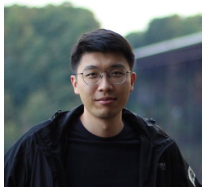
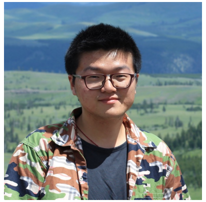
LREC-COLING 2024

SPLICE: A Singleton-Enhanced PipeLIne for Coreference REsolution

Yilun Zhu[1], Siyao Peng[2], Sameer Pradhan[3,4], Amir Zeldes[1]

1 GEORGETOWN UNIVERSITY

2 LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

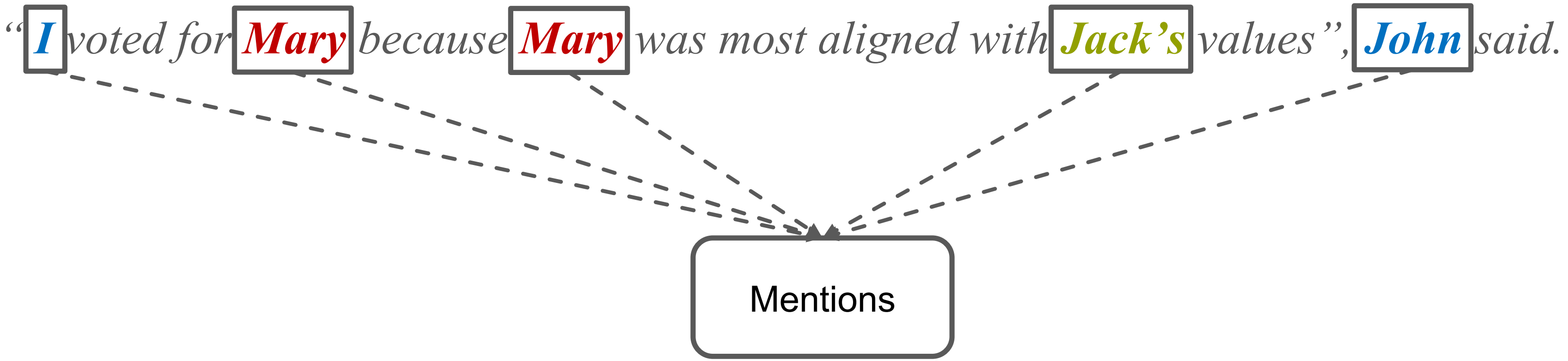
3 LDC Penn University of Pennsylvania

4 cemantix.org

Coreference Resolution and Mentions: An Example

*" **I** voted for **Mary** because **Mary** was most aligned with **Jack's** values", **John** said.*

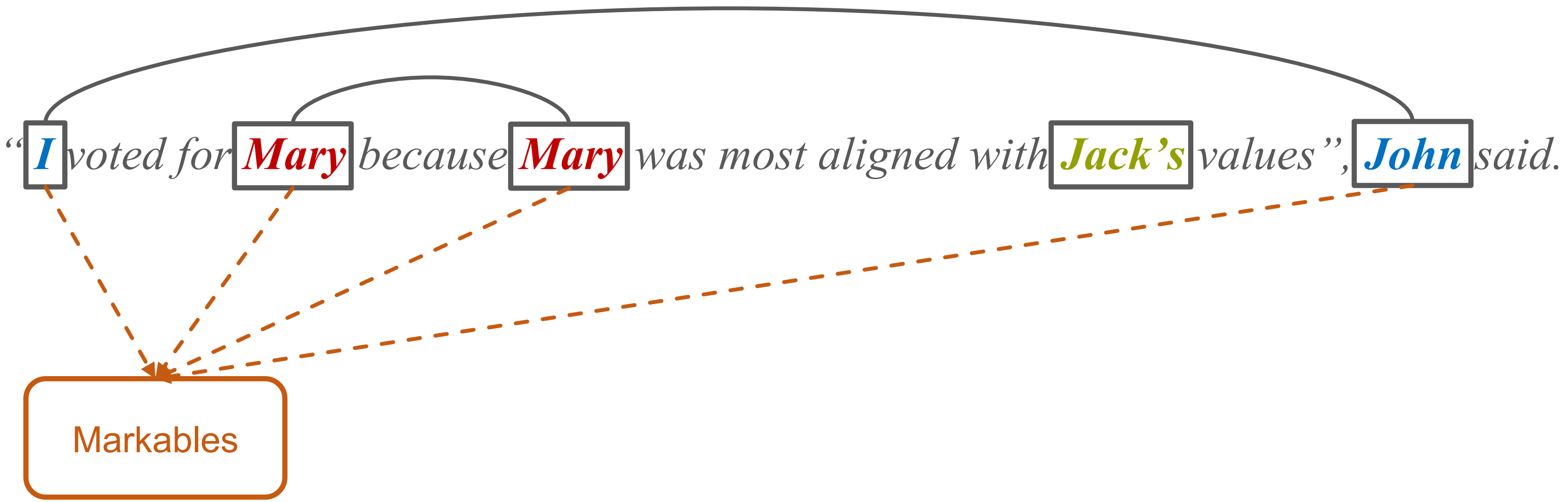
Coreference Resolution and Mentions: An Example



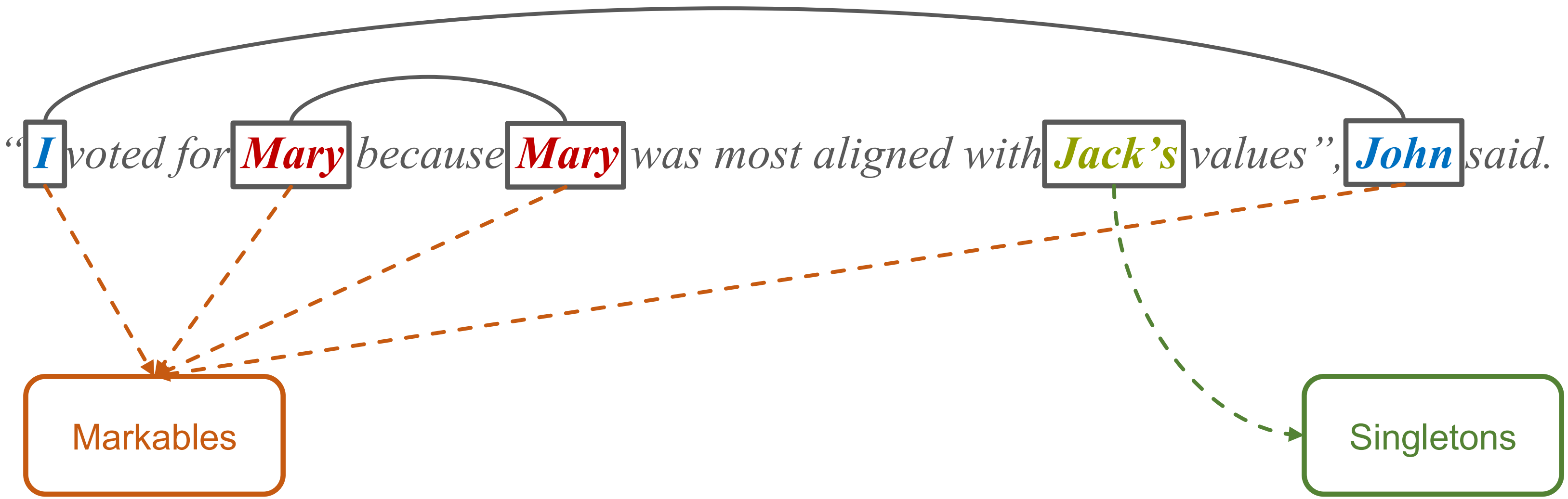
Coreference Resolution and Mentions: An Example

*“**I** voted for **Mary** because **Mary** was most aligned with **Jack’s** values”, **John** said.*

Coreference Resolution and Mentions: An Example



Coreference Resolution and Mentions: An Example



Discrepancy Between Linguistic Theory and Coref Models

- Singletons are important theoretically and empirically
  - Relate to how humans understand discourse and entity coherence (Grosz et al., 1995)
  - Singletons correspond to true negatives (Kübler and Zhekova, 2011)
  - Gold singletons improve coreference scores and help for generalization (Zhu et al, 2023)

- Existing datasets & models
  - OntoNotes lacks singleton annotation

    → models do pay attention to singleton spans
  - Limited interpretability of existing models

Utilizing Singletons from OntoNotes

- Use gold syntax structures (Raghunathan et al., 2010; Clark and Manning, 2015, 2016)
- Problems with these methods
  - Extracting NP subtrees → high recall in mention detection
    BUT generates a large number of precision errors (spans that are not valid mentions)
    - Generic you is a valid NP but is not a mention candidate for pair matching
    - See example in the next slide

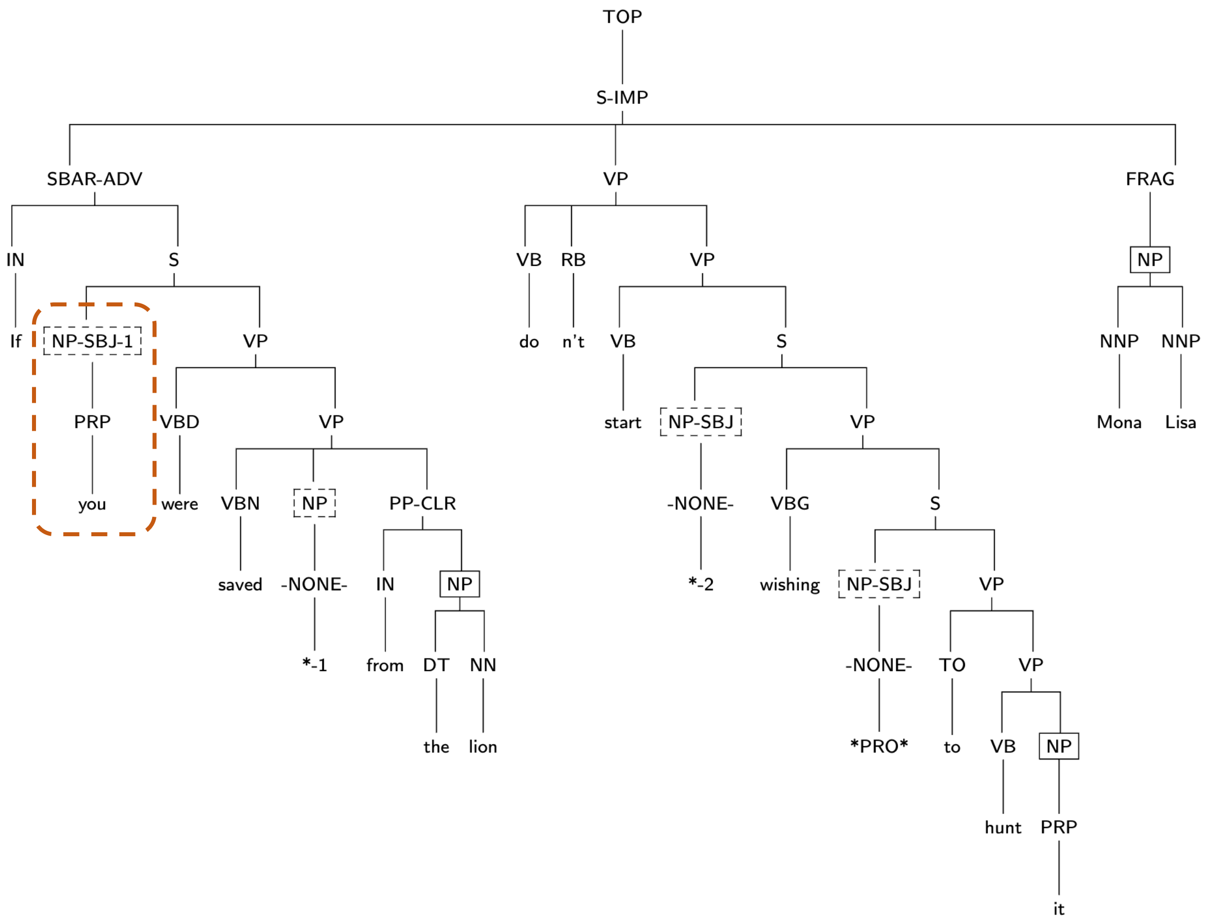
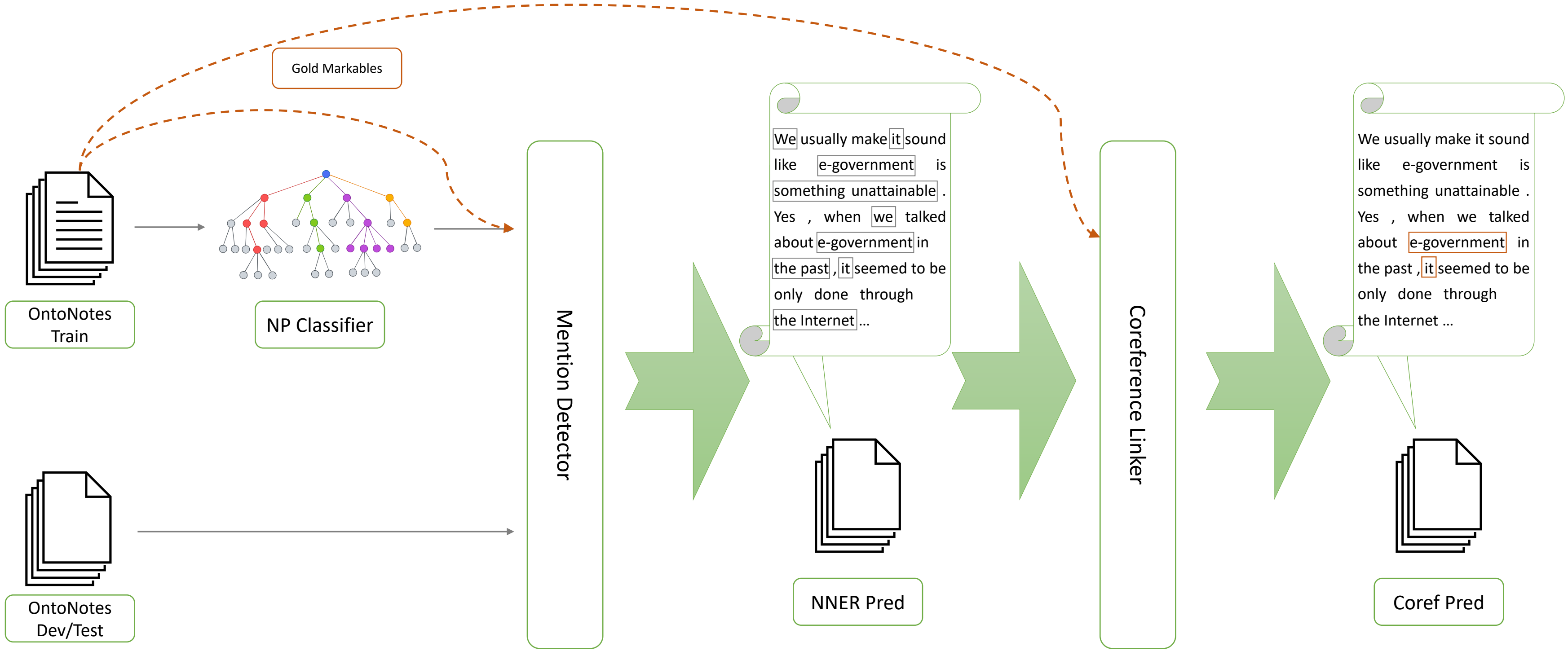


Figure 1: An example of the utilization of a syntax tree for the extraction of mentions. **Solid box:** NP is a candidate for coreference linking in OntoNotes. **Dashed box:** NP is not categorized as a mention.

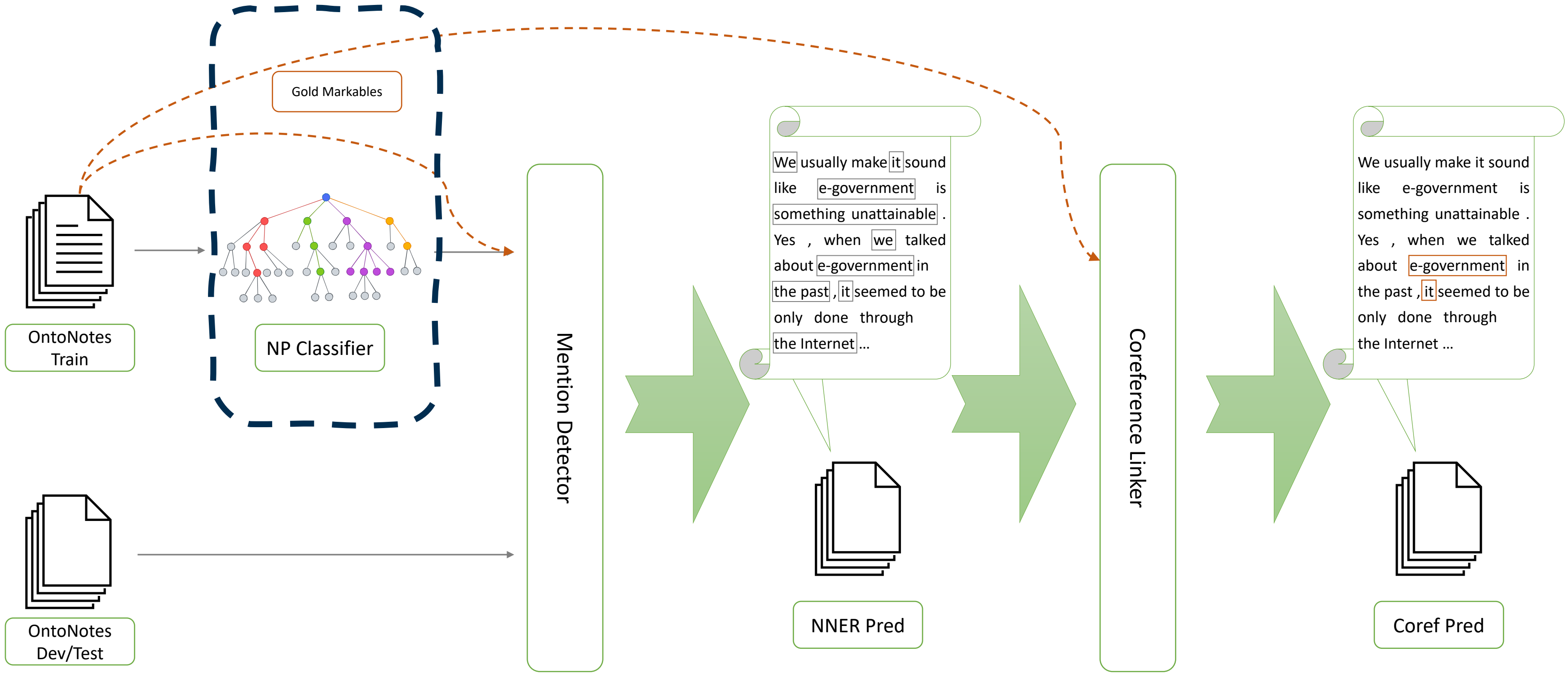
Utilizing Singletons from OntoNotes

- Generate silver singletons for the corpus (Recasens et al., 2013; Toshniwal et al., 2021)
- Problems with these methods
    - Biased pseudo-mentions
        - Missing atypical spans with semantic and syntactic disparities
    - Challenging evaluation
        - Unknown about the impact of mention detection to downstream coreference scores

Model Architecture



Model Architecture: Nominal Phrase Extraction



Nominal Phrase Extraction: Mention Classification

- Model
  - XGBoost
- Features
  - Mention-based features of the current NP, its parent phrases, and child phrases
    - POS tags
    - The usage of prepositions
    - Definite markers
    - Grammatical roles
    - Adverbial tags
    - ...
  - Features from other NPs that overlap with the current one
    - Their relative positions or hierarchical levels among other NPs
    - The largest and smallest interactive NP spans

Nominal Phrase Extraction: Mention Classification

- Dataset: required components
  - Gold syntax trees (constituency)
    - OntoNotes
    - ARRAU-RST news genre
  - Mention span annotation with OntoNotes
    - ARRAU super set (mostly)
    - OntoGUM
  - Singletons
    - ARRAU
    - OntoGUM
- Usage of the datasets
  - Training: ARRAU-RST

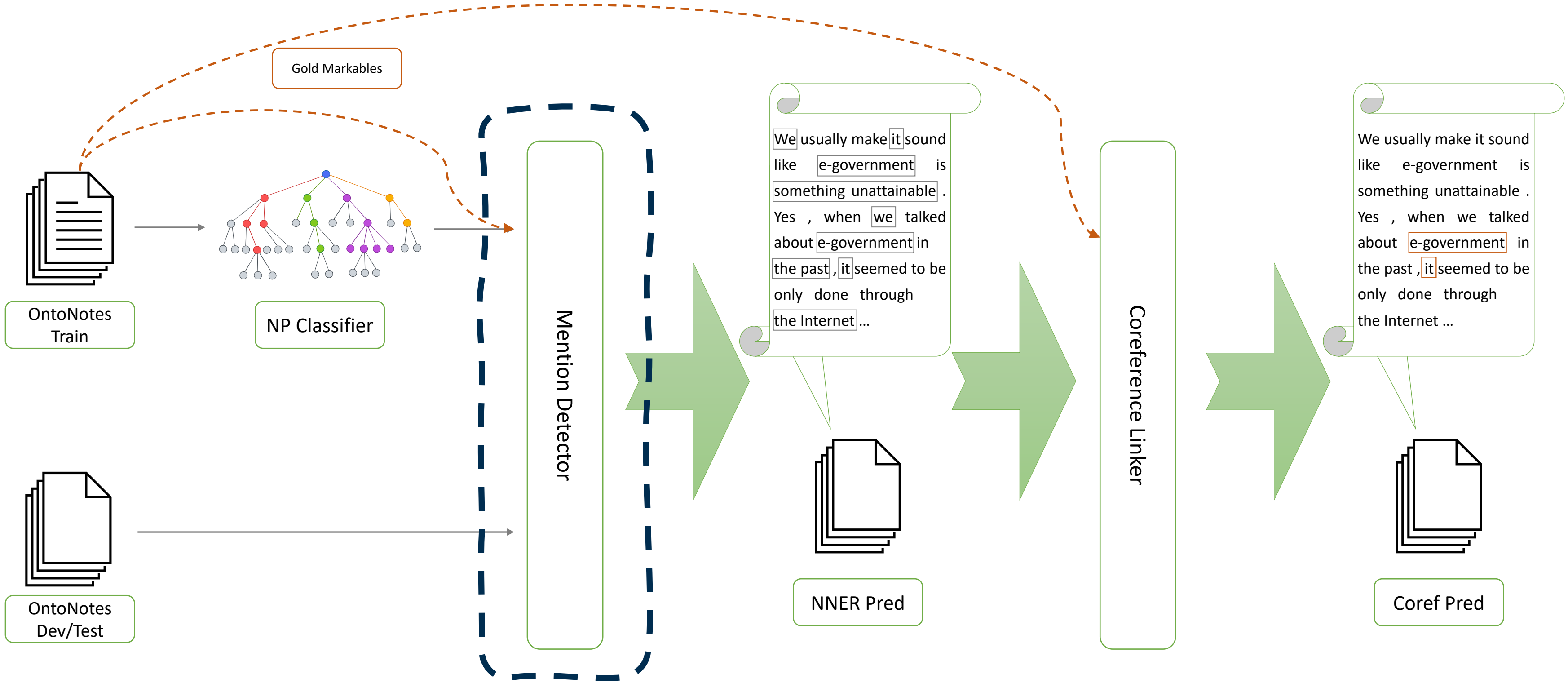
    → map gold NPs to near-gold singletons
  - Evaluation: ARRAU and OntoNotes

Dataset	P	R	F1
ARRAU	28.15	97.78	44.35
OntoNotes	39.46	91.65	55.16

Table 1: Results of coreference markables on ARRAU and OntoNotes test captured by the XGBoost classifier.

OntoNotes (Pradhan et al., 2013); ARRAU (Poesio et al., 2018); OntoGUM (Zhu et al, 2021)

Model Architecture: Mention Detection



Mention Detection

- Dataset
  - Training: OntoNotes
    - Use the classifier trained on ARRAU to predict positive and negative labels within the OntoNotes training dataset
    - Take the *union* of the classifier's outputs (positive labels) and gold coreference markables from the OntoNotes training set
  - Evaluation set
    - OntoNotes
    - OntoGUM

Mention Detection

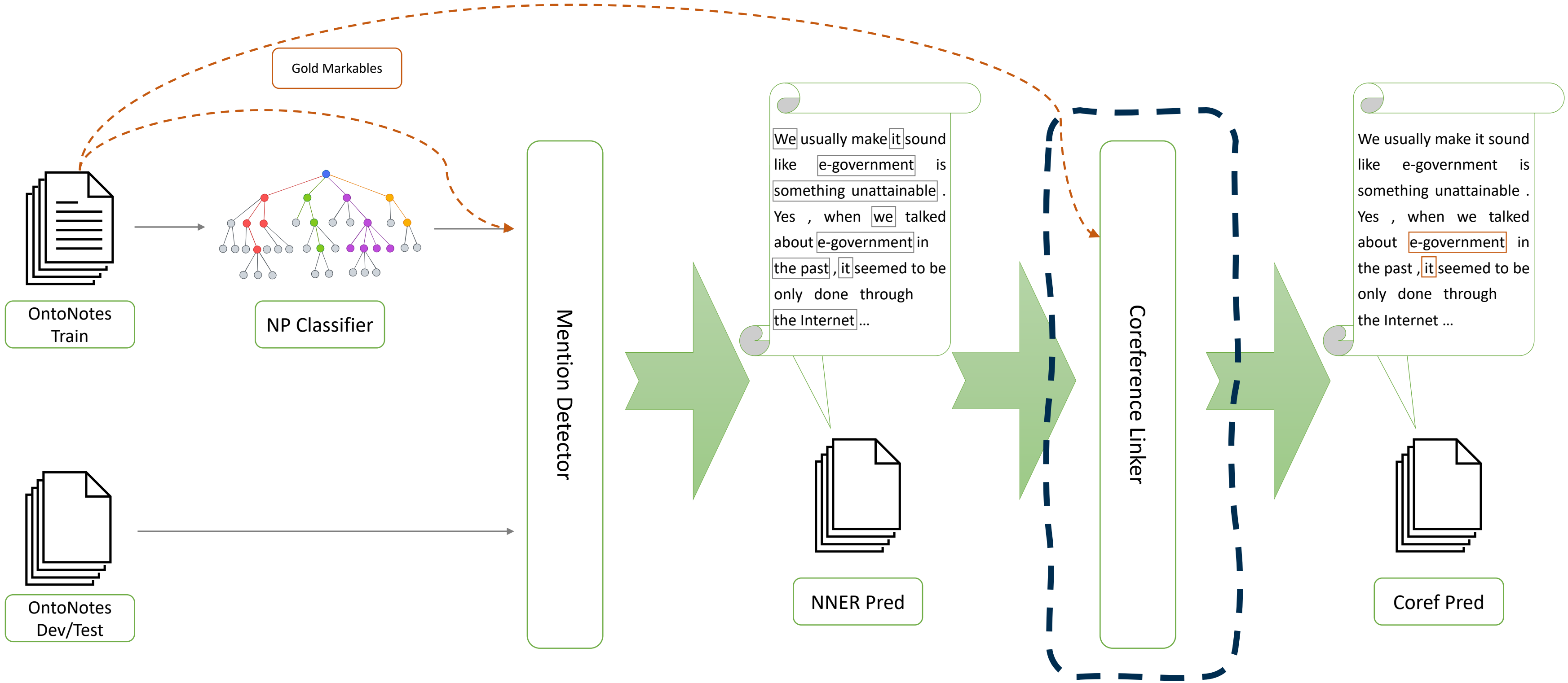
OntoNotes singleton data is publicly available!

- Model: Nested named-entity recognition (NNER) model
  - Sequence-to-set (Tan et al., 2021)
    - Focus on span
    - Ignore entity type, i.e., assign the same entity type *abstract* to every span

Data	Precision	Recall	F1
ONTONOTES-dev	37.84 (18,321/48,419)	95.64 (18,321/19,156)	54.22
ONTONOTES-test	37.75 (19,018/50,736)	96.23 (19,018/19,764)	54.23
ONTOGUM-test	37.21 (2,439/ 6,554)	91.66 (2,439/ 2,661)	52.94

Table 2: Mention detection performance on OntoNotes dev/test set and OntoGUM test set.

Model Architecture: Coreference



Coreference Model: Training

- Baseline end-to-end (Lee et al, 2017, 2018; Joshi et al, 2020)
  - Consider *all span possibilities* during coreference linking
  - Keep a *fixed* number of spans with top scores for coreference clustering

$$g_i = [x_{start(i)}, x_{end(i)}, \hat{x}_i, \varphi(i)]$$
$$s_m = FFNN_{\mathrm{m}}(g_i)$$

- SPLICE
  - Assign *identical mention scores* to all spans from mention detection
  - Utilize a *trainable parameter* w_m for the markable score

$$s_m = w_m$$

Coreference Model: Inference

- Inference (SPLICE)
  - ! mention spans and gold syntax trees cannot be used at test time
  - Two steps
    - Plain input → Mention detector → Nested mentions
    - Nested mentions → Coreference model → Coreference chains

Coreference Model: In-domain Results

- Comparable performance with the baseline model

	Mention Detection			MUC			B^3			$CEAF_{\phi4}$			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Joshi et al. (2020)	89.1	86.5	87.8	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Ours+MD	88.8	87.3	88.1	85.6	84.5	85.1	78.8	77.0	77.9	75.8	74.4	75.1	79.4
Ours+MD+GM (upperbound)	90.9	91.3	91.1	87.9	88.6	88.3	81.4	82.7	82.0	80.3	79.9	80.1	83.5

Table 3: Results on OntoNotes test set. **MD** denotes the model uses predictions from the mention detector; **GM** indicates the model uses gold coreference markables.

Coreference Model: In-domain Results

- Comparable performance with the baseline model
- Optimal scenario (gold markables) marks a nearly 4-point increase

	Mention Detection			MUC			B^3			$CEAF_{\phi4}$			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Joshi et al. (2020)	89.1	86.5	87.8	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Ours+MD	88.8	87.3	88.1	85.6	84.5	85.1	78.8	77.0	77.9	75.8	74.4	75.1	79.4
Ours+MD+GM (upperbound)	90.9	91.3	91.1	87.9	88.6	88.3	81.4	82.7	82.0	80.3	79.9	80.1	83.5

Table 3: Results on OntoNotes test set. **MD** denotes the model uses predictions from the mention detector; **GM** indicates the model uses gold coreference markables.

Coreference Model: Out-of-domain Results

- Improved mention detection scores

	Mention Detection			MUC			B^3			$CEAF_{\phi 4}$			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Joshi et al. (2020)	86.0	70.6	77.5	80.0	68.1	73.6	67.9	60.5	64.0	68.6	50.5	58.2	65.3
Ours+MD	85.3	73.5	78.9	78.8	70.6	74.5	66.5	63.5	64.9	68.3	52.0	59.0	66.4
Ours+GS (upperbound)	90.8	74.8	82.0	84.8	72.4	78.1	74.2	65.6	69.6	75.7	55.6	64.2	70.8

Table 4: Results on OntoGUM test set. **GS** indicates that our model uses gold singletons.

Coreference Model: Out-of-domain Results

- Improved mention detection scores
- Outperform the baseline model by 1.1 points

	Mention Detection			MUC			B^3			$CEAF_{\phi 4}$			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Joshi et al. (2020)	86.0	70.6	77.5	80.0	68.1	73.6	67.9	60.5	64.0	68.6	50.5	58.2	65.3
Ours+MD	85.3	73.5	78.9	78.8	70.6	74.5	66.5	63.5	64.9	68.3	52.0	59.0	66.4
Ours+GS (upperbound)	90.8	74.8	82.0	84.8	72.4	78.1	74.2	65.6	69.6	75.7	55.6	64.2	70.8

Table 4: Results on OntoGUM test set. **GS** indicates that our model uses gold singletons.

Coreference Model: Out-of-domain Results

- Improved mention detection scores
- Outperform the baseline model by 1.1 points
- Optimal scenario (gold mentions) marks a 5.5-point increase

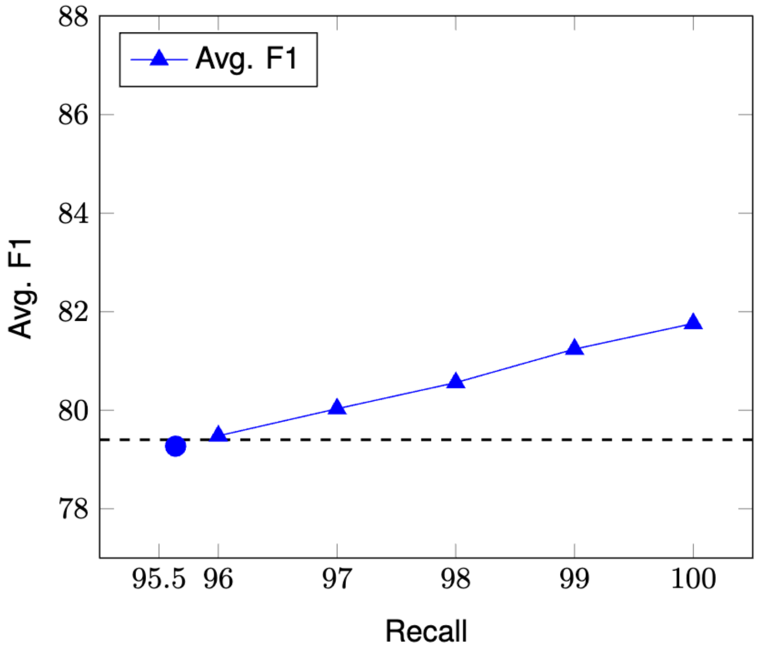
	Mention Detection			MUC			B^3			$CEAF_{\phi4}$			Avg. F1
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
Joshi et al. (2020)	86.0	70.6	77.5	80.0	68.1	73.6	67.9	60.5	64.0	68.6	50.5	58.2	65.3
Ours+MD	85.3	73.5	78.9	78.8	70.6	74.5	66.5	63.5	64.9	68.3	52.0	59.0	66.4
Ours+GS (upperbound)	90.8	74.8	82.0	84.8	72.4	78.1	74.2	65.6	69.6	75.7	55.6	64.2	70.8

Table 4: Results on OntoGUM test set. **GS** indicates that our model uses gold singletons.

Effect of Mention Detection: Recall

x-axis: mention detection
y-axis: coref

- Figure
  - Horizontal dashed line: <u>baseline</u>
  - Rounded data point: <u>F1 from SPLICE</u>
- Method
  - Randomly add gold coreference markables to increase recall score
- Optimal Scenario
  - Recall=100, Avg. F1 79 → 82

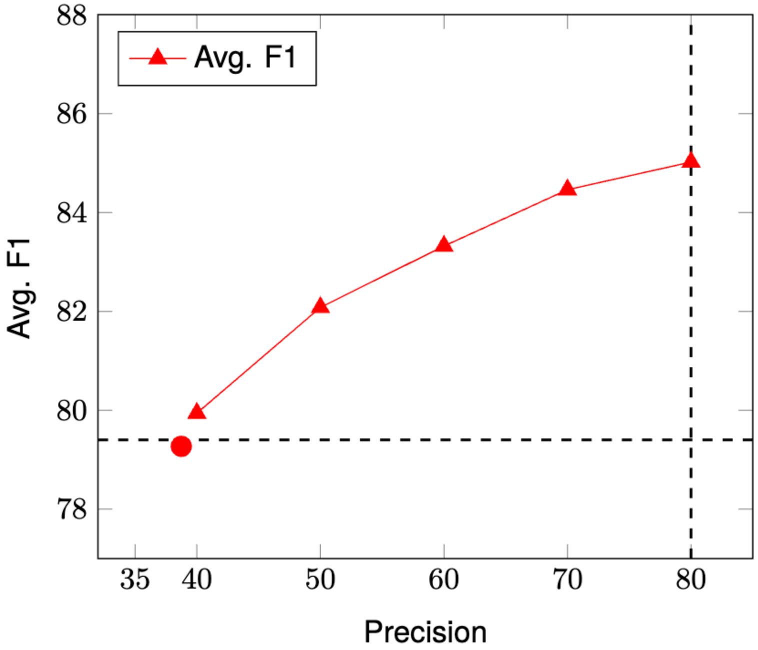


Effect of Mention Detection: Precision

x-axis: mention detection
y-axis: coref

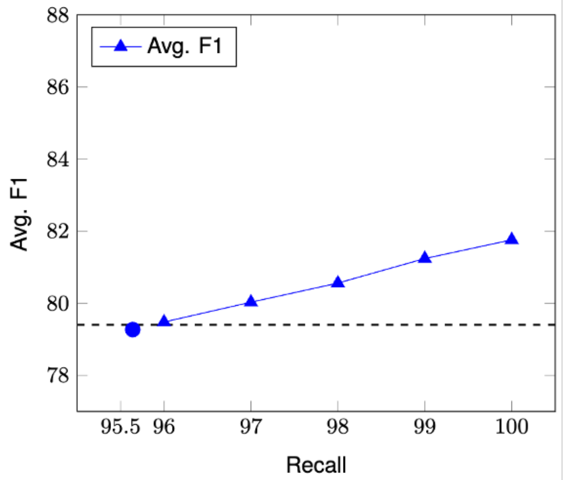
- Figure
  - Horizontal dashed line: baseline
  - Rounded data point: F1 from SPLICE
  - Vertical dashed line: Best precision with gold singletons

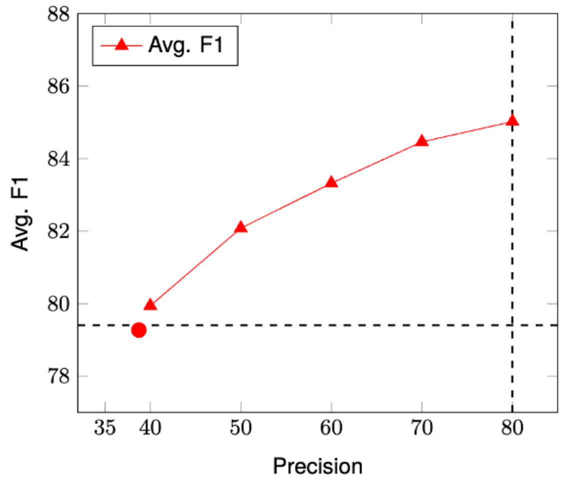
    Estimation of mentions: 19K×2 / 48K ≈ 80%
- Method
  - Randomly remove error predictions to increase precision score
- Optimal Scenario
  - Precision=80, Avg. F1 79 → 85



Effect of Mention Detection: Observation

- Reducing both mention precision and recall errors increase coreference resolution performance

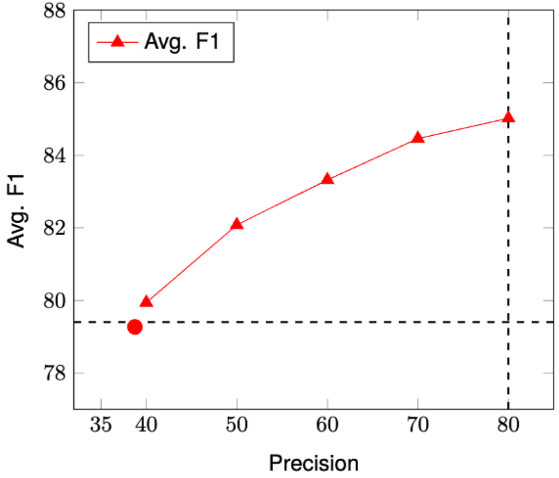




Effect of Mention Detection: Observation

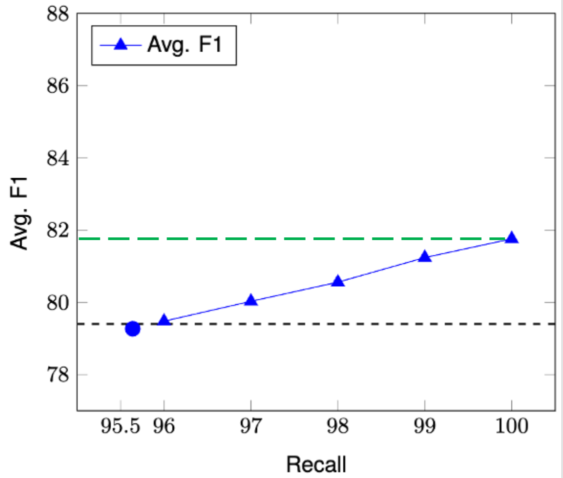
- Reducing both mention precision and recall errors increase coreference resolution performance

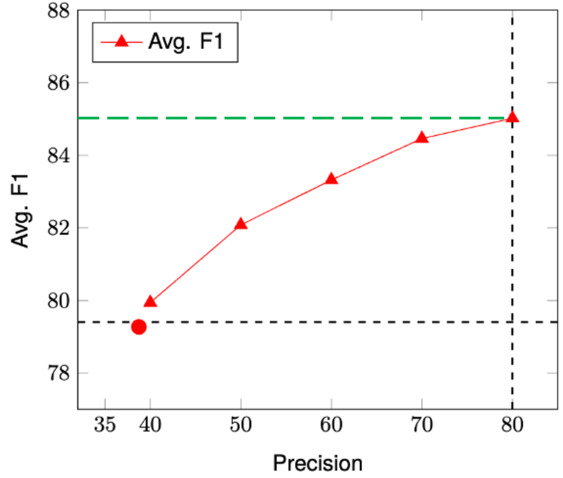
- Precision errors affect performance even with independent mention detector



Effect of Mention Detection: Observation

- Reducing both mention precision and recall errors increase coreference resolution performance
- Precision errors affect performance even with independent mention detector
- Precision improvement offers more significant benefits than recall for future coreference models





Effect of Mention Detection: Qualitative Analysis

Recall
Missing nested entity: Once the [Zhuhai - [Hong Kong] - Macao] bridge is built, it will no longer be a dream of tourists to enjoy gourmet food in Macao before having fun at Disneyland just an hour later .
Attachment of Prepositional Phrases: He just told [a story] uh from the beginning to the end.
Garden-path sentences: Like [the bones] xrays of his wisdom teeth also tell us something about his age.
Missing verbal referents: ... American military officials are now convinced that a unit of Marines [killed]$_{\#126}$ some 24 unarmed Iraqis ... One government official stated that [this atrocity]$_{\#126}$ showed " a total breakdown in morality . "
Gold Annotation Errors: They can volunteer at [any] [of thousands of non-profit institutions] , or participate in service programs required by high schools or encouraged by colleges or employers .
Precision
Redundant punctuations: [one .]
Redundant non-restrictive relative clauses: [5 p.m. EST – when stocks there plunged.]
Generic NPs: no media

Table 5: Major categories of recall and precision errors in OntoNotes dev set.

Conclusion

- A mention detection classifier that extracts mentions from syntactic structures and achieves ~94% recall
- A near-gold singleton annotated version of OntoNotes
- A pipeline-based neural coreference system, named SPLICE, using singletons, yielding results on par with the e2e approach in-domain and a +1.1 boost OOD
- Conduct a comprehensive analysis of the effect of mention detection to coreference linking
- **Release data and code at:** https://github.com/yilunzhu/splice

References

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In Proceedings of ACL-IJCNLP 2015.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In Proceedings of ACL 2016.

Association for Computational Linguistics.Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the

Association for Computational Linguistics.

Sandra Kübler and Desislava Zhekova. 2011. Singletons and coreference resolution evaluation. In Proceedings of RANLP 2011.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In Proceedings of EMNLP 2017.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of NAACL-HLT 2018.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In Proceedings of CRAC 2018.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In Proceedings of CoNLL 2013.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In Proceedings of EMNLP 2010.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition.

Yilun Zhu, Siyao Peng, Sameer Pradhan, and Amir Zeldes. 2023. Incorporating singletons and mention-based features in coreference resolution via multi-task learning for better generalization. In Proceedings of IJCNLP-AACL 2023.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In Proceedings of ACL-IJCNLP 2021.