Structured Prediction as Translation between Augmented Natural Languages

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto

Amazon Web Services

Presented by Björn Bebensee (2019-21343)

Department of Computer Science and Engineering, Seoul National University

What is structured prediction in NLP?

Instead of predicting some continuous value, the model's output space is structured.

Unlike in multiclass classification: size of the output space is often exponential, thus difficult.

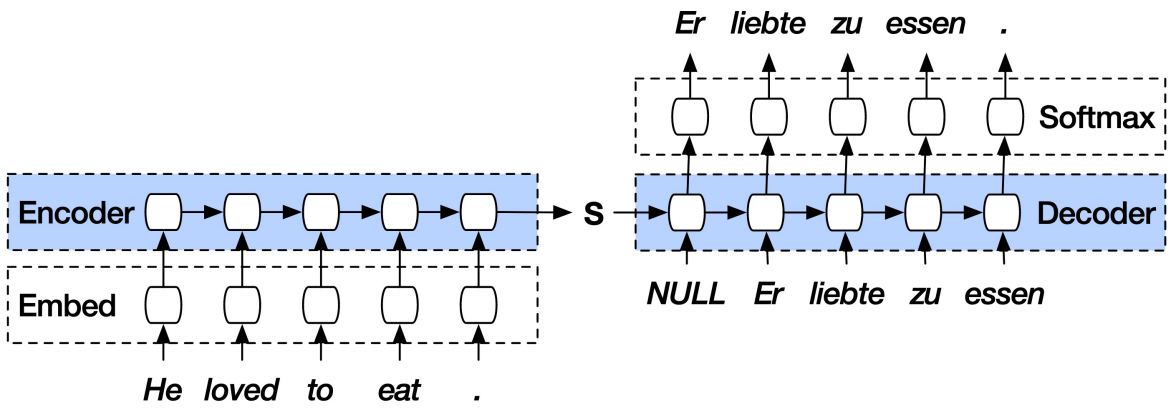


Image: Stephen Merity

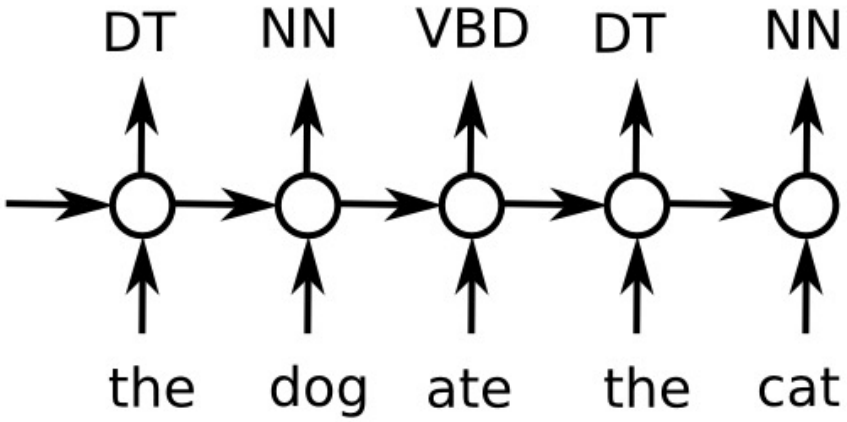


Image: Mat Kelcey, stackoverflow.com

Agenda

1. Brief overview of tasks
2. Motivation and related work
3. Proposed method
    a. Key idea
    b. Joint entity and relation extraction
    c. Named entity recognition
    d. Relation classification
    e. Semantic role labeling
    f. Coreference resolution
1. Experimental results

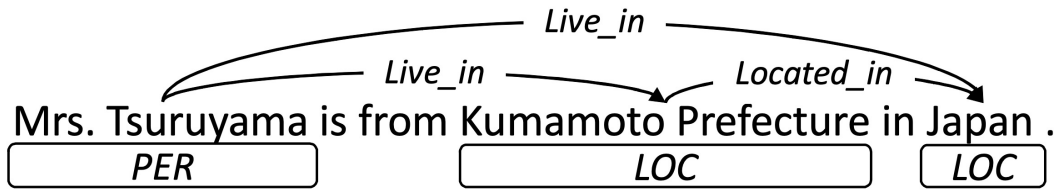
Structured prediction tasks

Let's have a look at some structured prediction tasks in NLP!

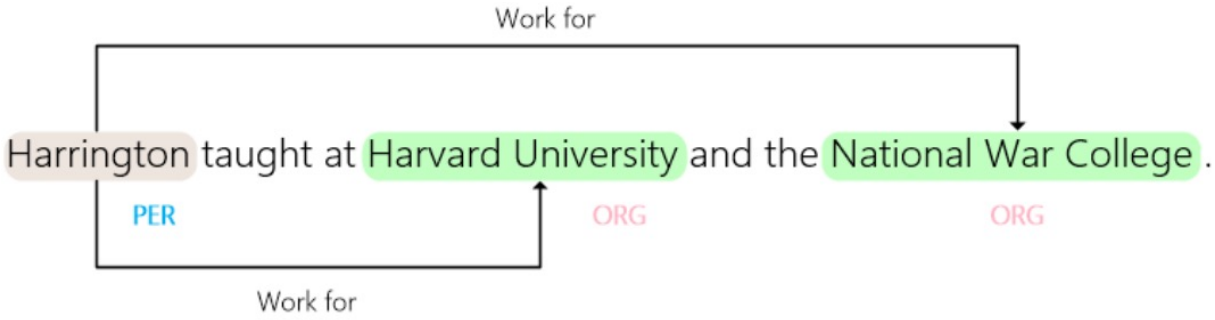
Joint entity and relation extraction

Given: one or multiple sentences

Output: entities, their categories, and how they are related



Miwa and Sasaki, EMNLP 2014.



Zhang et al., Applied Sciences 2021.

Named entity recognition

Given: one or multiple sentences

Output: predict which words are entities and their categories

A special case of joint entity and relation extraction w/o relations

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

Image: Huggingface

Semantic role labeling

Given: a sentence, a predicate

Output: how each noun, adverb relates to the predicate, i.e. answer the question "who did what to whom?"

Coreference resolution

Given: a text, typically consisting of multiple sentences

Output: identify all expressions referring to the same entity (i.e. “her”, “his”, etc.)

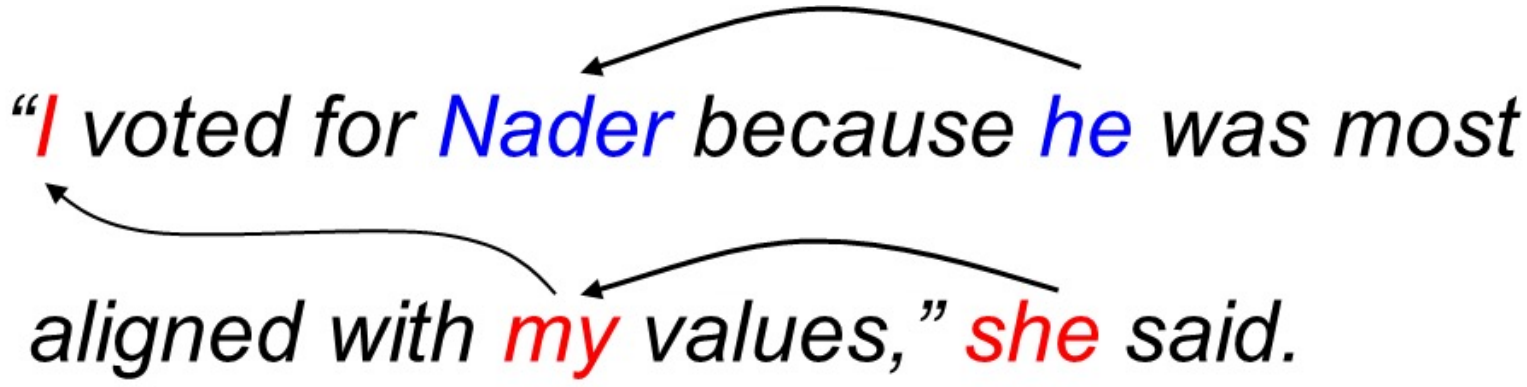


Image: Stanford Natural Language Processing Group

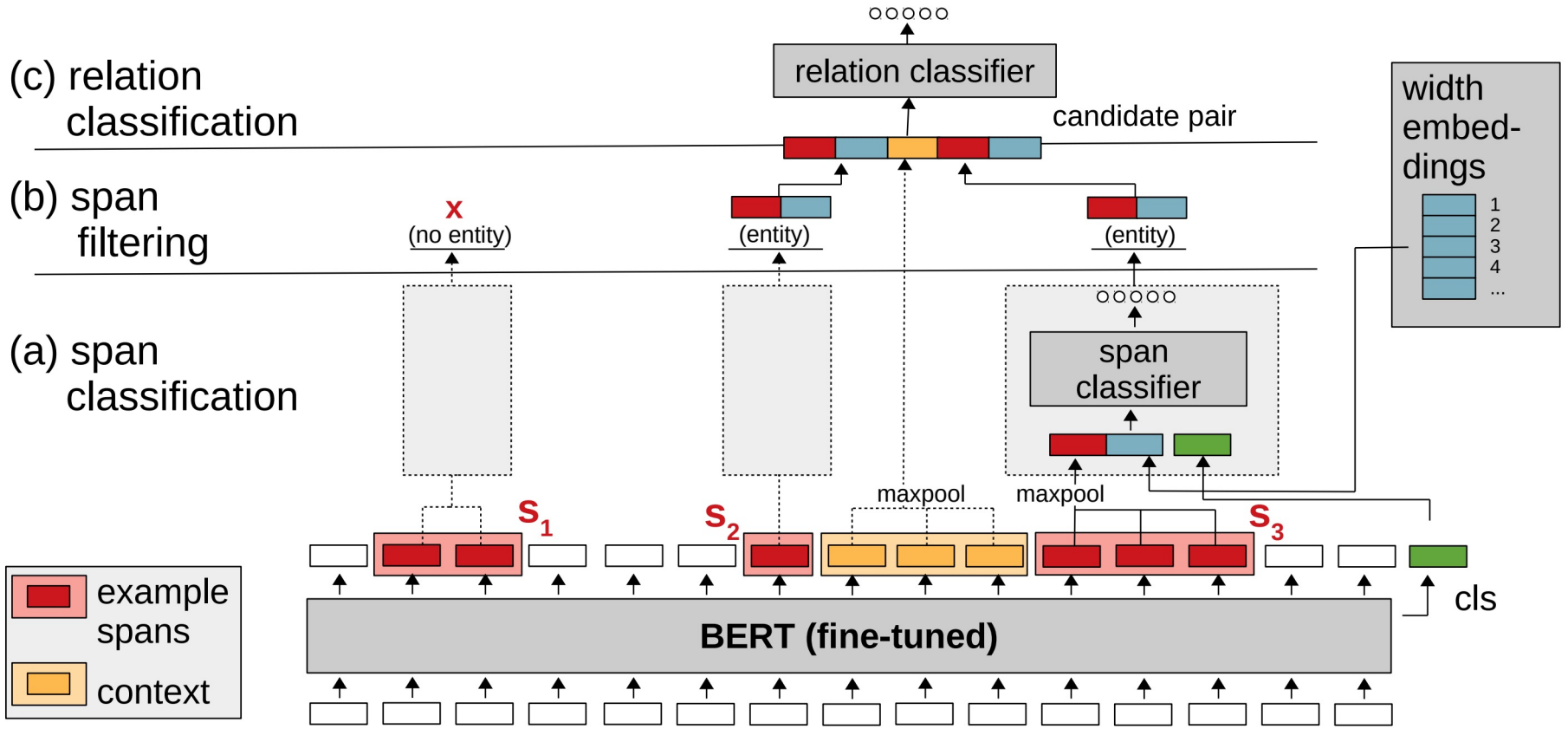
Motivation

Recent work: mostly task-specific approaches on top of features extractors like BERT.

Disadvantages:

- **Does not take advantage of semantics of class labels: the classifier does not know that an "author" is also a "person"**
- **Task-specific architectures needed**

Related work: joint entity and relation extraction



Athiwaratkun et al., arXiv 2020

Key idea

Augmented natural language: annotate language using special symbols, i.e. "[…]" for entities and "[… | category]" to categorize them

((AddToPlaylist)) Add [Kent James | artist] to the [Disney | playlist] soundtrack.

Athiwaratkun et al., arXiv 2020

Frame structured prediction task as text-to-text translation:

- **Input: source text, task**
- **Output: source text annotated with the answers, i.e. all entity categories and their relations**
- **Answers are easy to extract from bracket annotations**

Key idea

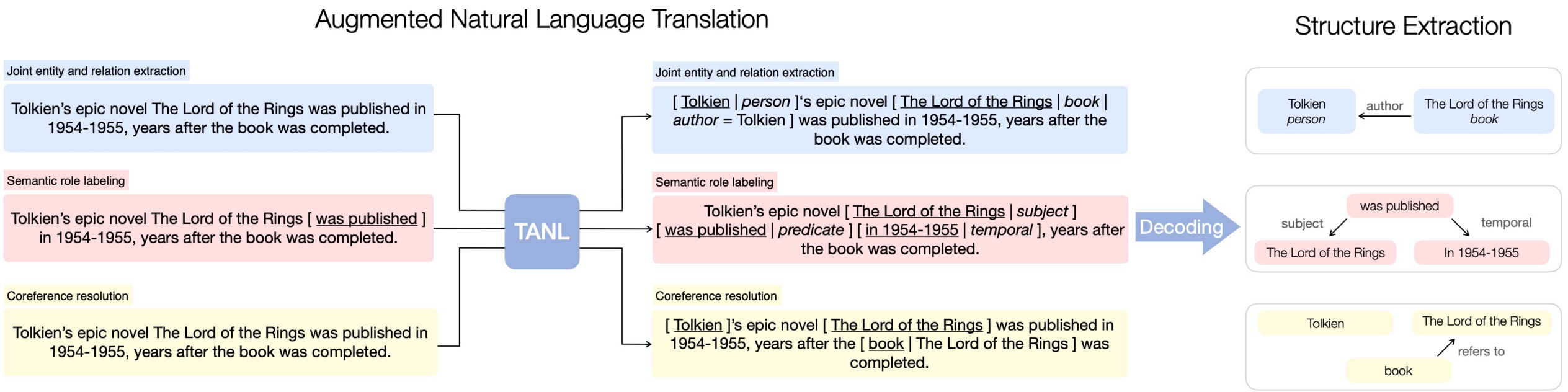
Advantages of this approach:

- Only requires a single architecture for translation
- Translation model understands and can take advantage of label semantics, i.e. a “person” is more likely to be “author” than a non-person

It’s possible to train a single multi-task model that can complete all tasks.

Knowledge is transferred from one task and dataset to the other, benefits low-resource tasks. ☺

Proposed method



Proposed method

Use Google's pre-trained T5-base model for translation (but: any generative pre-trained model can be used).

Single-task model: fine-tune to translate between source sentence and augmented natural language that contains answers.

Multi-task model: simultaneously train on all tasks, provide name of task and dataset so that model knows what to predict as prefix.

Proposed method: examples

Joint entity and relation extraction:

Input: Tolkien's epic novel The Lord of the Rings was published in 1954-1955, years after the book was completed.

Output: [Tolkien | *person*]'s epic novel [The Lord of the Rings | *book* | *author* = Tolkien] was published in 1954-1955, years after the book was completed.

Relation classification:

Input: Born in Bologna, Orlandi was a student of the famous Italian [soprano] and voice teacher [Carmen Melis] in Milan. The relationship between [Carmen Melis] and [soprano] is

Output: relationship between [Carmen Melis] and [soprano] = *voice type*

Proposed method: examples

Semantic role labeling:

Input: The luxury auto maker last year **[** sold **]** 1,214 cars in the U.S.

Output: **[** The luxury auto maker **|** *subject* **] [** last year **|** *temporal* **]** sold **[** 1,214 cars **|** *object* **] [** in the U.S. **|** *location* **]**

Coreference resolution:

Input: Barack Obama nominated Hillary Rodham Clinton as his secretary of state on Monday. He chose her because she had foreign affairs experience as a former First Lady.

Output: **[** Barack Obama **]** nominated **[** Hillary Rodham Clinton **]** as **[** his **|** *Barack Obama* **] [** secretary of state **|** *Hillary Rodham Clinton* **]** on Monday. **[** He **|** *Barack Obama* **]** chose **[** her **|** *Hillary Rodham Clinton* **]** because **[** she **|** *Hillary Rodham Clinton* **]** had foreign affairs experience as a former **[** First Lady **|** *Hillary Rodham Clinton* **]**.

Proposed method

Decoding procedure:

1. Extract entity types and relations (without special tokens)
2. Align input and output sentence on the token level (improves robustness).
3. For each relation in the output find the closest entity which exactly matches the tail entity, discard otherwise.
4. Discard entity and relations which do not belong to the predefined types.

Generated output: Six days after starting [Aciclovir | *drug*] she exhibited signs of [[lithium | *drug*] toxicity | *disease* | *effect* = Aciclovir | *effect* = lithium].

Cleaned output: Six days after starting Aciclovir she exhibited signs of lithium toxicity .

Experimental results

Entity Relation Extr.

	CoNLL04		ADE*		NYT		ACE2005	
	Entity	Rel.	Entity	Rel.	Entity	Rel.	Entity	Rel.
SpERT (Eberts & Ulges, 2019)	88.9	71.5	89.3	78.8				
DyGIE (Luan et al., 2019)							88.4	63.2
MRC4ERE (Zhao et al., 2020)	88.9	71.9					85.5	62.1
RSAN (Yuan et al., 2020)						84.6		
TANL	89.4	71.4	90.2	80.6	**94.9**	**90.8**	**88.9**	**63.7**
TANL (multi-dataset)	89.8	**72.6**	90.0	80.0	94.7	90.5	88.2	62.5
TANL (multi-task)	**90.3**	70.0	**91.2**	**83.8**	94.7	90.7		

NER

	CoNLL03	OntoNotes	GENIA*	ACE2005*
BERT-MRC (Li et al., 2019a)	93.0	91.1	**83.8**	**86.9**
BERT-MRC+DSC (Li et al., 2019b)	93.3	**92.1**		
Cloze-CNN (Baevski et al., 2019)	**93.5**			
GSL (Athiwaratkun et al., 2020)	90.7	90.2		
TANL	91.7	89.8	76.4	84.9
TANL (multi-dataset)	92.0	89.8	75.9	84.4
TANL (multi-task)	91.7	89.4	76.4	

Experimental results

Relation Class.

	TACRED	FewRel 1.0 (validation) 5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
BERT-EM (Soares et al., 2019)	70.1	88.9		82.8	
BERT_{EM}+MTB (Soares et al., 2019)	71.5	90.1		**83.4**	
DG-SpanBERT (Chen et al., 2020)	71.5				
BERT-PAIR (Gao et al., 2019)		85.7	89.5	76.8	81.8
TANL	**71.9**	**94.0 ± 4.1**	**96.4 ± 4.2**	82.6 ± 4.5	**88.2 ± 5.9**
TANL (multi-task)	69.1				

SRL

	CoNLL05 WSJ	CoNLL05 Brown	CoNLL2012
Dep and Span (Li et al., 2019d)	86.3	76.4	83.1
BERT SRL (Shi & Lin, 2019)	88.8	82.0	86.5
TANL	89.3	82.0	**87.7**
TANL (multi-dataset)	**89.4**	**84.3**	87.6
TANL (multi-task)	89.1	84.1	87.7

Experimental results

Event Extr.

	ACE2005			
	Trigger Id.	Trigger Cl.	Argument Id.	Argument Cl.
J3EE (Nguyen & Nguyen, 2019)	72.5	**69.8**	**59.9**	52.1
DyGIE++ (Wadden et al., 2019)		69.7	55.4	**52.5**
TANL	**72.9**	68.4	50.1	47.6
TANL (multi-task)	71.8	68.5	48.5	48.5

Coreference Res.

	CoNLL-2012* (BERT-base ¦ BERT-large)							
	MUC		B^3		$CEAF_{\phi_4}$		Avg. F1	
Higher-order c2f-coref (Lee et al., 2018)	80.4		70.8		67.6		73.0	
SpanBERT (Joshi et al., 2020)		85.3		78.1		75.3		79.6
BERT+c2f-coref (Joshi et al., 2019)	81.4	83.5	71.7	75.3	68.8	71.9	73.9	76.9
CorefQA+SpanBERT (Wu et al., 2020)	**86.3**	**88.0**	**77.6**	**82.2**	**75.8**	**79.1**	**79.9**	**83.1**
TANL	81.0		69.0		68.4		72.8	
TANL (multi-task)	78.7		65.7		63.8		69.4	

DST

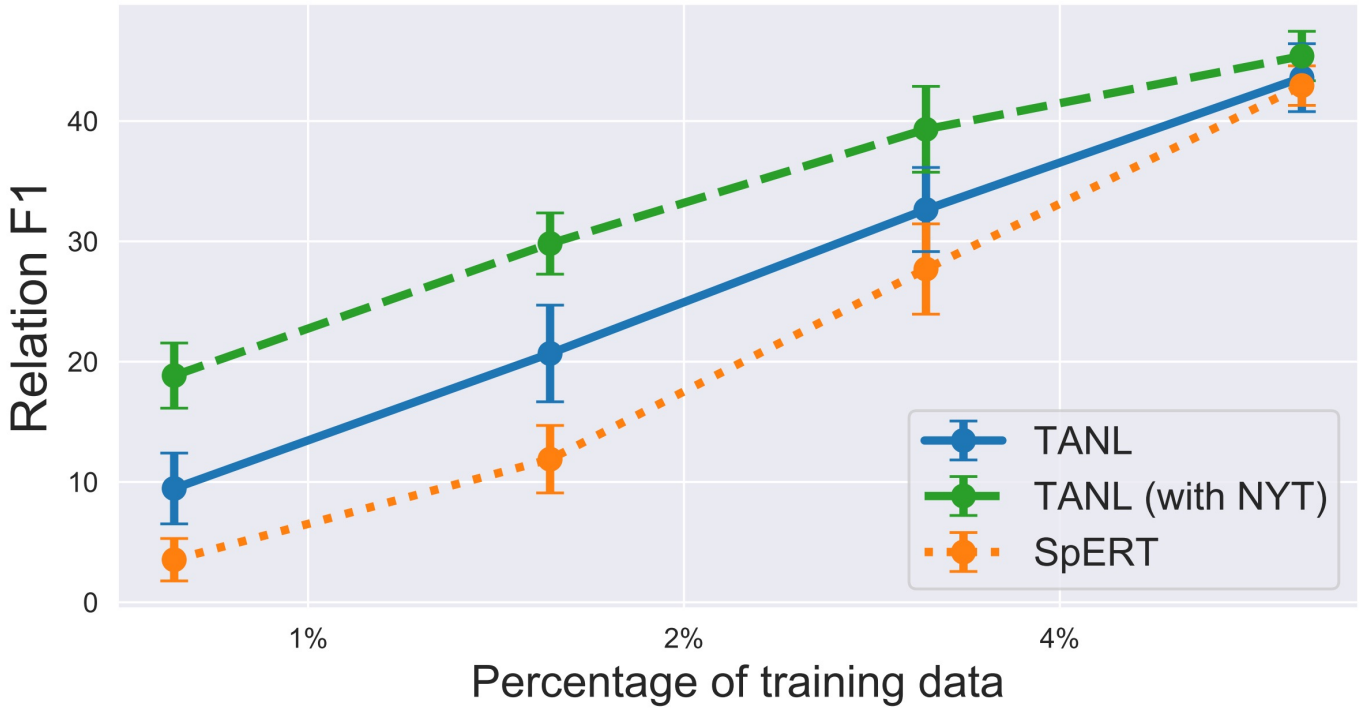
	MultiWOZ 2.1 (Joint Accuracy)
TRADE (Wu et al., 2019)	45.6
SimpleTOD (Hosseini-Asl et al., 2020)	**55.7**
TANL	50.5
TANL (multi-task)	51.4

Experimental results

Observations:

- **Across the board: TANL state-of-the art or competitive**
- **Multi-task model is competitive, significant improvements for low-resource datasets**
- **Poorer performance of multi-task on coreference resolution might be due to much longer texts (compared to other tasks)**
- **Good performance in few-shot classification**

Experimental results



(a) Low-resource scenarios

Discussion

Drawback of the generative approach as implemented here:

Time complexity of token generation is $\mathcal{O}(L^2)$ where L is the sequence length.

However: Reformer (Kitaev et al., 2020) and Linformer (Wang et al. 2020) reduce this to $\mathcal{O}(L \log L)$ and $\mathcal{O}(L)$ respectively.

Conclusion

To summarize:

- Authors propose a text-to-text translation framework which unifies many structured prediction tasks.
- New state-of-the-art for some tasks, competitive results on all tasks.
- Especially low-resource datasets can benefit from multi-task model.

Additional references (and figures)

Athiwaratkun, Ben, et al. "Augmented Natural Language for Generative Sequence Labeling." *arXiv preprint arXiv:2009.13272* (2020).

Eberts, Markus, and Adrian Ulges. "Span-based joint entity and relation extraction with transformer pre-training." *24th European Conference on Artificial Intelligence (ECAI)*. 2020.

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. "Reformer: The efficient transformer." *arXiv preprint arXiv:2001.04451* (2020).

Lee, Joohong, Sangwoo Seo, and Yong Suk Choi. "Semantic relation classification via bidirectional LSTM networks with entity-aware attention using latent entity typing." *Symmetry* 11.6 (2019): 785.

Miwa, Makoto, and Yutaka Sasaki. "Modeling joint entity and relation extraction with table representation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

Wang, Sinong, et al. "Linformer: Self-attention with linear complexity." *arXiv preprint arXiv:2006.04768* (2020).

Zhang, Haiyang, Guanqun Zhang, and Yue Ma. "Syntax-Informed Self-Attention Network for Span-Based Joint Entity and Relation Extraction." *Applied Sciences* 11.4 (2021): 1480.