

End-to-end Multilingual Coreference Resolution with Mention Head Prediction

Ondřej Pražák and Miloslav Konopík

Department of Computer Science and Engineering,
Faculty of Applied Sciences, University of West Bohemia, Technická 8, 306 14 Plzeň
Czech Republic

October 17, 2022

Introduction

- Based on the baseline system.
- End-to-end neural model.
- Joined model for all datasets finetuned on each dataset separately.
- Mention head predicted automatically from a span
- Simple syntax modeling.
- Sprint (implemented in three weeks).

Model

- Standard end-to-end coreference resolution [Xu & Choi, 2020], [Lee *et al.*, 2017]
- Operates over all possible spans up to maximum length
- Learning:
 1. score spans for being a mention.
 2. extract top k spans.
 3. score mention pairs to find best antecedents.
 4. extract top n best antecedent candidates for each mention.
 5. maximize marginal probability of all correct antecedents.

Learning

$$s(i, y_i) = \begin{cases} 0 & y_i = \epsilon \\ s_m(i) + s_m(y_i) + s_a(i, y_i) & y_i \neq \epsilon \end{cases},$$

$$P(y_i | D) = \frac{\exp(s(i, y_i))}{\sum_{y' \in Y(i)} \exp(s(i, y'))}$$

$$J(D) = \log \prod_{i=1}^N \sum_{\hat{y} \in Y(i) \cap \text{GOLD}(i)} P(\hat{y})$$

Multilingual training

- Joined model trained on all datasets
- XLM Roberta Large
- Finetuned on each dataset separately
- For small datasets we need pretraining of new parameters.

Model	Pretrained params	New params
mBERT	180M	40M
XLM-R	350M	50M

Table: Number of trainable parameters of the models

Mention head prediction

- Predicting whole spans probably does not optimize evaluation metric in the best way
- Better to output just the heads
- We do not know the rules for selecting the heads
- Rule-based selection of a syntactic head might not be the best solution.
- We train the model to predict the head from a span representation.

Partial matching

- Mentions considered correct if:
- The head of the gold cluster is included in predicted cluster.
- All word in the predicted mention is included in gold cluster.
- It is sufficient to predict only the heads.

Mention Head Prediction – Model

- Another classification head on the top of the model.
 1. Head position prediction.
 2. Binary classification of a cluster-word pair.
- Multiple words can be predicted as heads.
- If no head is predicted ($P < 0.5$ for all words):
 1. Most probable word
 2. All words

Tree representation

- Encode syntactic information into the model.
- Gold trees for some datasets.
- Needed to predict heads.

The Model

- For each word, the path to ROOT up to length 5 is added.
- Path to ROOT: Concatenation of word and relation embeddings.

Results

Dataset	BASELINE	Monoling	XLM-R	joined	+dev	+S2H	+Tree	CorPipe
ca_ancora	63.74	69.61	66.19	68.81	70.55	69.91	68.32	78.18
cs_pcedt	70	73.74	73.55	73.85	74.07	71.12	73.61	78.59
cs_pdt	67.27	69.81	70.99	70.63	71.49	72.42	70.99	77.69
de_parcorfull	33.75	43.04	33.75	68.91	73.9	68.3	65.29	65.52
de_potsdamcc	55.44	58.81	59.03	70.35	66.02	68.68	67.35	70.69
en_gum	62.59	68	66.27	68.16	68.31	66.88	67.39	72.50
en_parcorfull	36.44	25.84	36.44	30.21	31.9	23.45	40.05	39.00
es_ancora	65.98	60.12	67.99	71.24	71.48	72.32	72.04	81.39
fr_democrat	55.55	56.76	55.94	59.8	60.12	61.39	60.03	65.27
hu_szegedkoref	52.35	59.76	60.68	63.24	65.01	64.67	62.77	63.15
lt_lcc	64.81	66.93	64.81	66.34	68.05	67.49	64.01	69.92
pl_pcc	65.34	75.2	73.19	73.66	74.46	74.56	73.38	78.12
ru_rucor	67.66	69.33	77.5	75.5	74.82	76.02	75.94	79.34
avg	58.53	61.30	62.03	66.21	66.94	65.94	65.94	70.72

- Trees do not help (bug)
- Span-to-head helps, but not much (bad tree representation)

Post-evaluation Experiments

- Much worse results than CorPipe for some languages (Czech, Spanish, and Catalan).
- Reasons:
 - Bug in tree representations
 - Employed model does not optimize singletons.

$$J(D) = \log \prod_{i=1}^N \sum_{\hat{y} \in Y(i) \cap \text{GOLD}(i)} P(\hat{y})$$

- Loss over gold antecedents
 - Empty nodes
1. Fixed tree modeling.
 2. Singletons modeling added.
 3. Experiments with Rembert. Better for some datasets, but slightly worse in average.

Singletons – Solutions

- Many possibilities how to add singleton modeling.

Another Dummy Antecedent

- Singletons are mapped to dummy antecedent meaning that a span corresponds to a mention but has no real antecedent
- Discard binary score for singletons. It does not make sense to measure similarity to dummy antecedent.
- **Separate FFNN for Singletons** – The mention score for singletons can be predicted by separate layer.

Separate Mention Modeling

- The score for a span being a mention is added separately to the loss function.
- This variant does not treat singletons differently from other mention in the mention scoring step.

$$J(D) = \log \prod_{i=1}^N \sum_{\hat{y} \in Y(i) \cap \text{GOLD}(i)} P(\hat{y}) + y_m^{(i)} \cdot \sigma(s_m(i)) + (1 - y_m^{(i)}) \cdot \sigma(-s_m(i)) \quad (1)$$

where $y_m^{(i)}$ is 1 if span i corresponds to gold mention.

Post-evaluation Experiments – Results

	+Trees	+singletons	Best	CorPipe
ca_ancora	+1.21	+0.2	70.21	78.18
cs_pcedt	+1.53	+0	74.44	78.59
cs_pdt	+3.3	+1.12	74.64	77.69
de_parcorfull	+0	+3.24	70.76	65.52
de_potsdamcc	+0.5	+0.53	72.7	70.69
en_gum	+0	+2.00	71.31	72.50
en_parcorfull	+2.18	+0.5	37.56	39.00
es_ancora	+0.6	+0.4	72.59	81.39
fr_democrat	+1.34	+0.5	62.23	65.27
hu_szegedkoref	+2.00	+0.5	65.94	63.15
lt_lcc	+0	+1.43	66.5	69.92
pl_pcc	+0.9	+0.54	75.36	78.12
ru_rucor	+2.01	+0	76.82	79.34
avg	+1.18	+0.8	68.54	70.72

- Still much worse than CorPipe (for Czech, Catalan, and Spanish)
- Empty Nodes?

Discussion



CorefUD dataset	Total size					
	docs	sents	words	empty	singletons	discont.
Catalan-AnCora	1550	16,678	488,379	6,377	74.6%	0%
Czech-PDT	3165	49,428	834,721	33,086	35.3%	3.1%
Czech-PCEDT	2312	49,208	1,155,755	45,158	1.4%	4.1%
English-GUM	150	7,408	134,474	0	75%	0%
English-ParCorFull	19	543	10,798	0	6.1%	0.7%
French-Democrat	126	13,054	284,823	0	81.8%	0%
German-ParCorFull	19	543	10,602	0	5.8%	0.3%
German-PotsdamCC	176	2,238	33,222	0	76.5%	6.3%
Hungarian-SzegedKoref	400	8,820	123,976	4,849	7.9%	0.4%
Lithuanian-LCC	100	1,714	37,014	0	11.2%	0%
Polish-PCC	1828	35,874	538,891	864	82.6%	1.0%
Russian-RuCor	181	9,035	156,636	0	2.5%	0.5%
Spanish-AnCora	1635	17,662	517,258	8,111	73.4%	0%

Discussion & Conclusion

- Joined multilingual training helps a lot.
- Also syntactic information important for CorefUD.
- For official scoring metric, predicting heads only is better than predicting whole spans.
- Still surprisingly low results for some datasets.

Thank you for your attention.
Questions?

References I

-  Lee, Kenton, He, Luheng, Lewis, Mike, & Zettlemoyer, Luke. 2017. End-to-end Neural Coreference Resolution. *Pages 188–197 of: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics.
-  Xu, Liyan, & Choi, Jinho D. 2020. Revealing the Myth of Higher-Order Inference in Coreference Resolution. *Pages 8527–8533 of: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Online: Association for Computational Linguistics.