Label Verbalization and Entailment for Effective Zero- and Few-Shot Relation Extraction

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena and Eneko Agirre

EMNLP 2021

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Relation Extraction task

Given 2 entities e1 and e2 and a context c, predict the semantic relation (if any) holding between the two entities in the context.

⟨**Billy Mays**$_{\text{PERSON}}$, **Tampa**$_{\text{CITY}}$⟩

Billy Mays, the bearded, boisterous pitchman who, as the undisputed king of TV yell and sell, became an unlikely pop culture icon, died at his home in **Tampa**, Fla, on Sunday.

→ `per:city_of_death`

Current state of the art

- Mostly approached via **supervised learning** on large datasets or via **distant-supervision** when a Knowledge Base and a large set of documents are available.

- Supervised learning:
  - Pretrained language models (LM) fine-tuned on large amount of labeled data.
  - Focused on **models**: finding better pre-training objectives, relation representations or incorporating external knowledge.

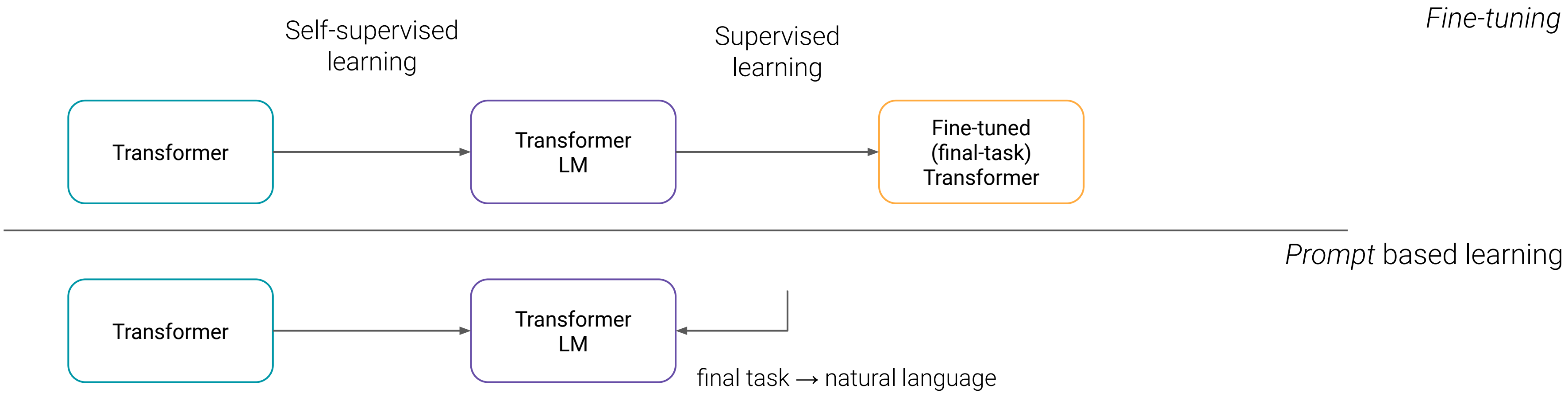
- Distant-supervision:
  - Pretrained language models (LM) fine-tuned on noisy large amounts of labeled data.
  - Focused on **data**: alleviating the noisy signal from the data and finding better bag of context representations.

- How about focusing on <u>a model that works with a small amount of data?</u>

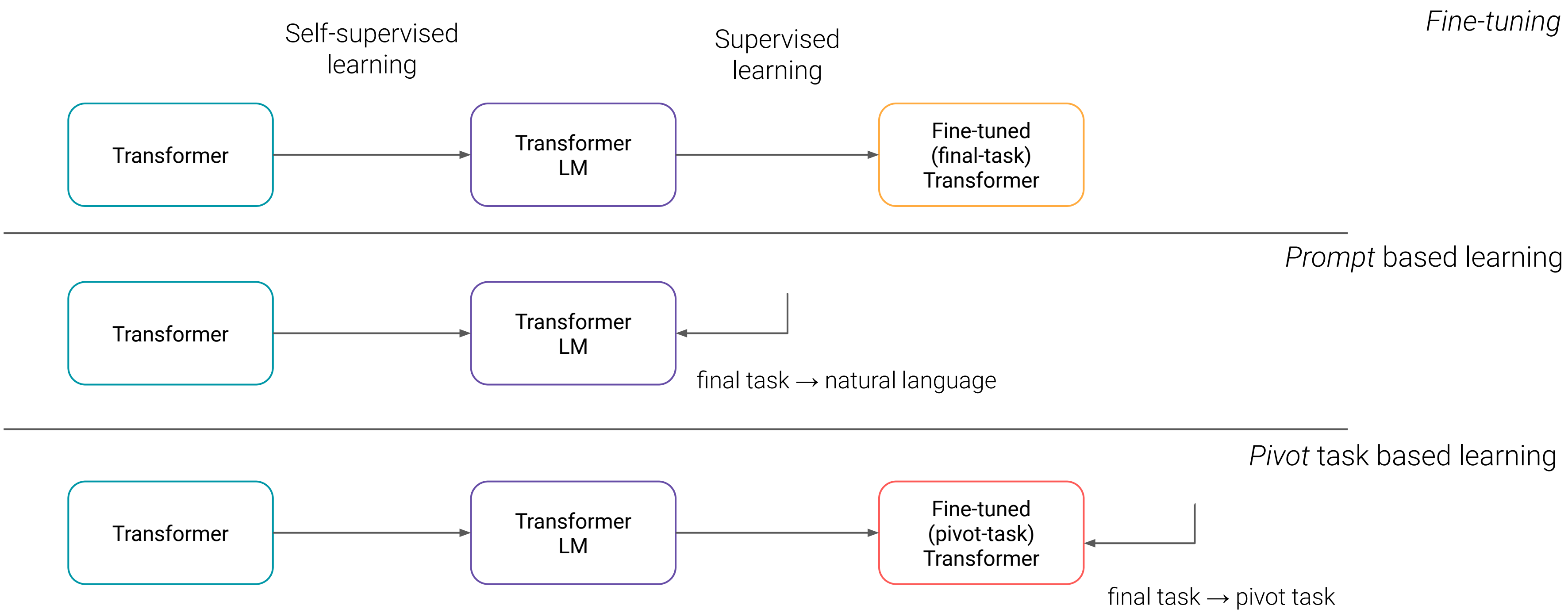
Alternative paradigms to fine-tuning

Fine-tuning

Alternative paradigms to fine-tuning

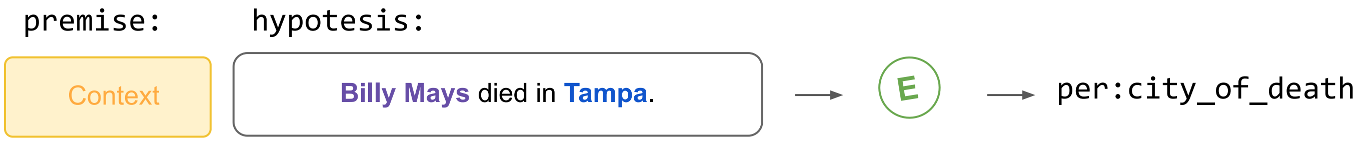


Alternative paradigms to standard fine-tuning



Entailment-based Relation Extraction

- We propose to reformulate Relation Extraction as a Textual Entailment (similar to Obamuyide and Vlachos (2018)[1]).

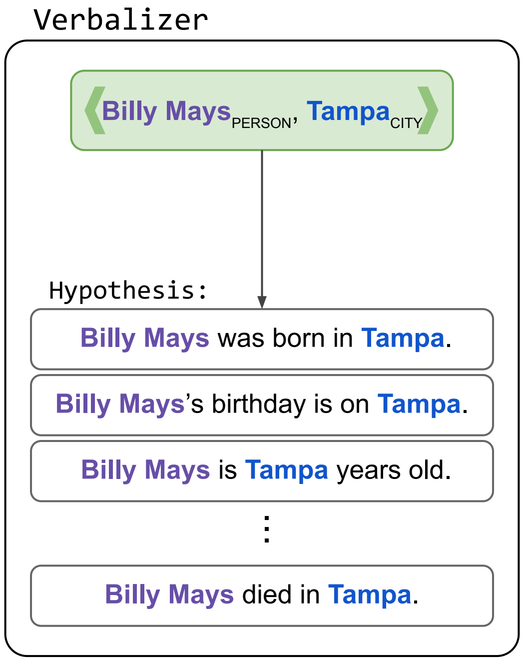


- We found Textual Entailment (aka **NLI**) to be a robust **pivot task** for zero- and few-shot learning.
- We propose a simple yet effective **inference strategy** based on NLI pretrained models to achieve competent results even with no training examples.

[1]Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Approach

Entailment based Relation Extraction



- Function that combines entity pairs with **templates** to generate textual hypotheses for relations:

$$hyp = \text{VERBALIZE}(t, x_{e1}, x_{e2})$$

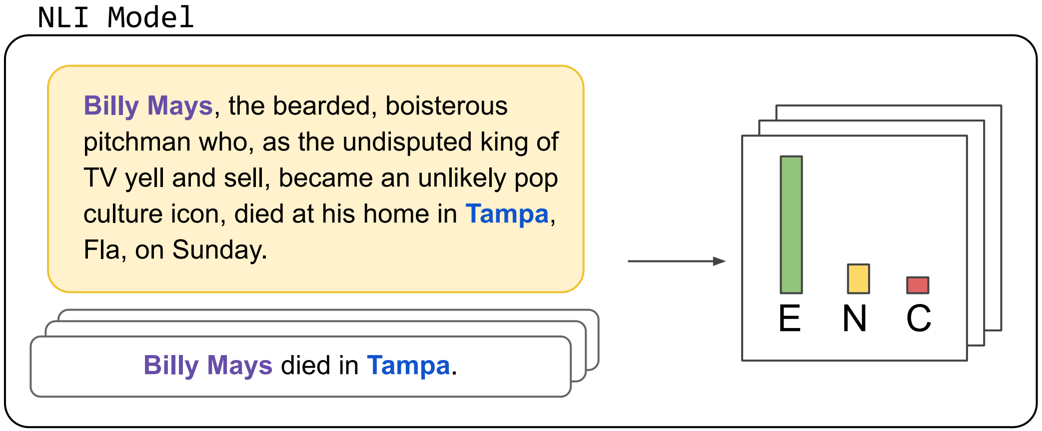
- N:M relation between templates and relations
- Also, type constraints for entities

Relation	Templates	Valid argument types
per:alternate_names	{subj} is also known as {obj}	PERSON, MISC
per:date_of_birth	{subj}'s birthday is on {obj}	DATE
	{subj} was born on {obj}	
per:age	{subj} is {obj} years old	NUMBER, DURATION
per:country_of_birth	{subj} was born in {obj}	COUNTRY
per:stateorprovince_of_birth	{subj} was born in {obj}	STATE_OR_PROVINCE
per:city_of_birth	{subj} was born in {obj}	CITY, LOCATION

Entailment based Relation Extraction

- Next, we compute the entailment probabilities for each of the hypothesis independently.

$$\mathrm{P}_{NLI}(x, hyp)$$



Entailment based Relation Extraction

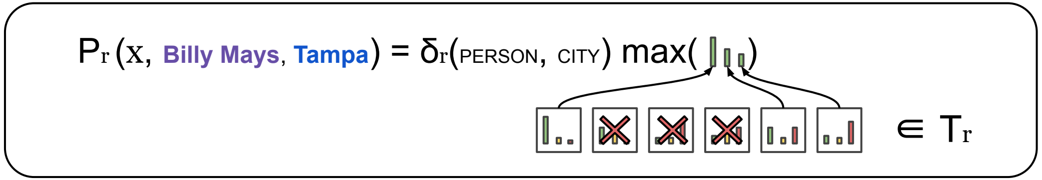
- We compute the probability of relation r based on the hypothesis probabilities and entity constraints:

$$\mathsf{P}_r(x, x_{e1}, x_{e2}) = \delta_r(e_1, e_2) \max_{t \in T_r} \mathsf{P}_{NLI}(x, hyp)$$

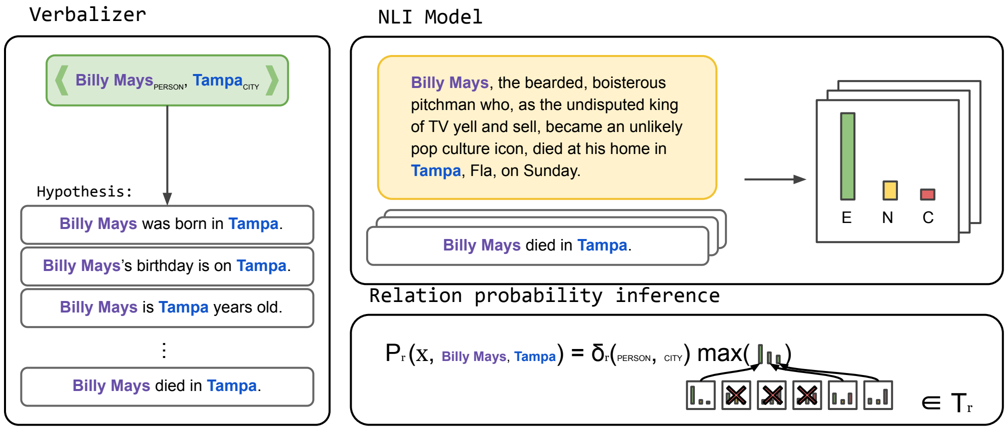
- The δ_r function describes the entity constraints of the relation r:

$$\delta_r(e_1, e_2) = \begin{cases} 1 & e_1 \in E_{r1} \land e_2 \in E_{r2} \\ 0 & \text{otherwise} \end{cases}$$

Relation probability inference



Entailment based Relation Extraction



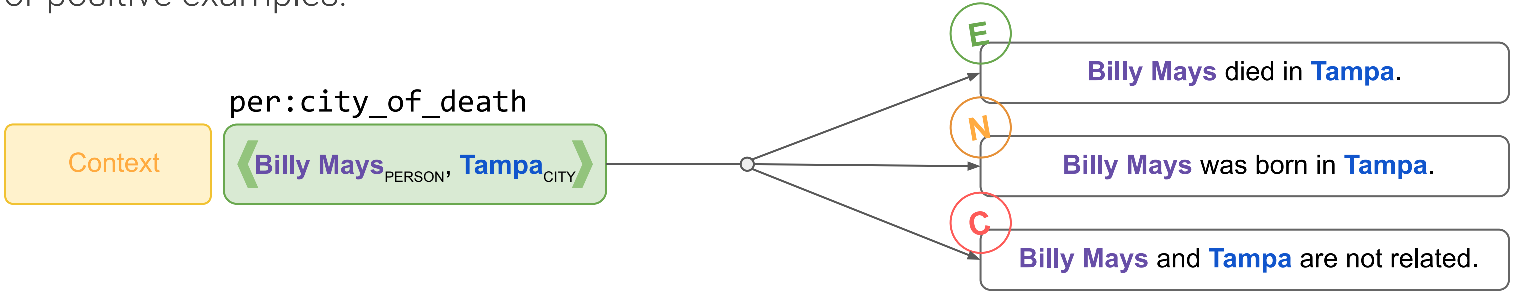
Finally, we return the relation with the higher probability:

$$\hat{r} = \arg\max_{r \in R} \mathrm{P}_r(x, x_{e1}, x_{e2})$$

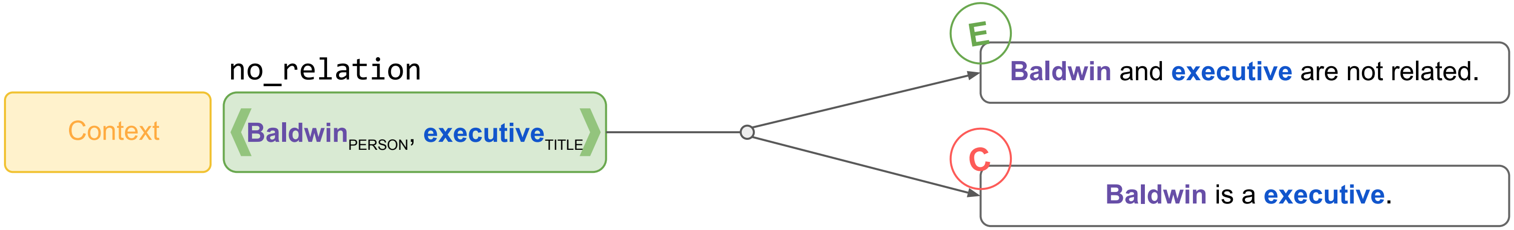
If no relation has higher entailment prob. than threshold (T, default 0.5), r = `no_relation`.

Fine-tuning NLI on Relation Extraction data

For positive examples:



For negative examples:



Results

Zero-Shot

NLI Model	# Param.	MNLI Acc.
$\text{ALBERT}_{\text{xxLarge}}$	223M	90.8
RoBERTa	355M	90.2
BART	406M	89.9
$\text{DeBERTa}_{\text{xLarge}}$	900M	91.7
$\text{DeBERTa}_{\text{xxLarge}}$	1.5B	91.7

Zero-Shot relation extraction:

Zero-Shot

NLI Model	# Param.	MNLI Acc.	No Dev ($\mathcal{T} = 0.5$) Pr.	Rec.	F1
$\text{ALBERT}_{\text{xxLarge}}$	223M	90.8	32.6	**79.5**	46.2
RoBERTa	355M	90.2	32.8	75.5	45.7
BART	406M	89.9	39.0	63.1	48.2
$\text{DeBERTa}_{\text{xLarge}}$	900M	91.7	40.3	77.7	53.0
$\text{DeBERTa}_{\text{xxLarge}}$	1.5B	91.7	**46.6**	76.1	**57.8**

Zero-Shot relation extraction:

- Default threshold for `no_relation` produces low precision

Zero-Shot

NLI Model	# Param.	MNLI Acc.	No Dev ($\mathcal{T} = 0.5$) Pr.	Rec.	F1	1% Dev Pr.	Rec.	F1
$\text{ALBERT}_{\text{xxLarge}}$	223M	90.8	32.6	**79.5**	46.2	55.2	58.1	56.6 $\pm$1.4
RoBERTa	355M	90.2	32.8	75.5	45.7	58.5	53.1	55.6 $\pm$1.3
BART	406M	89.9	39.0	63.1	48.2	60.7	46.0	52.3 $\pm$1.8
$\text{DeBERTa}_{\text{xLarge}}$	900M	91.7	40.3	77.7	53.0	**66.3**	59.7	**62.8** $\pm$1.7
$\text{DeBERTa}_{\text{xxLarge}}$	1.5B	91.7	**46.6**	76.1	**57.8**	63.2	**59.8**	61.4 $\pm$1.0

Zero-Shot relation extraction:

- Default threshold for `no_relation` produces low precision
- With 1% of Dev (2 examples per relation, 100 negative examples) threshold can be tuned for each relation, yielding better results

Zero-Shot

NLI Model	# Param.	MNLI Acc.	No Dev ($\mathcal{T} = 0.5$) Pr.	Rec.	F1	1% Dev Pr.	Rec.	F1
$\text{ALBERT}_{\text{xxLarge}}$	223M	90.8	32.6	**79.5**	46.2	55.2	58.1	56.6 $\pm$1.4
RoBERTa	355M	90.2	32.8	75.5	45.7	58.5	53.1	55.6 $\pm$1.3
BART	406M	89.9	39.0	63.1	48.2	60.7	46.0	52.3 $\pm$1.8
$\text{DeBERTa}_{\text{xLarge}}$	900M	91.7	40.3	77.7	53.0	**66.3**	59.7	**62.8** $\pm$1.7
$\text{DeBERTa}_{\text{xxLarge}}$	1.5B	91.7	**46.6**	76.1	**57.8**	63.2	**59.8**	61.4 $\pm$1.0

Zero-Shot relation extraction:

- Default threshold for `no_relation` produces low precision
- With 1% of Dev (2 examples per relation, 100 negative examples) threshold can be tuned for each relation, yielding better results
- DeBERTa achieves the best results, maybe due to the number of parameters.
- Note that minor variations in MNLI (±2) produce large variations in F1.

Few-Shot

Model	1%			5%			10%		
	Pr.	Rec.	F1	Pr.	Rec.	F1	Prec.	Rec.	F1
SpanBERT	0.0	0.0	0.0 ±0.0	36.3	23.9	28.8 ±13.5	3.2	1.1	1.6 ±20.7
RoBERTa	56.8	4.1	7.7 ±3.6	52.8	34.6	41.8 ±3.3	61.0	50.3	55.1 ±0.8
K-Adapter	73.8	7.6	13.8 ±3.4	56.4	37.6	45.1 ±0.1	62.3	50.9	56.0 ±1.3
LUKE	61.5	9.9	17.0 ±5.9	57.1	47.0	51.6 ±0.4	60.6	60.6	60.6 ±0.4

Few-Shot relation extraction:

- State of the art systems have difficulties to learn the task where very small amount of data is annotated. Indeed, SpanBERT does not even work.

Few-Shot

Model	1% Pr.	1% Rec.	1% F1	5% Pr.	5% Rec.	5% F1	10% Prec.	10% Rec.	10% F1
SpanBERT	0.0	0.0	0.0 ±0.0	36.3	23.9	28.8 ±13.5	3.2	1.1	1.6 ±20.7
RoBERTa	56.8	4.1	7.7 ±3.6	52.8	34.6	41.8 ±3.3	61.0	50.3	55.1 ±0.8
K-Adapter	73.8	7.6	13.8 ±3.4	56.4	37.6	45.1 ±0.1	62.3	50.9	56.0 ±1.3
LUKE	61.5	9.9	17.0 ±5.9	57.1	47.0	51.6 ±0.4	60.6	60.6	60.6 ±0.4
$\text{NLI}_{\text{RoBERTa}}$ (ours)	56.6	55.6	56.1 ±0.0	60.4	68.3	64.1 ±0.2	**65.8**	69.9	67.8 ±0.2
$\text{NLI}_{\text{DeBERTa}}$ (ours)	**59.5**	**68.5**	**63.7** ±0.0	**64.1**	**74.8**	**69.0** ±0.2	62.4	**74.4**	**67.9** ±0.5

Few-Shot relation extraction:

- State of the art systems have difficulties to learn the task where very small amount of data is annotated. Indeed, SpanBERT does not even work.
- NLI systems instead, achieve very good results from the beginning, and, as the rest do, the results improve with training data.
- As in the zero-shot setting, DeBERTa model score the best.

Full training

Full trained relation extraction

- NLI sistems perform in pair when large amount of annotated data is available (RoBERTa vs $NLI_{RoBERTa}$).
- The performance gap between NLI systems is maintained even after fine-tuned with the whole dataset ($NLI_{RoBERTa}$ vs $NLI_{DeBERTa}$).
- We outperformed the state of the art with $NLI_{DeBERTa}$. But it is true that similar performance is expected using a vanilla DeBERTa trained on whole TACRED (Zhang et al. 2017)[3].

Model	Pr.	Rec.	F1
SpanBERT	70.8	70.9	70.8
RoBERTa	70.2	72.4	71.3
K-Adapter	70.1	74.0	72.0
LUKE	70.4	75.1	72.7
$NLI_{RoBERTa}$ (ours)	71.6	70.4	71.0
$NLI_{DeBERTa}$ (ours)	**72.5**	**75.3**	**73.9**

[3] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), pages 35–45.

Conclusions and Future Work

Conclusions and Future Work

Conclusions:

- Prompt or pivot-task based methods are effective for zero- or few-shot relation-extraction.
- Specially NLI based relation extraction yields very good results.
- NLI based methods are robust discriminating positive relations, but have difficulties deciding when a relation exists or not. Better methods for `no_relation` identification need to be developed.

Future Work:

- Develop a better method for `no_relation` (negative class) identification.
- Extend the framework to other tasks such as Event Argument Extraction.

Label Verbalization and Entailment for Effective Zero- and Few-Shot Relation Extraction

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena and Eneko Agirre

EMNLP 2021

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology