Google DeepMind

Multilingual Coreference Resolution with seq2seq Models

Bernd Bohnet

based on joined work with Chris Alberti, Michael Collins

06/12/2023

Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- Failures and Insights
- Training Schema
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- Trying to predict the Future

Outline

- **Why a text-to-text paradigm**
- Transition-based Coreference Resolution
- Failures and Insights
- Training Schema
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- Trying to predict the Future

Why Coreference Resolution via a text-to-text paradigm

Google DeepMind

- Enabling most advanced LLMs
  - We developed a text-based system to predict coreferences.
- Simplifying coreference resolution
  - We use joint prediction of mention and links
  - Text-2-text without complicated higher-order model
- Top accuracy for a large number of languages
  - SotA for CoNLL-2012: English, Chinese, and Arabic
  - SotA for SemEval-2010 datasets: Catalan, German, Dutch, etc.
  - High Zero-Shot performance for many other languages: 100+

How does the transition system work

- We iterate sentence-wise over the input documents
- For each sentence, we predict the coreference links
- We formalize and implement the coreference resolution as a transition system
    - **Link**: Create a reference to an antecedent
    - **Append**: Add a references to a coreference cluster
    - **Shift**: Continue with the next sentence

Outline

- Why a text-to-text paradigm
- **Transition-based Coreference Resolution**
- Failures and Insights
- Training Schema
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- Trying to predict the Future

Example

Input: *Speaker-A* I_2 still have n't gone to that fresh French restaurant by your house *Speaker-A* I_{17} 'm like dying to go there. *Speaker-B* You mean the one right next to the apartment *Speaker-B Yeah Yeah Yeah*

The system predicts the coreferences for a single sentence at a time.

Example: Link-Append Transition system

Google DeepMind

Input: | # *Speaker-A* I_2 still have n’t gone to that fresh French restaurant by your house ** # *Speaker-A* I_{17} ’m like dying to go there *Speaker-B* You mean the one right next to the apartment # *Speaker-B* *Yeah Yeah Yeah*

We add symbols to mark the focus sentence start | and end ** and the speaker with #

Example: Link-Append Transition system

Input: | # *Speaker-A* I_2 still have n't gone to that fresh French restaurant by your house ** # *Speaker-A* I_{17} 'm like dying to go there. *Speaker-B* You mean the one right next to the apartment # *Speaker-B Yeah Yeah Yeah*

Prediction: **SHIFT**

As the first sentence has no coreferences to previous text, the systems predicts a SHIFT-transition

Example: Link-Append Transition system

Input: # *Speaker-A* I$_{2}$ still have n’t gone to that fresh French restaurant by your house | # *Speaker-A* I$_{17}$ ’m like dying to go there ** *Speaker-B* You mean the one right next to the apartment # *Speaker-B Yeah Yeah Yeah*

The focus shifts to the next sentence

Example: Link-Append Transition system

Google DeepMind

Input: # *Speaker-A* I_2 still have n’t gone to that fresh French restaurant by your house | # *Speaker-A* I_{17} ’m like dying to go there ** *Speaker-B* You mean the one right next to the apartment # *Speaker-B Yeah Yeah Yeah*

Predictions: $I_{17} \rightarrow I_2$
SHIFT

The system predicts a link-transition

Example: Link-Append Transition system

Input: # *Speaker-A* [1 I] still have n’t gone to that fresh French restaurant by your house # *Speaker-A* [1 I] ’m like dying to go there | # *Speaker-B* You mean the one right next to the apartment ** # *Speaker-B Yeah Yeah Yeah*

Predictions:

The input is annotated with the coreferences cluster [1 I]

Example: Link-Append Transition system

Input: # *Speaker-A* [1 I] still have n't gone to that fresh French restaurant by your house # *Speaker-A* [1 I] 'm like dying to go there | # *Speaker-B* You mean the one right next to the apartment ** # *Speaker-B Yeah Yeah Yeah*

Predictions: You → [I

the apartment → your house

the one right next to the apartment → that fresh French restaurant by your house

SHIFT

Example: Link-Append Transition system

Input: # *Speaker-A* [1 I] still have n't gone to [3 that fresh French restaurant by [2 your house]] # *Speaker-A* [1 I] 'm like dying to go there # *Speaker-B* You [3 mean the one right next to [2 the apartment]] *| # Speaker-B Yeah Yeah Yeah ***

Example: Link-Append Transition system

Input: # *Speaker-A* [1 I] still have n't gone to [3 that fresh French restaurant by [2 your house]] # *Speaker-A* [1 I] 'm like dying to go there # *Speaker-B* You [3 mean the one right next to [2 the apartment]] # *Speaker-B Yeah Yeah Yeah*

Predictions: **SHIFT**

Let’s take a step back and investigate why Iterative Coreference Resolution

Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- **Failures and Insights**
- Training Schema
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- Trying to predict the Future

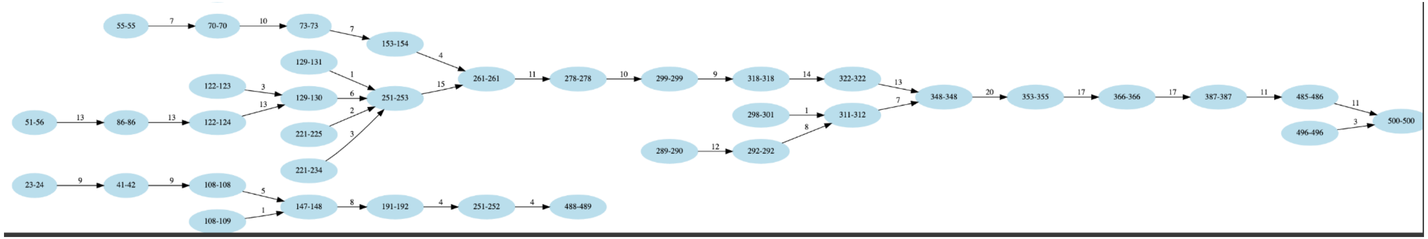
Simple things are sometimes difficult

We tried a number of solutions before arriving at a simpler and better one.

- Initially we worked on predicting mentions and coreference links separately then computing a graph cover (Hamilton path) over all mentions to obtain coreference chains

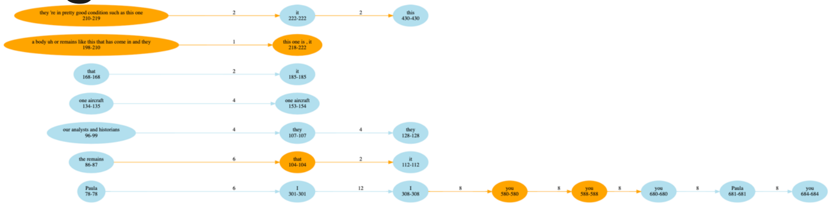
  → Filtering of overpredicted mentions via links and computing a graph cover did not work well probably due to the lack of joint training.

  → Higher order approach (Lee et al 2018) require span-based scoring and hence are not a good fit to a text-2-text paradigm to use



Reached only 79% F1 and we gave up on it

Insight: failurs are curcial for progress



- Text-2-text models are not ideal for mention predictions and subsequent linking as they try to balance precision and recall

  => The model might leave out mentions that seem less likely

- Predicting mentions and chains given a coreference link (or chain) yields higher has a higher probability.

$$P(Y \leftarrow Z \,|\, X \leftarrow Y) >\sim P(Y \leftarrow Z)$$

The elephant in the room:
Predicting the full sequence vs. a step at a time

- The transition-based is iterative: taking carefully a step at a time
- We tried to predict sequence at once and got lower accuracy

Hypothesis on lower performance:
- **Sequence version**: The encoder sees only the input and does not know about previous introduced coreferences
- **Interative/**Stepwise: the new input is encoded including previous corefernces. The encoder sees previous corefernece chains

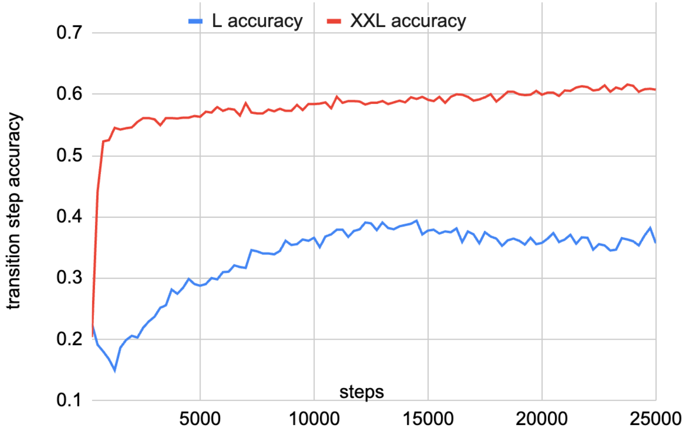
Might have additional reasons too such as T5 was trained on shorter outputs and inputs.

Training Schema: Model Size

Mystery: Why do we see pure performance on smaller models

- L-mT5 has quite pure performance; even hill like
- XXL-mT5 has a nice steady increasing accuracy curve
- Unusual high gain from xl:78 -> xxl: **83** F1

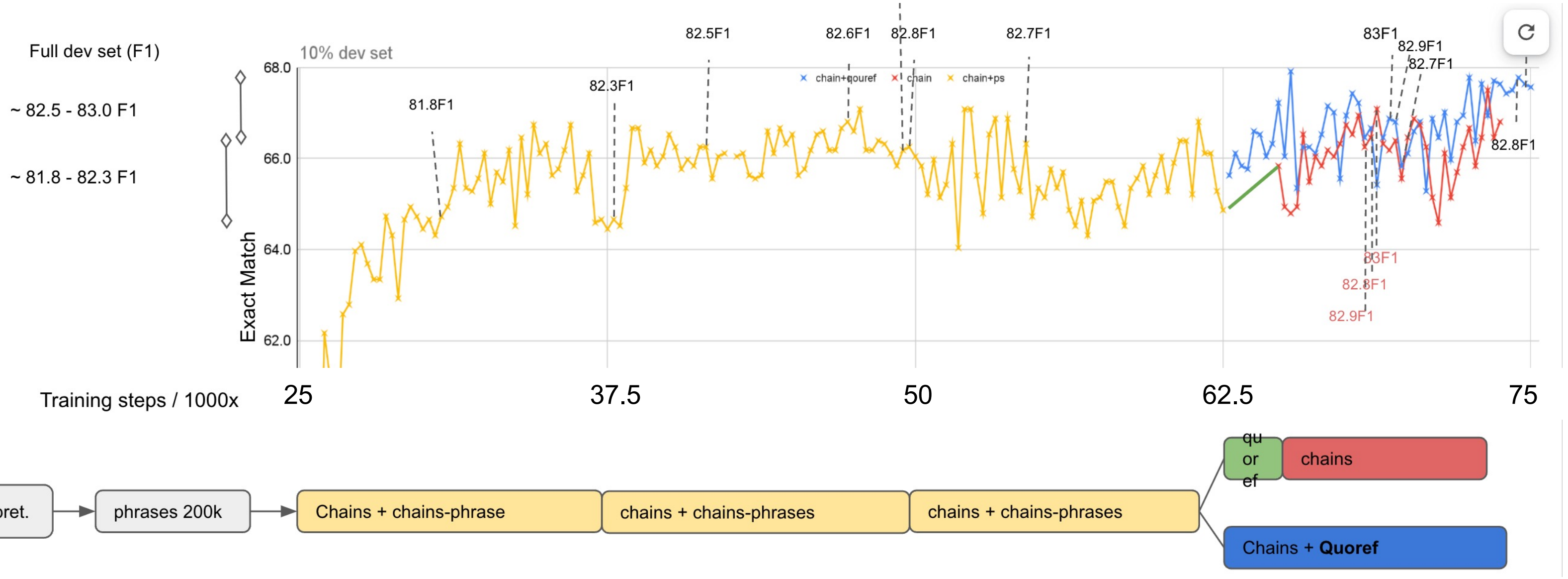
Simplified evaluation schema for feasability during training on 10% dev set; predition accuracy; ~ LAS (Labeled Accuracy Score)



Training steps of 128 samples

Does syntax and QA-dataset help when using mT5 XXL?

Google DeepMind

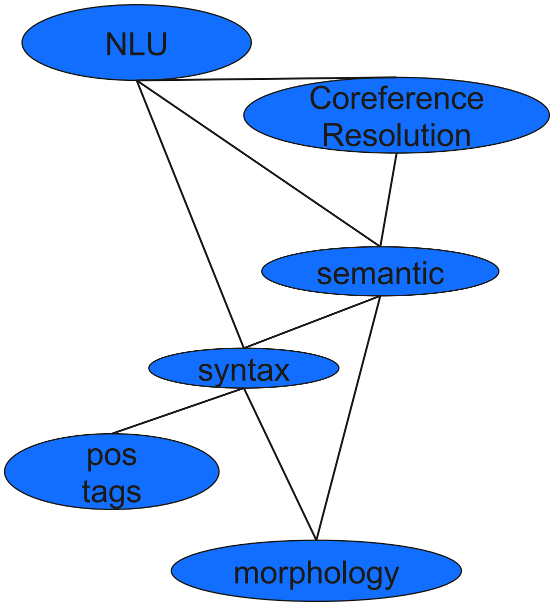


- Inspired by CorefQA (Wu et al. 2020), we pretrained with QuoRef and Squad
- We pretrained with phrase structures (NER as well)

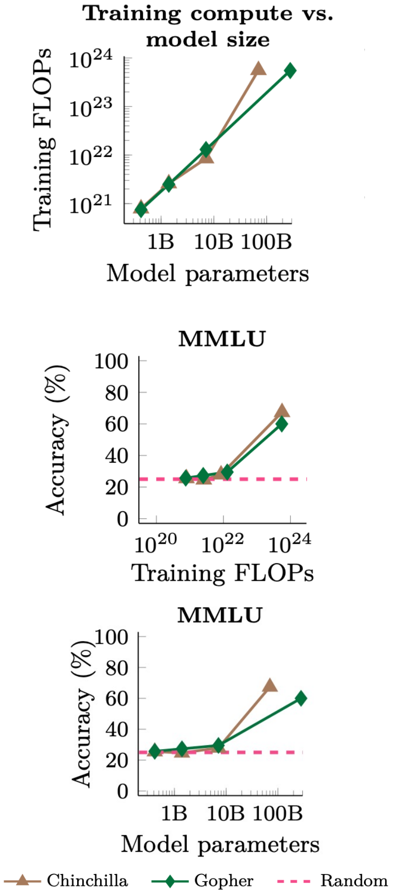
Unfortunatley, no statistically significant gains.
=> We did not use this pretraining schema finally

Emergent Properties of LMs

- Models need to learn skill before they can solve tasks (Wei et al. 2022)
- With larger models and more traingin skill emerge.



From Wei et al. 2022



Working Hypothesis

- The pretraining and joint training with
  phrases structures and **QA** might **not** have yield gains
  as the model had already this skill

- mT5-XXL might have already (partly) the skill to identify coreferences

- The training re-inforced this abilities and adapted to the specifica of the training data, e.g. mentitions as phrase-bounderies

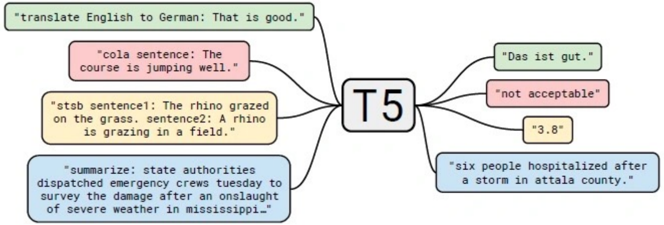
Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- Failures and Insights
- **Training Schema**
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- Trying to predict the Future

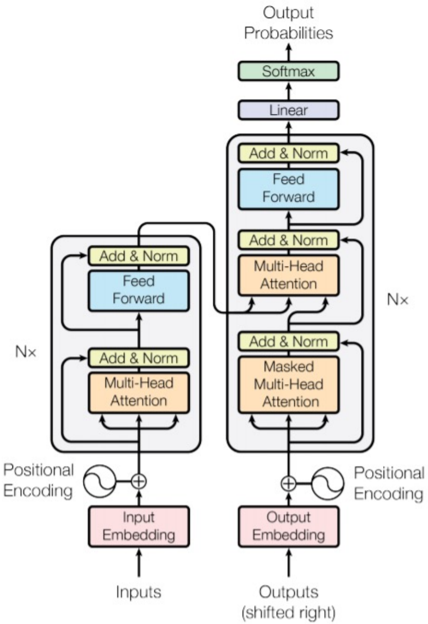
Training Schema

Google DeepMind

- We use multilingual T5 (Xue et al 2021) for training which is an encoder-decoder model (see graphic on the right bottom)



- mT5 is a **text-2-text** model as nowadays any LLM
- Training **input context size 2k** tokens and 384 output tokens
- ~ 100k steps, 128 TPUs-v4 @ 2-4 days depending on evaluation details
- We tried L (1.2b), XL (3b) and XXL (13B): 5 points gain by using one number larger model xl -> xxl



Final training schema adopted

- Training on coreference chains only:
  input: stepwise annotated chains with marked focus sentence
  target: transition for focus sentence

  => **no** phrase structure pretraining or QA data
  => **no** other extra data set

- Batch size 128; learning reate 0.001
- Input sequence size 2048 sentence piece tokens
- Target: 384
- Training for 100k steps

Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- Failures and Insights
- Training Schema
- **Ablation Study & Test Set Results**
- Multilingual Coreference Resolution
- Trying to predict the Future

Test set results

	LM	Decoder	MUC			B^3			$CEAF_{\Phi_4}$			Avg.
			P	R	F1	P	R	F1	P	R	F1	F1
English												
Lee et al. (2017)	-	neural e2e	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Lee et al. (2018)	Elmo	c2f	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
Joshi et al. (2019)	BERT	c2f	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
Yu et al. (2020)	BERT	Ranking	82.7	83.3	83.0	73.8	75.6	74.7	72.2	71.0	71.6	76.4
Joshi et al. (2020)	SpanBERT	c2f	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
Xia et al. (2020)	SpanBERT	transitions	85.7	84.8	85.3	78.1	77.5	77.8	76.3	74.1	75.2	79.4
Wu et al. (2020)	SpanBERT	QA	88.6	87.4	88.0	82.4	82.0	82.2	79.9	78.3	79.1	*83.1**
Xu and Choi (2020)	SpanBERT	hoi	85.9	85.5	85.7	79.0	78.9	79.0	76.7	75.2	75.9	80.2
Kirstain et al. (2021)	LongFormer	bilinear	86.5	85.1	85.8	80.3	77.9	79.1	76.8	75.4	76.1	80.3
Dobrovolskii (2021)	RoBERTa	c2f	84.9	87.9	86.3	77.4	82.6	79.9	76.1	77.1	76.6	81.0
Link-Append	mT5	transition	87.4	88.3	87.8	81.8	83.4	82.6	79.1	79.9	79.5	**83.3**
Arabic												
Aloraini et al. (2020)	AraBERT	c2f	63.2	70.9	66.8	57.1	66.3	61.3	61.6	65.5	63.5	63.9
Min (2021)	GigaBERT	c2f	73.6	61.8	67.2	70.7	55.9	62.5	66.1	62.0	64.0	64.6
Link-Append	mT5	transition	71.0	70.9	70.9	66.5	66.7	66.6	68.3	68.6	68.4	**68.7**
Chinese												
Xia and Durme (2021)	XLM-R	transition	-	-	-	-	-	-	-	-	-	69.0
Link-Append	mT5	transition	81.5	76.8	79.1	76.1	69.9	72.9	74.1	67.9	70.9	**74.3**

Ablation study

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

Many training steps and long contex
yields to better results.

Ablation study

Google DeepMind

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

Shorter context does not capture all coreferences links

Ablation study

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

Many training steps and long contex
yields to better results.

Ablation study

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

Context beyond the focus sentence is important

Ablation study

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

The English only model performce less well as the multilingual mT5 model

Ablation study

System	Ablation	F1
Link-Append	100k steps/3k pieces	83.2
Link-Append	2k sentence pieces	83.1
Link-Append	50k steps	82.9
Link-Append	no context beyond i	82.8
Link-Append	xxl-T5.1.1	82.7
Link-Append	xl-mT5	78.0
Mention-Link-Append	3k pieces	82.6
Mention-Link-Append	2k pieces	82.2
Link-only	link transitions only	81.4

Huge Accuracy loss for smaller model e.g. 11B vs 3b: -5 points

Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- Failures and Insights
- Training Schema
- Ablation Study & Test Set Results
- **Multilingual Coreference Resolution**
- Trying to predict the Future

Multilingual Coreference Resolution

- We used SemEval-2010 datasets for multilingual coreference resolution as well as
- **Without** finetuning, we see transfer to other language
  => We get high performance comparable with the winning systems of SemEval-2010
- Continuing training from English model, we obtain SoTA for all none English languages tested (Catalan, German, Arabic, Chinese, etc.)

	Training		Development		Test	
Language	docs	tokens	docs	tokens	docs	tokens
OntoNotes / CoNLL-2012 datasets						
English	1940	1.3M	343	160k	348	170k
Chinese	1729	750k	254	110k	218	90k
Arabic	359	300k	44	30k	44	30k
SemEval 2010 data						
Catalan	829	253k	142	42k	167	49k
Dutch	145	46k	23	9k	72	48k
German	900	331k	199	73k	136	50k
Italian	80	81k	18	16k	46	41k
Spanish	875	284k	140	44k	168	51k

Table 1: Sizes of the SemEval Shared Task data sets and OntoNotes (CoNLL-2012).

Results for Catalan and Dutch

Google DeepMind

Systems	Sing. P	Sing. E	# training docs./[illegible]	Avg.
Catalan				
Attardi et al. (2010)	Y	Y	all	48.2
Mention-Link-Append	Y	Y	all	**83.5**
Xia and Durme (2021)	N	Y	all	51.0
Mention-Link-Append	N	Y	all	**59.2**
Bitew et al. (2021)	N	N	∅/Translation	**48.0**
Link-Append	N	N	∅/Zero-shot	47.7
Link-Append	N	N	10/Few-shot	68.9
Dutch				
Kobdani and Schütze (2010)	Y	Y	all	19.1
Mention-Link-Append	Y	Y	all	**66.6**
Xia and Durme (2021)	N	Y	all	55.4
Mention-Link-Append	N	Y	all	**59.9**
Bitew et al. (2021)	N	N	∅/Translation	37.5
Link-Append	N	N	∅/Zero-shot	**57.6**
Link-Append	N	N	10/Few-shot	65.7

Evaluation in literature differs whether they include singletons (singeltons=mentions without coreference)

Results for Catalan and Dutch

Google DeepMind

Systems	Sing. P	Sing. E	# training docs./method	Avg. F1
Catalan				
Attardi et al. (2010)	Y	Y	all	48.2
Mention-Link-Append	Y	Y	all	**83.5**
Xia and Durme (2021)	N	Y	all	51.0
Mention-Link-Append	N	Y	all	**59.2**
Bitew et al. (2021)	N	N	∅/Translation	**48.0**
Link-Append	N	N	∅/Zero-shot	47.7
Link-Append	N	N	10/Few-shot	68.9
Dutch				
Kobdani and Schütze (2010)	Y	Y	all	19.1
Mention-Link-Append	Y	Y	all	**66.6**
Xia and Durme (2021)	N	Y	all	55.4
Mention-Link-Append	N	Y	all	**59.9**
Bitew et al. (2021)	N	N	∅/Translation	37.5
Link-Append	N	N	∅/Zero-shot	**57.6**
Link-Append	N	N	10/Few-shot	65.7

Same evaluation as SemEval2010: predicting single mentions and evaluating single mentions

Results for Catalan and Dutch

Systems	Sing. P	Sing. E	# training docs./method	Avg. F1
Catalan				
Attardi et al. (2010)	Y	Y	all	48.2
Mention-Link-Append	Y	Y	all	**83.5**
Xia and Durme (2021)	N	Y	all	51.0
Mention-Link-Append	N	Y	all	**59.2**
Bitew et al. (2021)	N	N	∅/Translation	**48.0**
Link-Append	N	N	∅/Zero-shot	47.7
Link-Append	N	N	10/Few-shot	68.9
Dutch				
Kobdani and Schütze (2010)	Y	Y	all	19.1
Mention-Link-Append	Y	Y	all	**66.6**
Xia and Durme (2021)	N	Y	all	55.4
Mention-Link-Append	N	Y	all	**59.9**
Bitew et al. (2021)	N	N	∅/Translation	37.5
Link-Append	N	N	∅/Zero-shot	**57.6**
Link-Append	N	N	10/Few-shot	65.7

Same evaluation as SemEval2010: predicting single mentions and evaluating single mentions

We see for instance for Catalan and Dutch (and other SemEval languages) far better performance

Results for Catalan and Dutch

Systems	Sing. P	Sing. E	# training docs./method	Avg. F1
Catalan				
Attardi et al. (2010)	Y	Y	all	48.2
Mention-Link-Append	Y	Y	all	**83.5**
Xia and Durme (2021)	N	Y	all	51.0
Mention-Link-Append	N	Y	all	**59.2**
Bitew et al. (2021)	N	N	∅/Translation	**48.0**
Link-Append	N	N	∅/Zero-shot	47.7
Link-Append	N	N	10/Few-shot	68.9
Dutch				
Kobdani and Schütze (2010)	Y	Y	all	19.1
Mention-Link-Append	Y	Y	all	**66.6**
Xia and Durme (2021)	N	Y	all	55.4
Mention-Link-Append	N	Y	all	**59.9**
Bitew et al. (2021)	N	N	∅/Translation	37.5
Link-Append	N	N	∅/Zero-shot	**57.6**
Link-Append	N	N	10/Few-shot	65.7

CoNLL-2012 setup: Predicting only Coreference Clusters.
=> much higher performance as well

We see for instance for Catalan and Dutch (and other SemEval languages) far better performance

Outline

- Why a text-to-text paradigm
- Transition-based Coreference Resolution
- Failures and Insights
- Training Schema
- Ablation Study & Test Set Results
- Multilingual Coreference Resolution
- **Trying to predict the Future**

How could the Future of Coreference Resolution look like

- LLM have even better language abilities such as strong reasoning and “understanding”
- The larger models are usually decoder only
- Models have already a notion of coreferences
- Few shot prompting yields results
- Mixin into LLMs

Few shot prompt to Palm 2

Google DeepMind

In **[1 the summer of 2005]** , a picture that people have long been looking forward to started emerging with frequency in various major **[2 Hong Kong]** media . With **their** unique charm , **these well - known cartoon images** once again caused **[2 Hong Kong]** to be a focus of worldwide attention . The **[3 world 's fifth] Disney park** will soon open to the public here . The most important thing about **Disney** is that **it** is a global brand . Well , for several years , although **it** was still under construction and , er , not yet open , it can be said that many people have viewed **[2 Hong Kong]** with new respect . Then welcome to the official writing ceremony of **[4 Hong Kong Disneyland]** . The construction of **[4 Hong Kong Disneyland]** began two years ago , in **[5 2003]** . In January of that year , the **[2 Hong Kong]** government turned over to **[6 Disney Corporation] 200 hectares of land at the [7 foot of Lantau Island]** that was obtained following the largest land reclamation project in recent years . One . Since then , **this area** has become …

=> Still substantial mistakes

With a little bit of Fine Tuning Palm 2

Google DeepMind

In the summer of 2005 , **[1 a picture that people have long been looking forward to]** started emerging with frequency in various major **[2 Hong Kong]** media . With **[1 their]** unique charm , **[1 these well - known cartoon images]** once again caused **[2 Hong Kong]** to be a focus of worldwide attention . **[3 The world 's fifth Disney park]** will soon open to the public here . The most important thing about **[4 Disney]** is that **[4 it]** is a global brand . Well , for several years , although **[3 it]** was still under construction and , er , not yet open , it can be said that many people have viewed **[2 Hong Kong]** with new respect . Then welcome to the official writing ceremony of **[3 Hong Kong Disneyland]** . The construction of **[3 Hong Kong Disneyland]** began two years ago , in **[5 2003]** . In January of **[5 that year]** , the **[2 Hong Kong]** government turned over to **[4 Disney Corporation]** **[6 200 hectares of land at the foot of [7 Lantau Island] that was obtained following the largest land reclamation project in recent years]** . One . Since then , **[6 this area]**

Conclusions

- **Simplicity**: We use greedy seq2seq prediction without a separate mention detection step and do not employ a higher order decoder to identify links.
- **Accuracy**: The accuracy of the method exceeds the previous state of the art.
- **Text-to-text (seq2seq) based**: the method can make direct use of modern generation models that employ the generation of text strings as the key primitive.