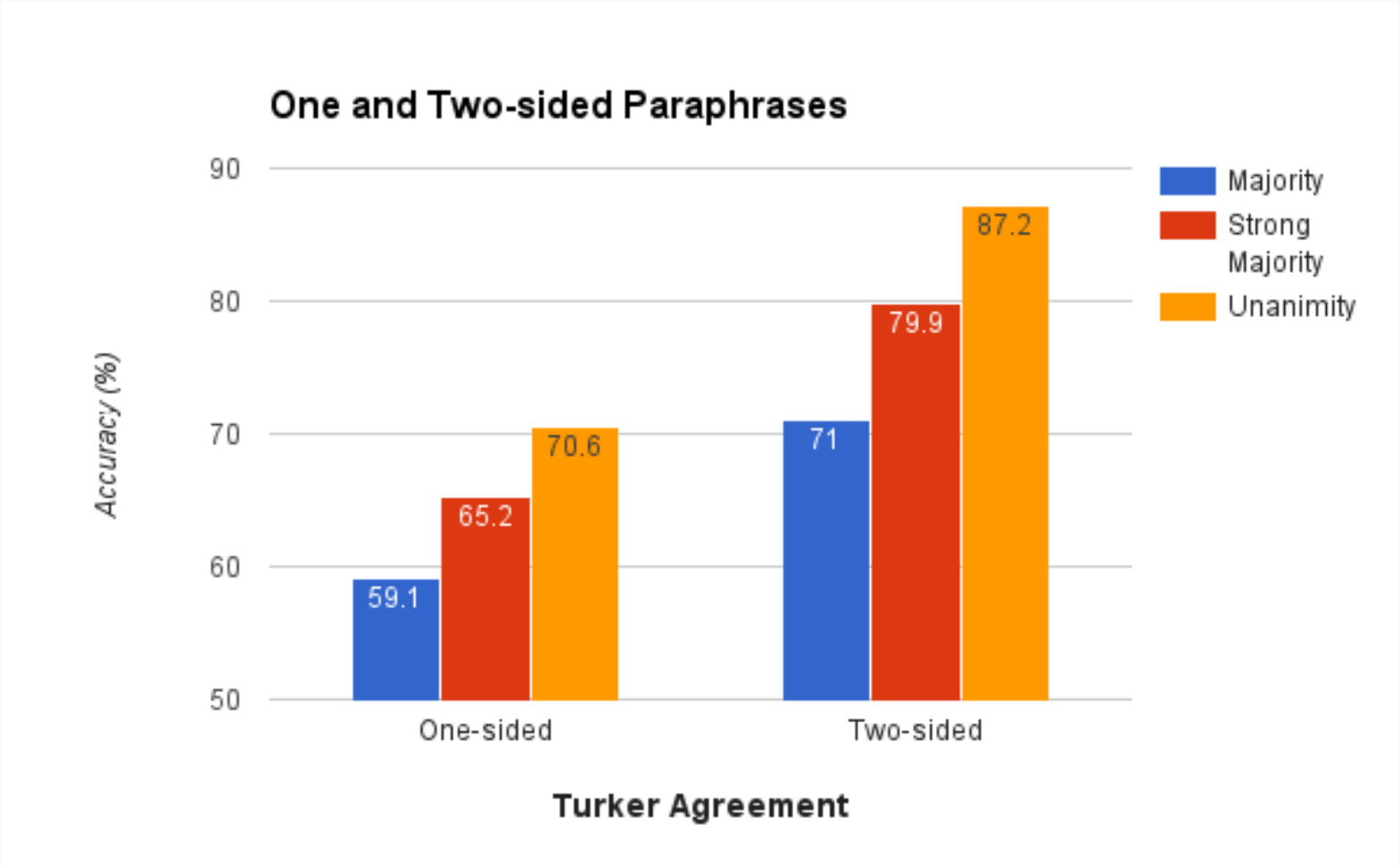
- 5 judgments for each sentence were collected from AMT workers and the judgments of identical sentences were collapsed
- “Neither” cases were excluded from analysis
- Comprehension questions were asked to prevent random choosing
- Agreement levels among the AMT workers:

Majority	$>$ 50% agreement
Strong Majority	$>$ 75%
Unanimity	$>$ 90%

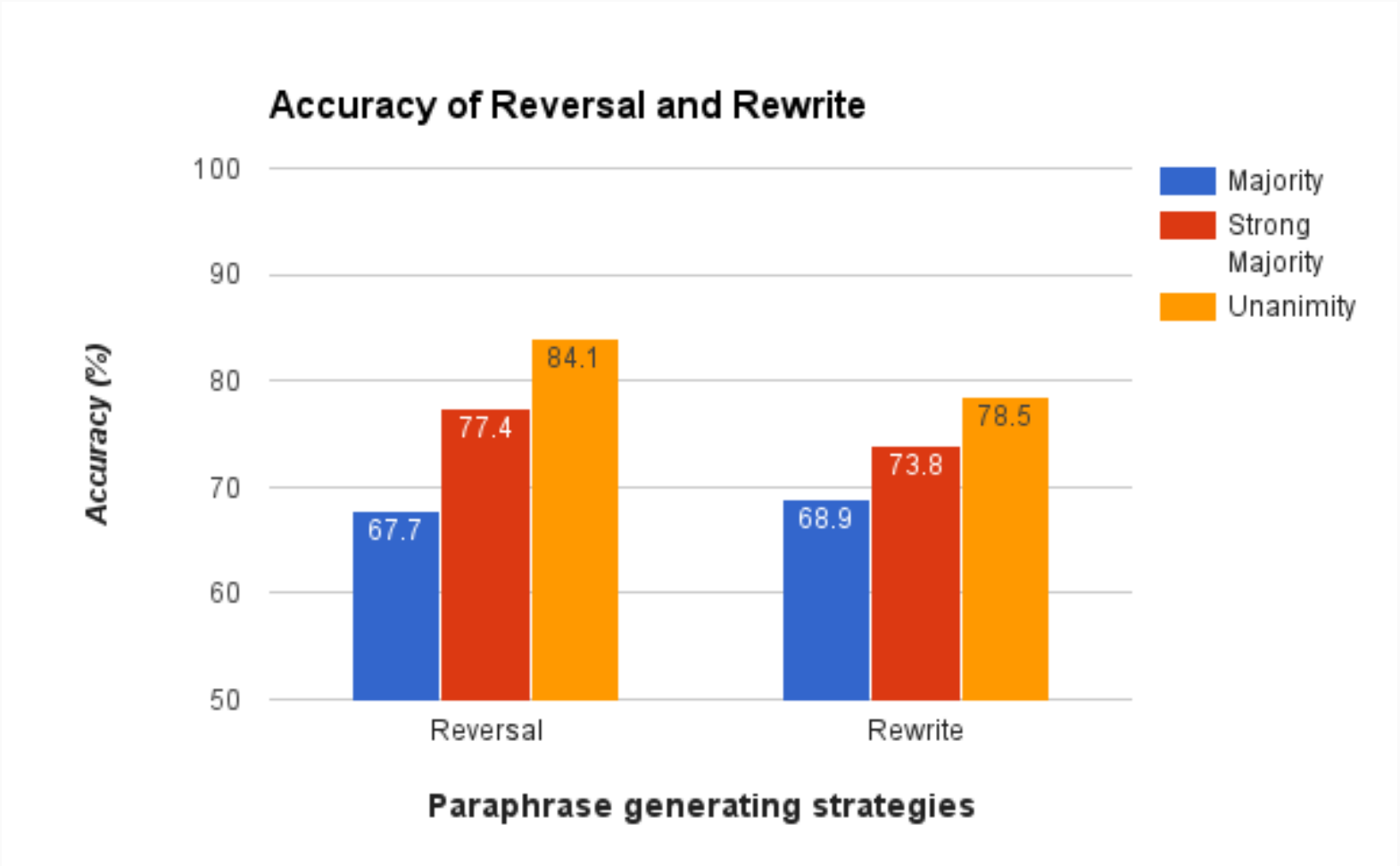
Coverage vs. Accuracy: Higher accuracy (but lower coverage) with greater agreement



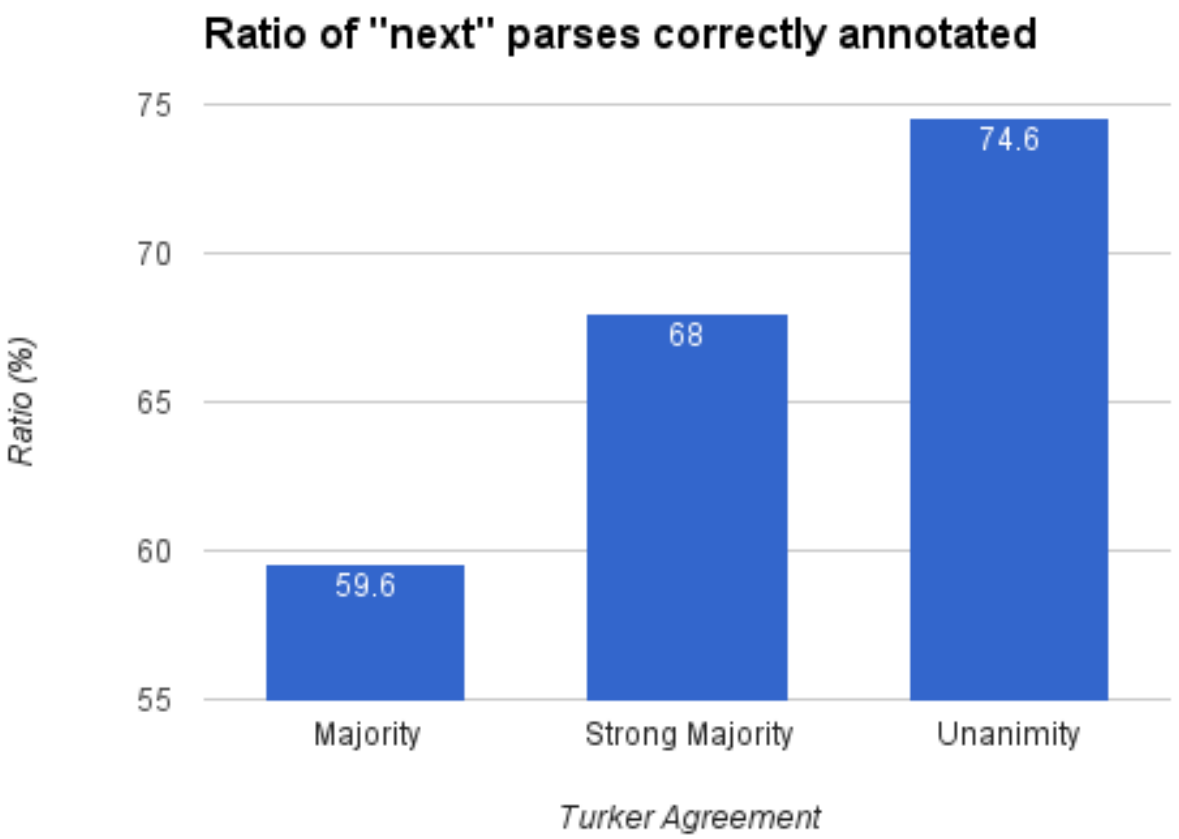
One-sided vs. Two-sided: Two-sided much more reliable



Reversals vs. Rewrites: Reversals at least as accurate



Potential correction to current parser



Manual analysis

Examined 43 sentences where unanimous AMT workers judgments did not agree with gold annotations and located the following **reasons for error**:

- Incompetent or broken realizations (29/43)
- Bad parses (11/43)
- Lack of context (3/43)

Preliminary parser retraining experiment

- Trained OpenCCG Parser with majority **AMT worker annotations** (along with original CCGbank data)
- Trained the parser separately in the two domains
- Evaluated the parser with 10-fold cross validation

Evaluation of retrained parser: an example

Parses were considered correct if the *top* and *next* dependencies occur **in the same order** as in gold: e.g., for the sentence *I saw the girl with the telescope*, if (saw, with) is annotated as the correct dependency,

n-best parses	Correct	Incorrect
1	...	...
2	(saw, with)	...
3	...	...
4	...	(girl, with)
5	(girl, with)	...
6	...	...
...	...	(saw, with)
25	...	...

Parser retraining results

	Dinosaur	Football
Train size	471	356
Eval size	291	226
Original acc.	0.701	0.668
Retrained acc.	**0.749**	**0.717**
Correction rate	0.243	0.32

- MacNemars chi-square test shows a significant improvement in the dinosaur domain ($p = 0.02$)
- No significant improvement on football data due to the smaller data size
- The retrained parsers do not differ significantly from the original parser ($p > 0.05$ for both) on the CCGbank development set

Conclusions

Conclusions and future work

- It is possible to obtain accurate crowd-sourced judgments from naive annotators with no instruction — pointing the way towards collecting parser training data on a **massive scale**

Conclusions and future work

- It is possible to obtain accurate crowd-sourced judgments from naive annotators with no instruction — pointing the way towards collecting parser training data on a **massive scale**
- The preliminary parsing experiment already suggests that automatic parsers can be retrained to achieve better parsing accuracy

Conclusions and future work

- It is possible to obtain accurate crowd-sourced judgments from naive annotators with no instruction — pointing the way towards collecting parser training data on a **massive scale**
- The preliminary parsing experiment already suggests that automatic parsers can be retrained to achieve better parsing accuracy
- In the future, we plan to experiment with parser adaptation with **multiple parsers and larger data sets**
- We also plan to experiment with generating paraphrases with **sentence splitting and simplification** (Siddharthan, 2006; Siddharthan, 2011)

Acknowledgments

We thank James Curran, Eric Fosler-Lussier, the OSU Clippers Group and the anonymous reviewers for helpful comments and discussion. This work was supported in part by NSF grant 1319318.

Thank you!

Incompetent realizations

Realization ok, but fails to reliably capture the different meaning in the parses

Usually involved just adding or deleting punctuation

Incompetent realizations: An example

The teeth were adapted to crush bivalves, gastropods and other animals with a shell or exoskeleton.

(animals, with): Same as the original sentence

(crush, with): The teeth were adapted to crush bivalves, gastropods and other animals , with a shell or exoskeleton.

Broken realizations

- Inappropriate heavy NP shift
- Long adverbials moved between verbs and their (other) complements
- Wrong modifier-modificand word order
- Wrong position of the particle for phrasal verbs
- Wrong preposition-complement position

Broken realizations: An example

They are thought to have gone extinct during the Triassic-Jurassic extinction event.

(gone, during): They are thought to have gone during the Triassic-Jurassic extinction event extinct.

(thought, during): They are thought during the Triassic-Jurassic extinction event to have gone extinct.

Bad parses

Although one parse is better than the other one for the disputed dependency, the rest of both parses are so broken that the realization cannot reliably capture the meaning difference

- Parsing *in* as a conjunction
- Bad parse in general

Bad parses: An example

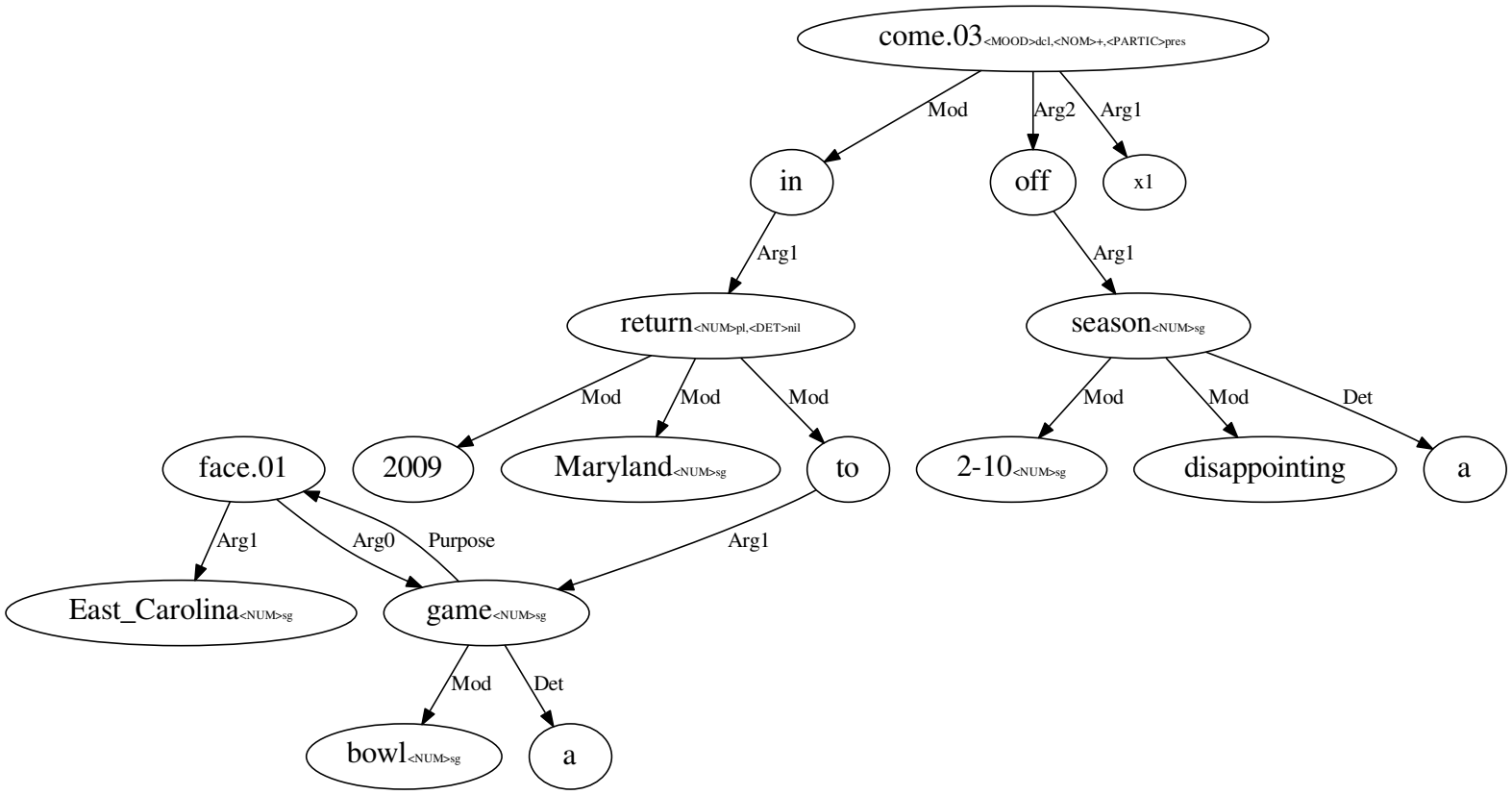
Coming off a disappointing 2-10 season in 2009 Maryland returns to a bowl game to face East Carolina.

(returns, to): Coming off a disappointing 2-10 season in 2009 returns to a bowl game to face East Carolina Maryland.

(Coming, to): Coming off a disappointing 2-10 season to a bowl game to face East Carolina in 2009 Maryland returns.

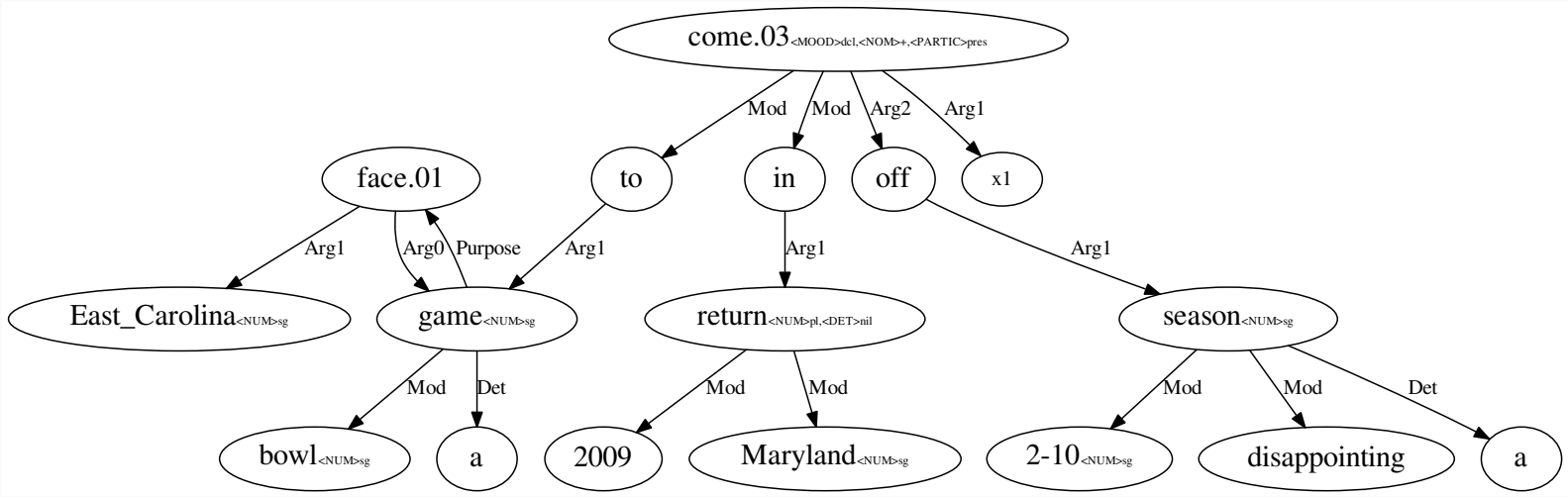
Bad parses: top parse

Coming off a disappointing 2-10 season in 2009 Maryland returns to a bowl game to face East Carolina.



Bad parses: next meaningfully distinct

Coming off a disappointing 2-10 season in 2009 Maryland returns to a bowl game to face East Carolina.



Lack of context

Turkers fail to choose the correct parse because of lack of context

Lack of context: An example

Michigan's backup center, Gerald_Ford, expressed a desire to attend the fair while in Chicago.

(attend, while): Michigan's backup center, Gerald_Ford, expressed a desire to attend while in Chicago the fair.

(expressed, while): Michigan's backup center, Gerald_Ford, expressed while in Chicago a desire to attend the fair.

Regression analysis

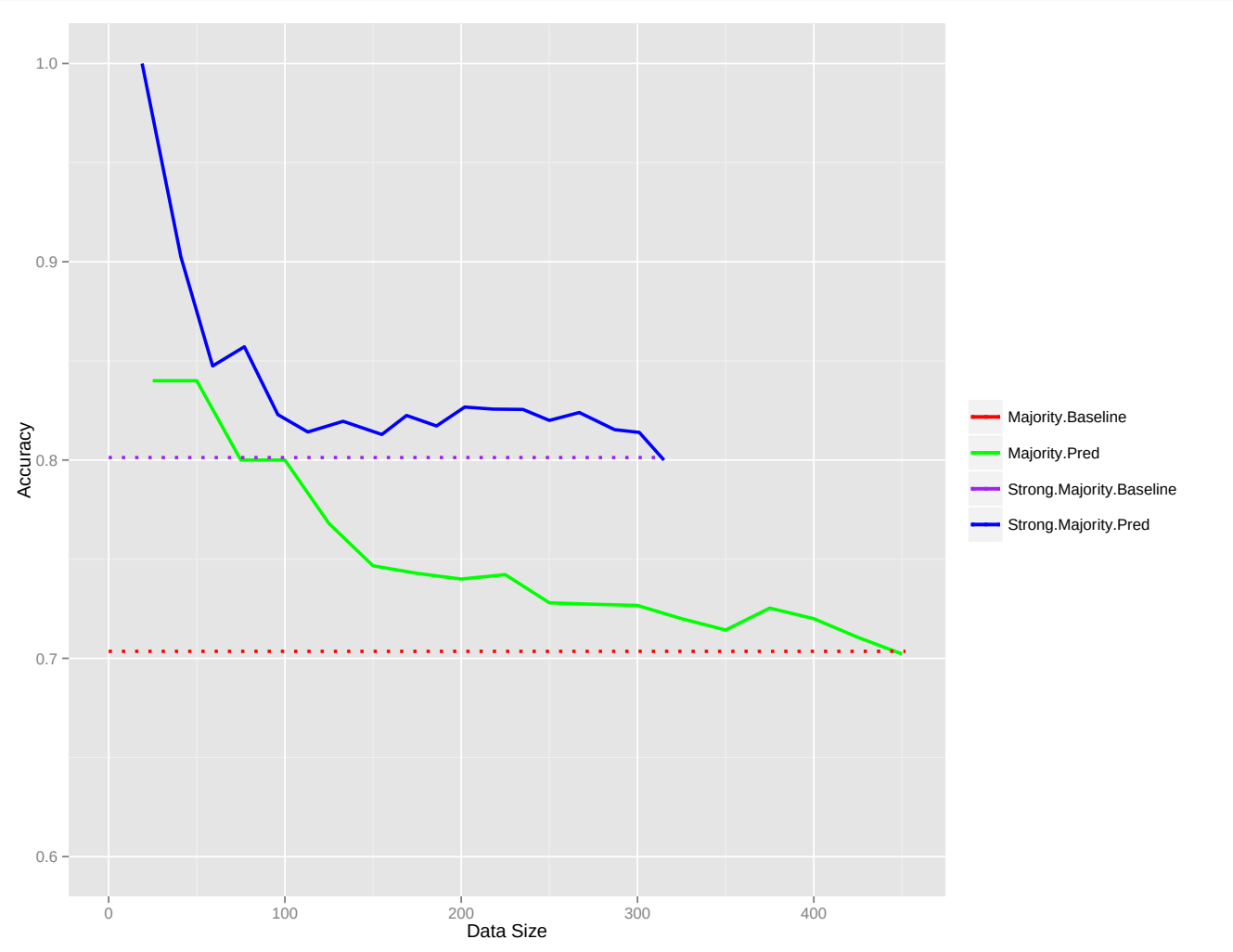
A regression analysis to determine the factors affecting AMT workers choices:

	One-sided		Two-sided	
	Maj	S. Maj	Maj	S. Maj
parse	-0.03	-0.05	0.01	0.01
bleu	3.05*	4.38**	1.68*	3.07**
rlz.glb	0.01	0.01	0.07**	0.103***

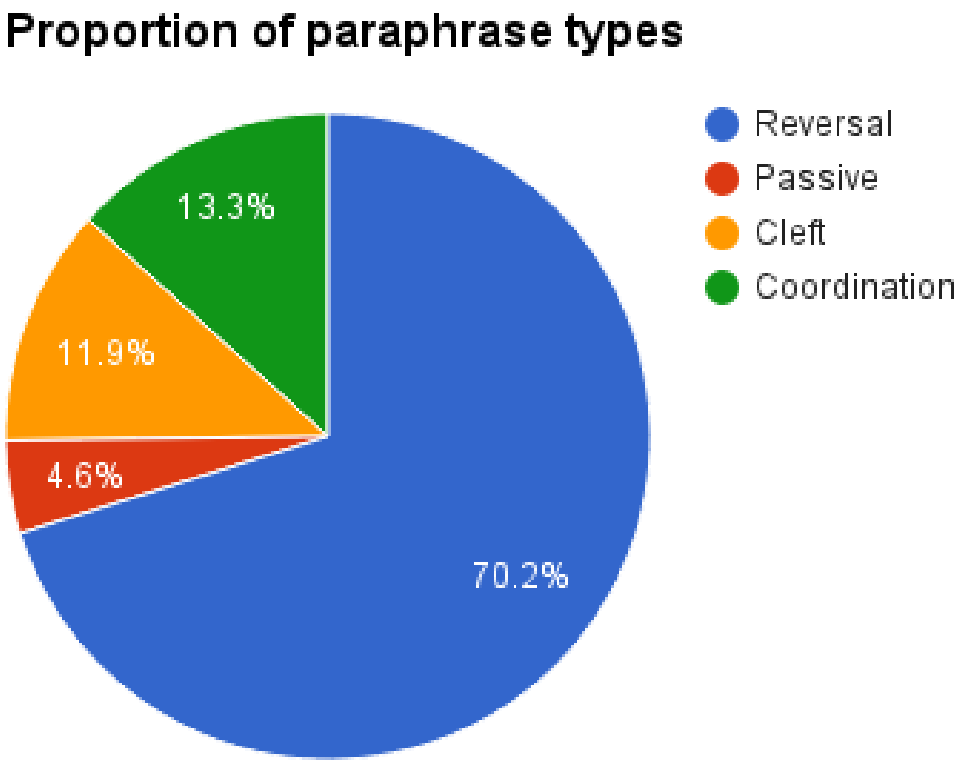
AMT workers tend to choose:

- the paraphrases **similar to** the original sentence
- the paraphrases with **higher fluency scores**

Regression analysis for coverage and accuracy trade-off



Distribution of test data



Data preparation

1. We collected 6,335 sentences from *Prehistoric Reptiles* and 7,779 from *Big 10 Conference Football*
2. After parsing the sentences and filtering sentences too short or too long, 5,063 sentences were found to be ambiguous
3. Valid paraphrases were generated for 3,605 sentences
4. 515 sentences from each domain were selected for validation experiment