Few-shot Information Extraction

Pre-train, Prompt, **Entail**

Eneko Agirre

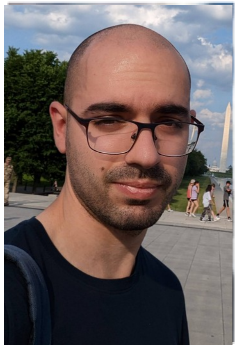
Director of HiTZ

Basque Center for Language Technology (UPV/EHU)

@eagirre

http://hitz.eus/eneko/ (slide deck)

In collaboration with



Oscar
Sainz

Oier Lopez
de Lacalle

Gorka
Labaka

Ander
Barrena

Itziar
Gonzalez-Dios

eman ta zabal zazu

Universidad del País Vasco | Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

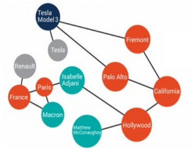
Bonan
Min

Haoling
Qiu

Raytheon
BBN

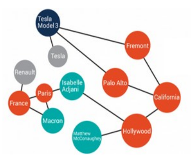
Few-shot Information Extraction?

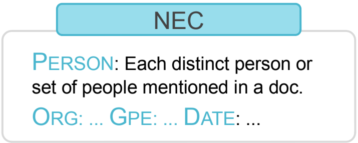
- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



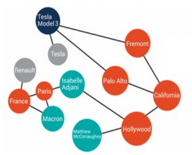


Named-entity Classification (NEC)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC

PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

Named-entity Classification (NEC)

EVENT

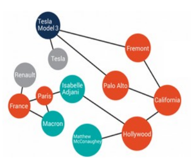
LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.

Event Extraction (EE)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC

PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

Named-entity Classification (NEC)

EVENT

LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.

Event Extraction (EE)

RELATION

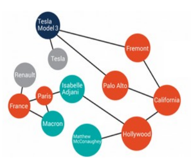
EMPLOYEEOF: Employment captures the relationship between Persons and their employers. This Relation is only taggable when it can be reasonably assumed that the PER is paid by the ORG or GPE.

Relation Extraction (RE)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC
PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

Named-entity Classification (NEC)

EVENT
LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.

Event Extraction (EE)

RELATION
EMPLOYEEOF: Employment captures the relationship between Persons and their employers. This Relation is only taggable when it can be reasonably assumed that the PER is paid by the ORG or GPE.

Relation Extraction (RE)

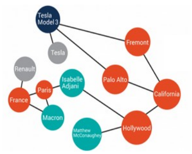
EVENT ARGUMENT
VICTIM-ARG: The person(s) who died
PLACE-ARG: Where the death takes place

Event Argument Extraction (EAE)

https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf
https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-relations-guidelines-v6.2.pdf

Few-shot Information Extraction?

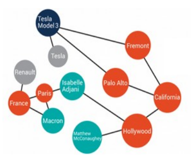
- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



John Smith, an executive at XYZ Co., died in Florida on Sunday.

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



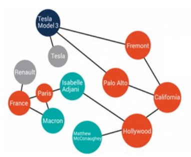
NEC

PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

John Smith$_{\text{PERSON}}$, an executive at XYZ Co.$_{\text{ORGANIZATION}}$, died in Florida$_{\text{GPE}}$ on Sunday$_{\text{DATE}}$.

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC

PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

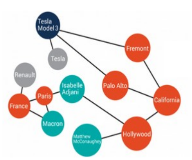
EVENT

LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.

John Smith$_{\text{PERSON}}$, an executive at XYZ Co.$_{\text{ORGANIZATION}}$, died$_{\text{LIFE.DIE}}$ in Florida$_{\text{GPE}}$ on Sunday$_{\text{DATE}}$.

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC

PERSON: Each distinct person or set of people mentioned in a doc.
ORG: ... GPE: ... DATE: ...

EVENT

LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.

RELATION

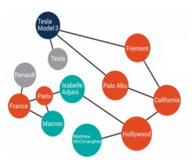
EMPLOYEEOF: Employment captures the relationship between Persons and their employers. This Relation is only taggable when it can be reasonably assumed that the PER is paid by the ORG or GPE.

John Smith$_{\text{PERSON}}$, an executive at XYZ Co.$_{\text{ORGANIZATION}}$, died$_{\text{LIFE.DIE}}$ in Florida$_{\text{GPE}}$ on Sunday$_{\text{DATE}}$.

EMPLOYEEOF

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train



NEC	EVENT	RELATION	EVENT ARGUMENT
PERSON: Each distinct person or set of people mentioned in a doc. ORG: ... GPE: ... DATE: ...	LIFE.DIE: A DIE Event occurs whenever the life of a PERSON Entity ends.	EMPLOYEEOF: Employment captures the relationship between Persons and their employers. This Relation is only taggable when it can be reasonably assumed that the PER is paid by the ORG or GPE.	VICTIM-ARG: The person(s) who died PLACE-ARG: Where the death takes place

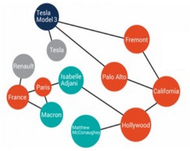
VICTIM-ARG PLACE-ARG

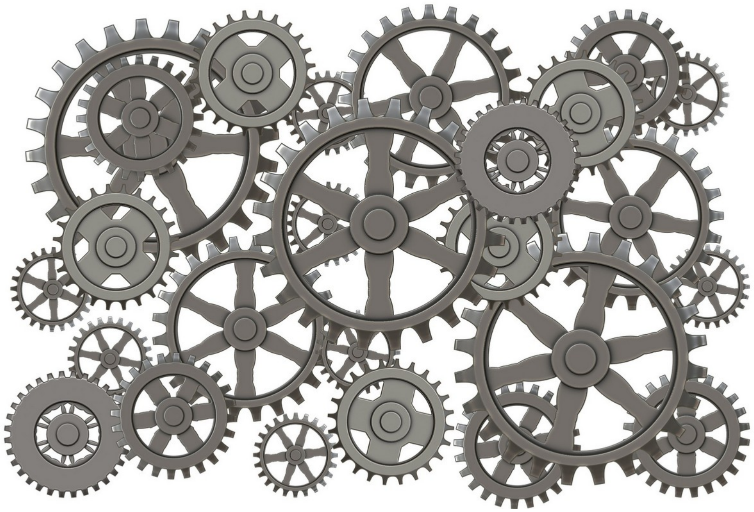
John Smith$_{\text{PERSON}}$, an executive at XYZ Co.$_{\text{ORGANIZATION}}$, died$_{\text{LIFE.DIE}}$ in Florida$_{\text{GPE}}$ on Sunday$_{\text{DATE}}$.

EMPLOYEEOF TIME-ARG

Few-shot Information Extraction?

- Adoption of NLP in companies deterred because of high effort of domain experts
  - In the case of Information Extraction, define non-trivial schemas with entities and relations of interest, annotate corpus, train supervised ML system
  - Define, annotate, train





mastro-h2020.eu/project-committees/

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

NEC VERBALIZATIONS

{X} is a person. → PERSON

{X} is a date. → DATE

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

NEC VERBALIZATIONS

{X} is a person. → PERSON

{X} is a date. → DATE

EVENT VERBALIZATIONS

{E} refers to a death. → LIFE.DIE

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

NEC VERBALIZATIONS

{X} is a person. → PERSON

{X} is a date. → DATE

EVENT VERBALIZATIONS

{E} refers to a death. → LIFE.DIE

RELATION VERBALIZATIONS

{X} is employed by {Y}. → EMPLOYEEOF

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

NEC verbalizations

{X} is a person. → Person

{X} is a date. → Date

Event verbalizations

{E} refers to a death. → Life.Die

Relation verbalizations

{X} is employed by {Y}. → EmployeeOf

Event Argument verbalizations

{X} died. → Victim-arg

Someone {E} in {X}. → Place-arg

Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

NEC verbalizations

{X} is a person. → Person

{X} is a date. → Date

Event verbalizations

{E} refers to a death. → Life.Die

Relation verbalizations

{X} is employed by {Y}. → EmployeeOf

Event Argument verbalizations

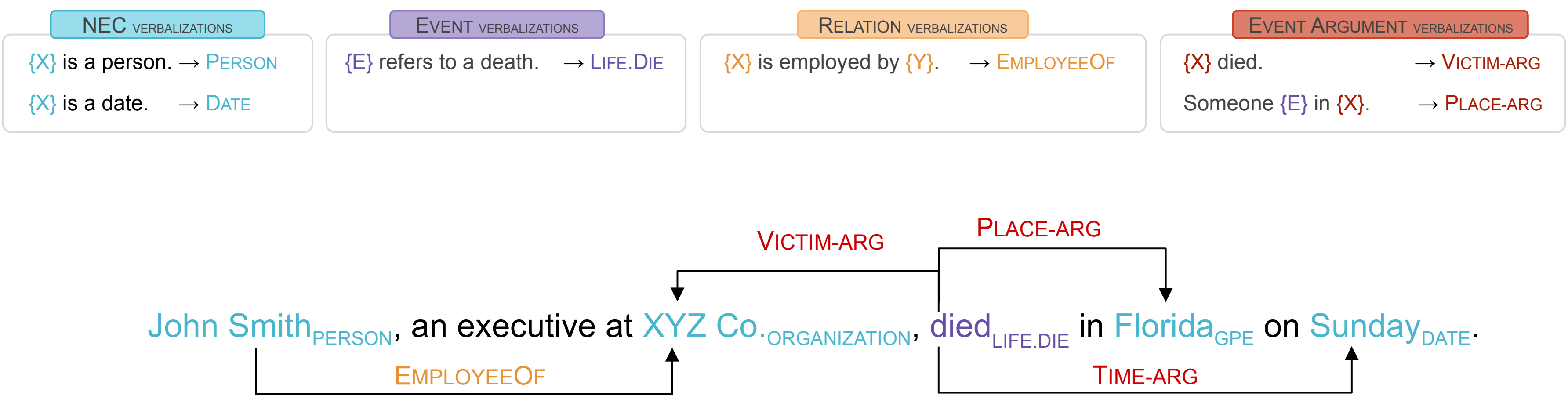
{X} died. → Victim-Arg

Someone {E} in {X}. → Place-Arg

John Smith, an executive at XYZ Co., died in Florida on Sunday.

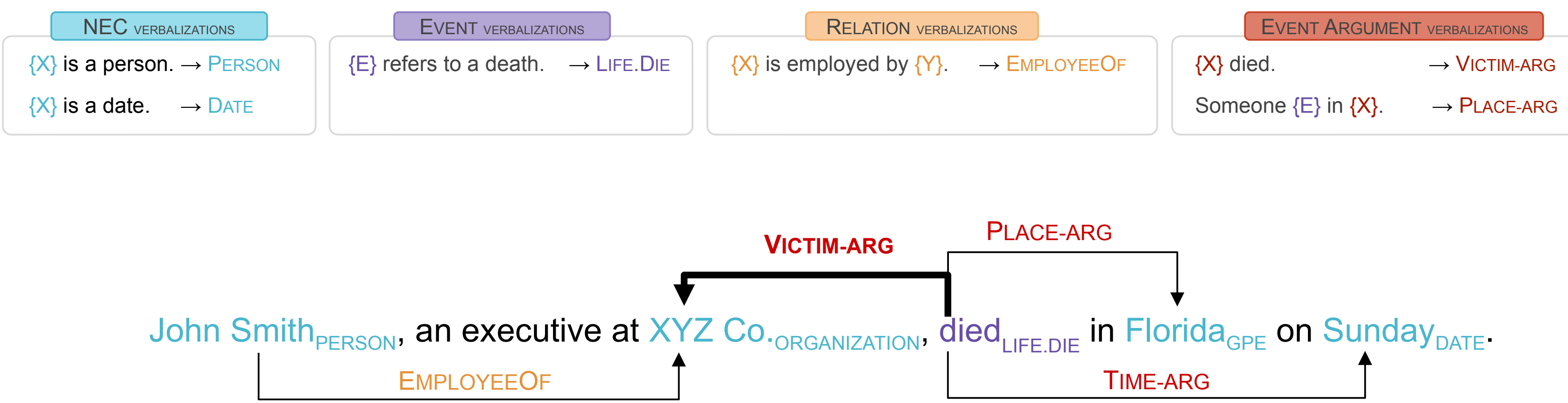
Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates



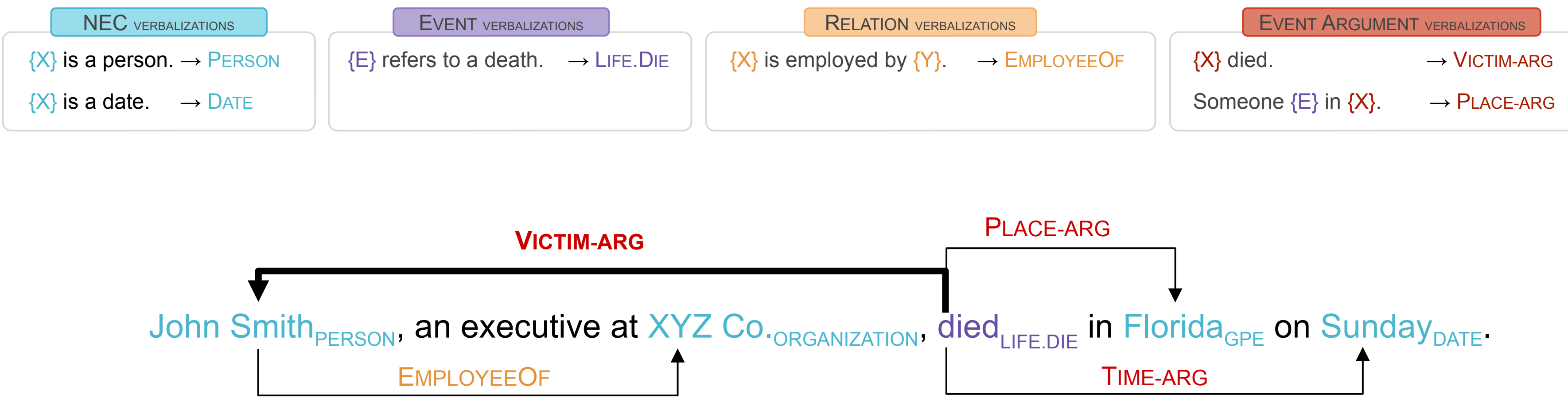
Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates



Few-shot Information Extraction?

- Interactive workflow: verbalize while defining
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates



Few-shot Information Extraction?

Define, annotate, train

vs.

Interactive workflow: verbalize while defining

- 10 times more effective (time of domain experts)

insider.com/

- Friendlier for domain experts

freepik.com/

Few-shot Information Extraction?

Thanks to latest advances:

- Large pre-trained language models (LM)
- Recast IE into natural language instructions and prompts

But LMs have limited inference ability

Few-shot Information Extraction?

Thanks to latest advances:

- Large pre-trained language models (LM)
- ~~Recast IE into natural language instructions and prompts~~
- Enhance inference abilities of LM with entailment datasets
- Recast IE as an entailment problem

Plan for the talk

- **Pre-trained Language Models**
- Prompting
- Entailment
- Few-shot Information Extraction

Pre-trained Language Models

1) Self-supervised LM pre-training

- Unlabelled data: HUGE corpora:
  Wikipedia, news, web crawl, social media, etc.
- Train some variant of a Language Model

Pre-trained Language Models

1) Self-supervised LM pre-training

- Unlabelled data: HUGE corpora:
  Wikipedia, news, web crawl, social media, etc.
- Train some variant of a Language Model

2) Supervised pre-training

- Very common in vision (ImageNet), standalone.
  NLP in-conjuction with self-supervised LM.
- Task-specific: e.g. transfer from one Q&A dataset to another
- Entailment for improved inference
  (e.g. Sainz et al. 2021; Wang et al. 2021)
- All available tasks (e.g. T0, Sahn et al. 2021)

Self-supervised LM pre-training

Informally, learn parameters Θ using some variant of

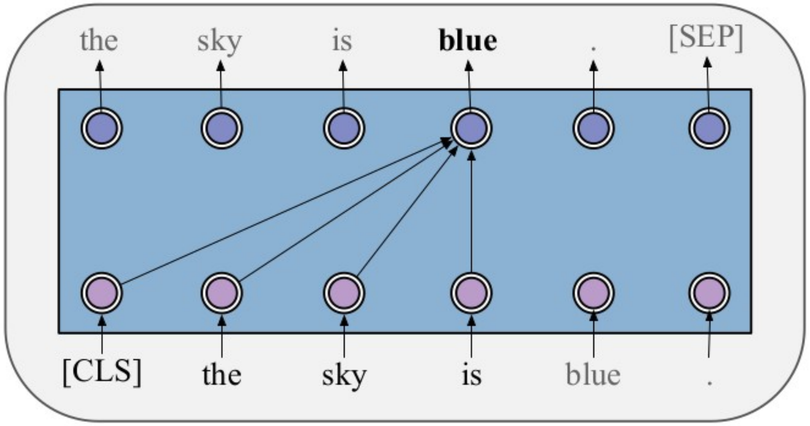
$P_{\Theta}(\textit{text} \mid \textit{some other text})$

Self-supervised LM pre-training

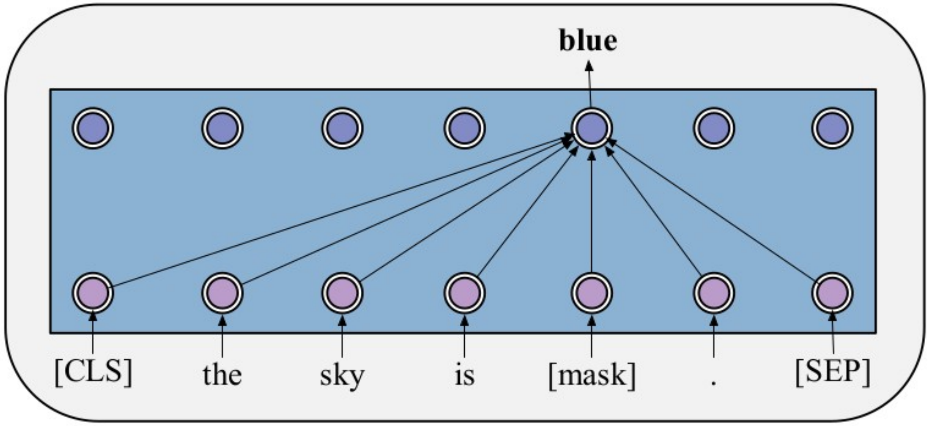
Informally, learn parameters *ϴ* using some variant of

$P_{\Theta}(\text{text} \mid \text{some other text})$

(Causal) Language Model (GPT)



Masked Language Model (BERT)



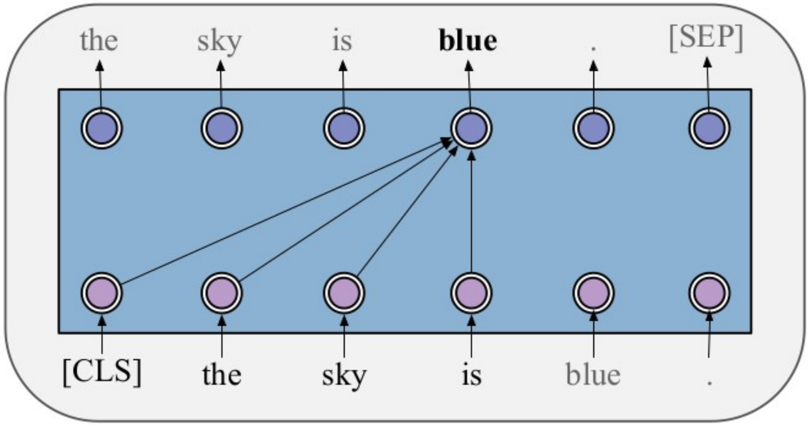
Pre-Trained Models: Past, Present and Future (Han et al. 2021)

Self-supervised LM pre-training

Informally, learn parameters Θ using some variant of

$P_{\Theta}(\textit{text} \mid \textit{some other text})$

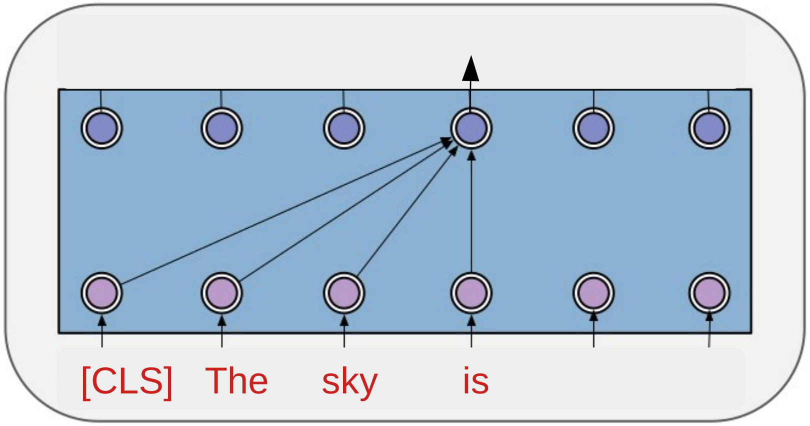
(Causal) Language Model (GPT)



- Self-attention: left
- Loss: next word

Pre-Trained Models: Past, Present and Future (Han et al. 2021)

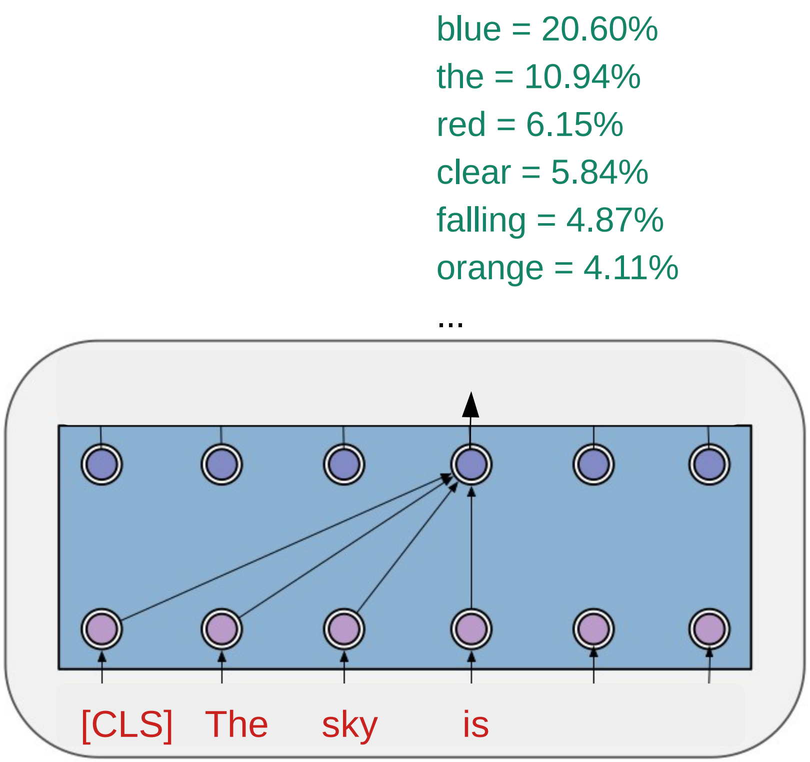
Self-supervised LM pre-training



- Self-attention: left
- Loss: next word
- At inference: generates text conditioned on prefix

OpenAI Playground DaVinci

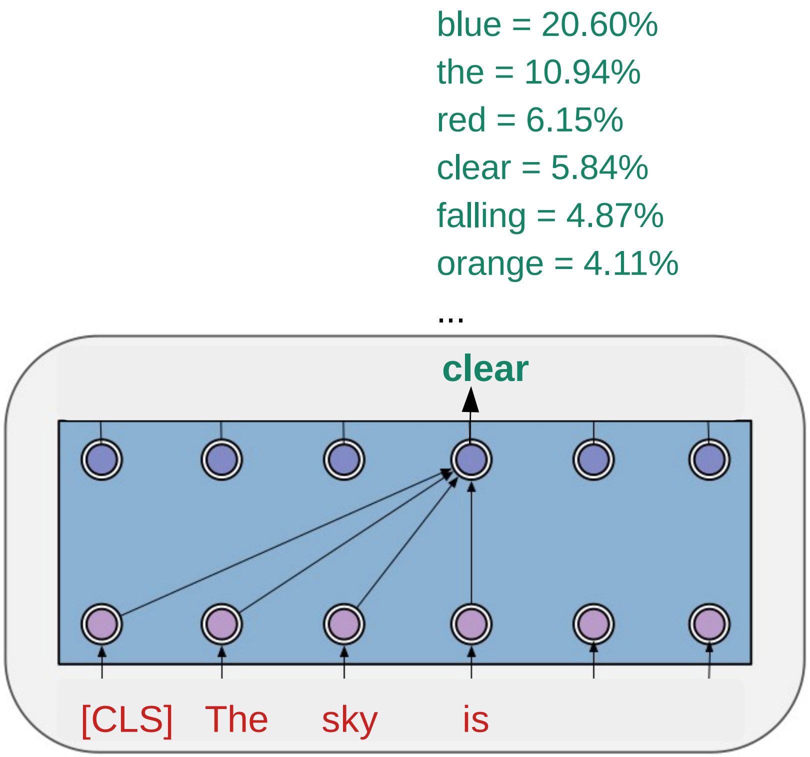
Self-supervised LM pre-training



- Self-attention: left
- Loss: next word
- At inference: generates text conditioned on prefix

OpenAI Playground DaVinci

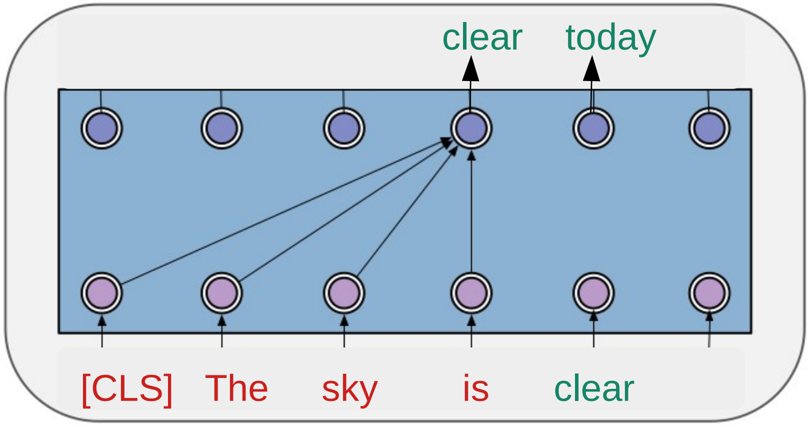
Self-supervised LM pre-training



- Self-attention: left
- Loss: next word
- At inference: generates text conditioned on prefix

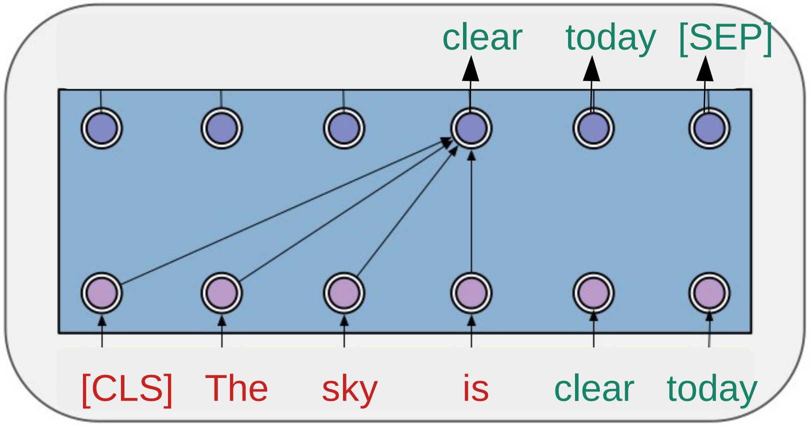
OpenAI Playground DaVinci

Self-supervised LM pre-training



- Self-attention: left
- Loss: next word
- At inference: generates text conditioned on prefix

Self-supervised LM pre-training



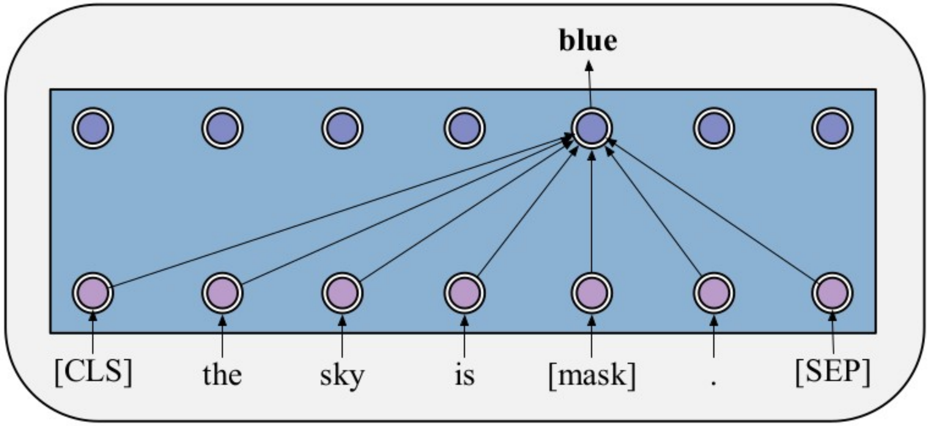
- Self-attention: left
- Loss: next word
- At inference: generates text conditioned on prefix

Self-supervised LM pre-training

Informally, learn parameters ϴ using some variant of

$P_{\Theta}(\textit{text} \mid \textit{some other text})$

Masked Language Model (BERT)



Pre-Trained Models: Past, Present and Future (Han et al. 2021)

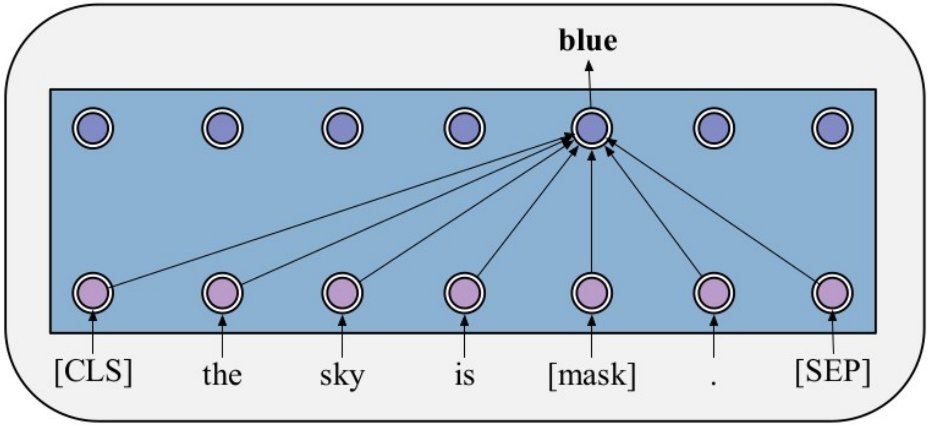
Self-supervised LM pre-training

Informally, learn parameters Θ using some variant of

$$P_{\Theta}(\textit{text} \mid \textit{some other text})$$

- Self-attention:
  left and right
- Loss:
  masked words

Masked Language Model (BERT)

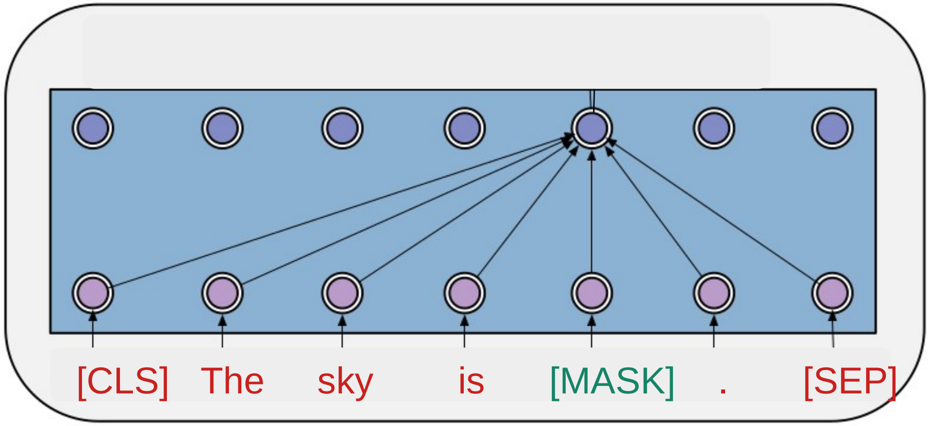


Pre-Trained Models: Past, Present and Future (Han et al. 2021)

Self-supervised LM pre-training

blue = 20.60%
~~the = 10.94%~~
red = 6.15%
clear = 5.84%
~~falling = 4.87%~~
orange = 4.11%
...

- Self-attention:
  left and right
- Loss:
  masked words
- At inference it can fill explicitly masked tokens



Pre-Trained Models: Past, Present and Future (Han et al. 2021)

Fine-tuning on a specific task

Sentence classification:

Add a classification head
on top of the [CLS] token

Sentiment
Analysis

Training example:
(The sky is fantastic,Positive)

Pre-Trained Models: Past, Present and Future (Han et al. 2021)

Fine-tuning on a specific task

Sentence classification:

Add a classification head

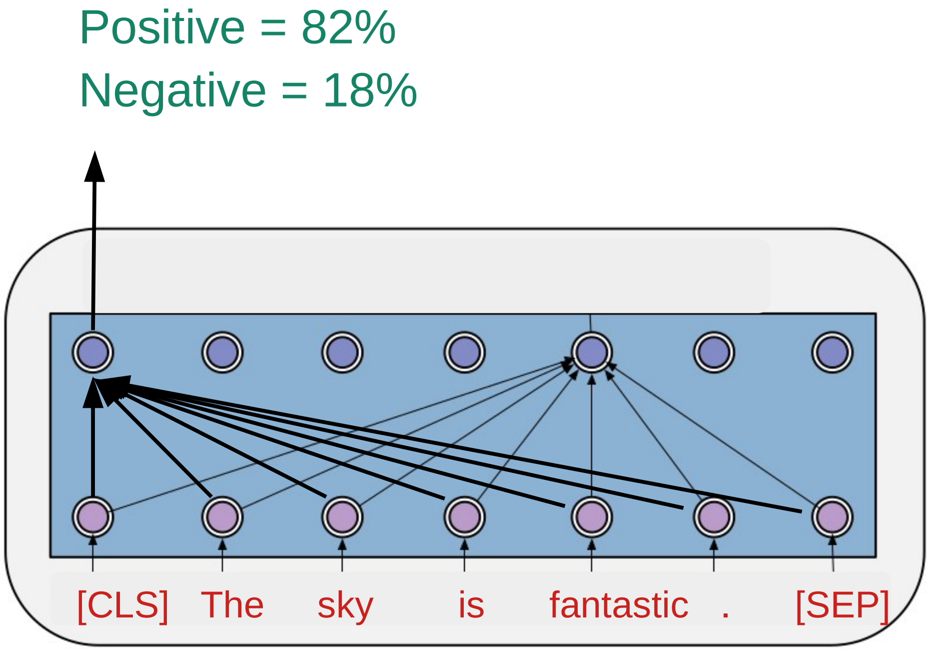
on top of the [CLS] token

Sentiment

Analysis

Training example:

(The sky is fantastic,Positive)



Pre-Trained Models: Past, Present and Future (Han et al. 2021)

NLP performance improvement

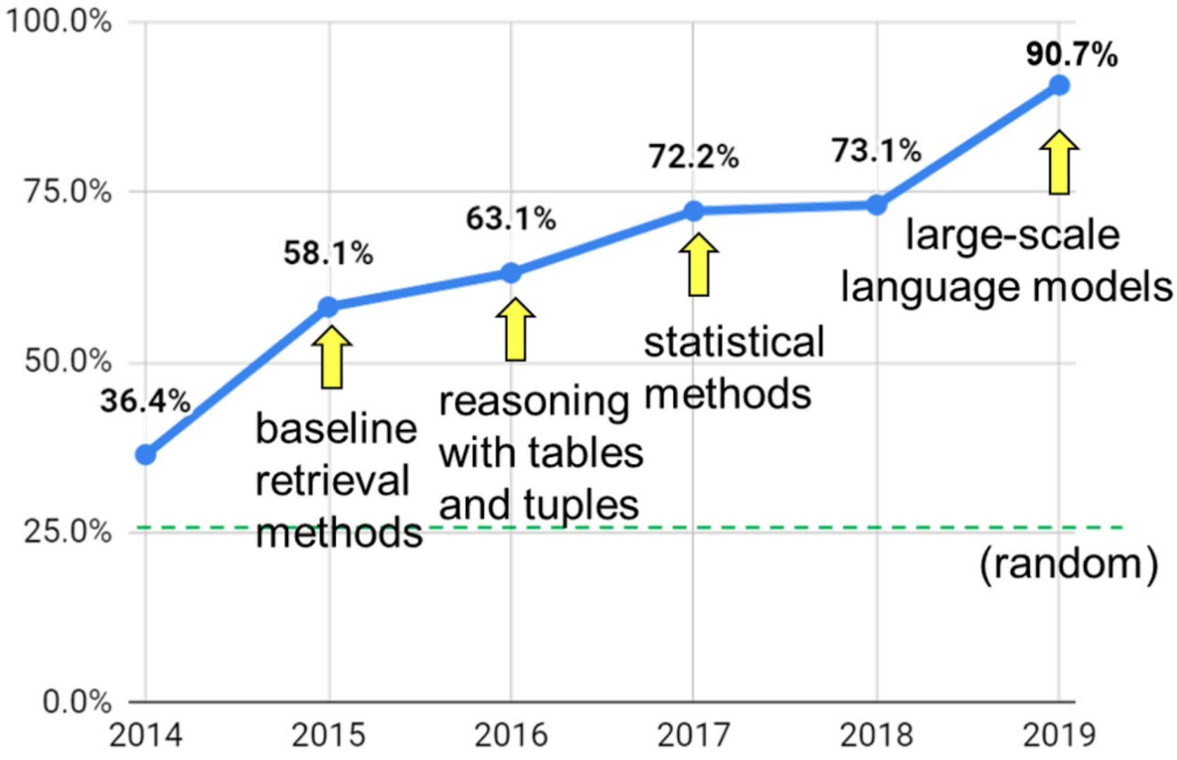
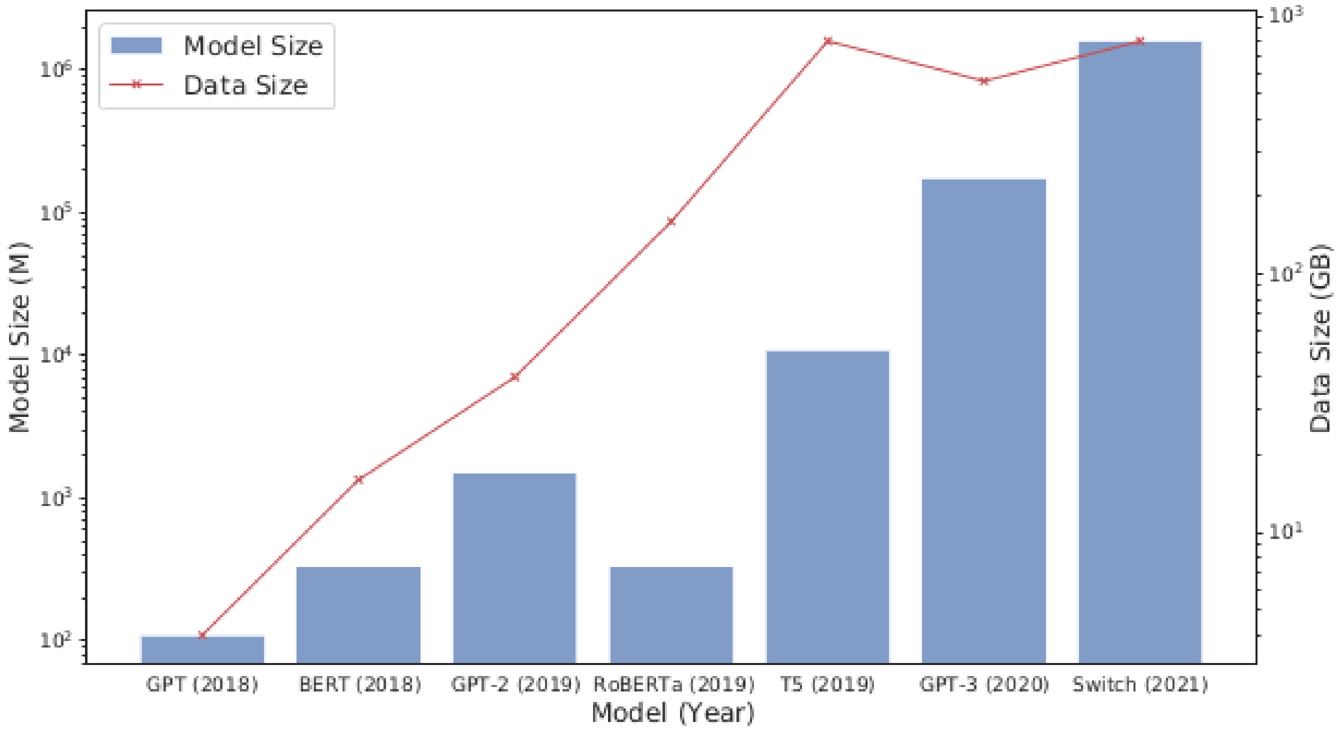


Figure 2: Aristo's scores on Regents 8th Grade Science (non-diagram, multiple choice) over time (held-out test set).

Clark et al. AI Magazine 41 (4) 2020

Scaling up pretraining



(b) The model size and data size applied by recent NLP PTMs. A base-10 log scale is used for the figure.

Pre-Trained Models: Past, Present and Future (Han et al. 2021)

Why do Pre-trained LMs work so well?

- LM is a very difficult task, even for humans.
  - LMs compress any possible context into a vector that generalizes over possible completions.
  - Forced to learn syntax, semantics, encode facts about the world, etc.
- LM consume huge amounts of data
- The fine-tuning stage can exploit all knowledge in LM, instead of starting from scratch

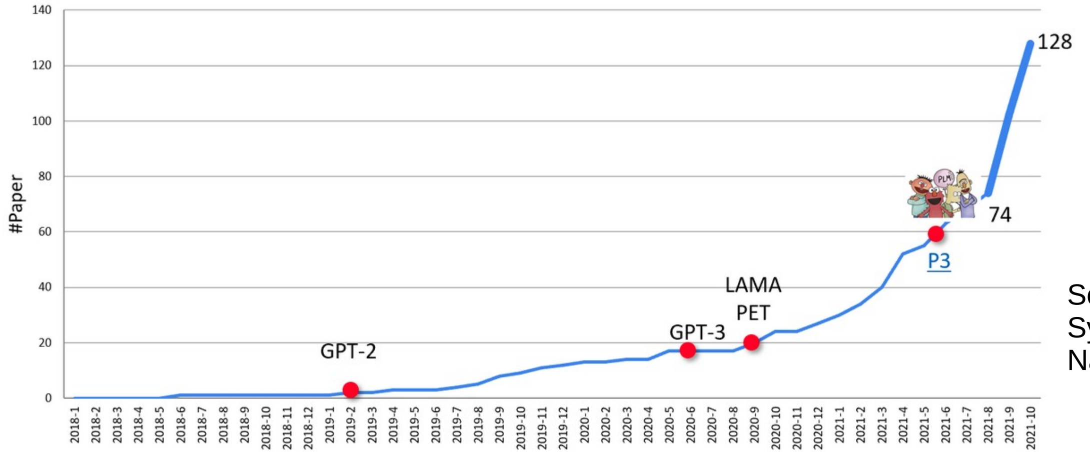
Plan for this session

- Pre-trained LM
- **Prompting**
- Entailment
- Few-shot Information Extraction

What is prompt learning?

Encourage a pre-trained model
to make particular predictions by providing
a “prompt” specifying the task to be done

A new paradigm: Pre-train, prompt, predict



Source: Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (Liu et al. 2022)

What is prompt learning?

Rationale:

Recast NLP tasks into natural language, so Pretrained Language Models can apply their knowledge about language and the world

What is prompt learning?

Rationale:

Recast NLP tasks into natural language, so Pretrained Language Models can apply their knowledge about language and the world

Related ideas, zero-shot and few-shot

Learn a task with minimal task description:

- Instructions on what the task is
- Present task to LM as a prompt
- If few-shot: prepend handful of labeled examples

Sentiment analysis

The sky is fantastic .

Positive

Negative

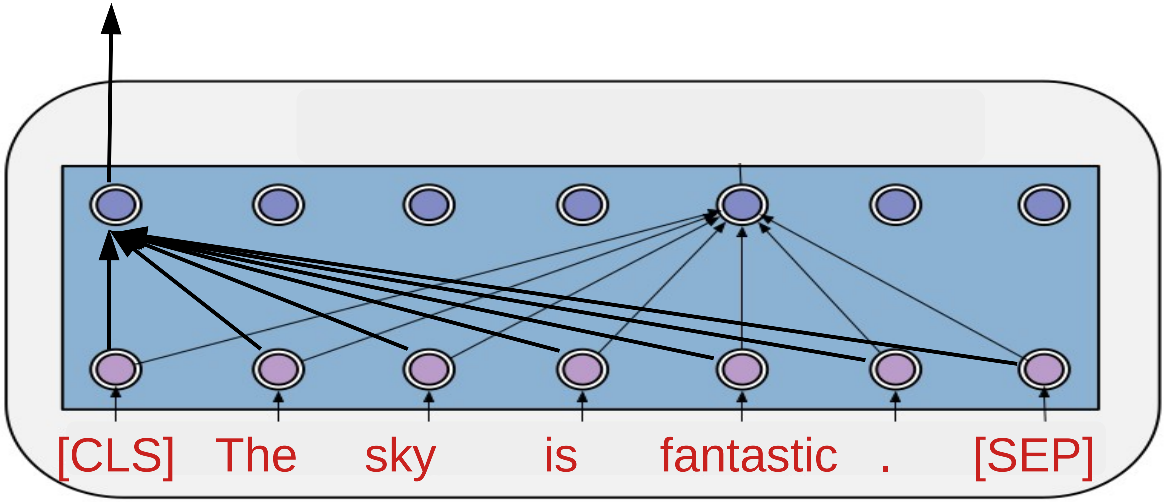
Sentiment analysis

Positive = 82%

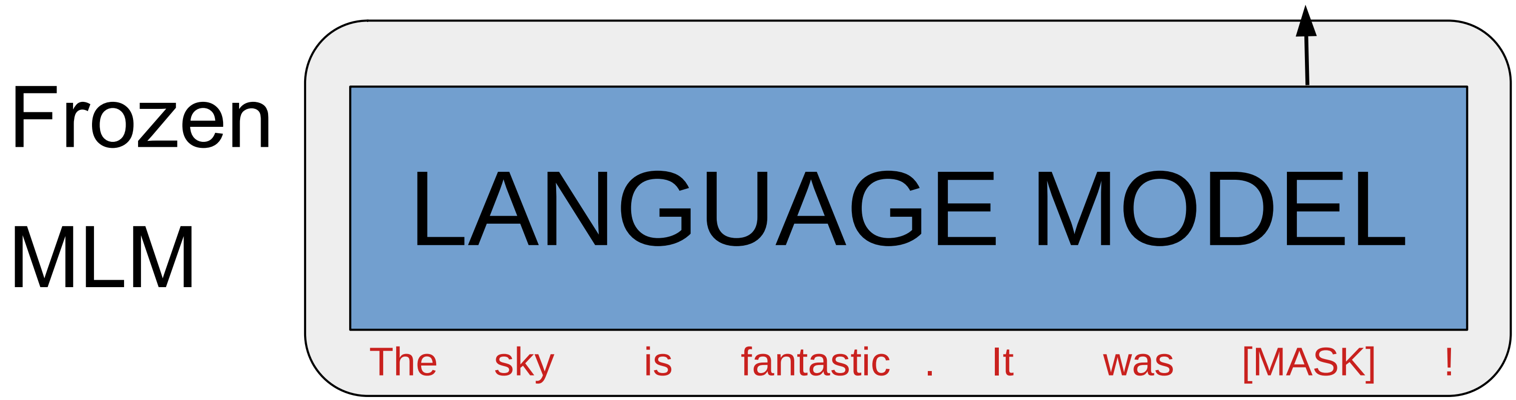
Negative = 18%

Fine-tuned

LM



LM prompting (zero-shot)

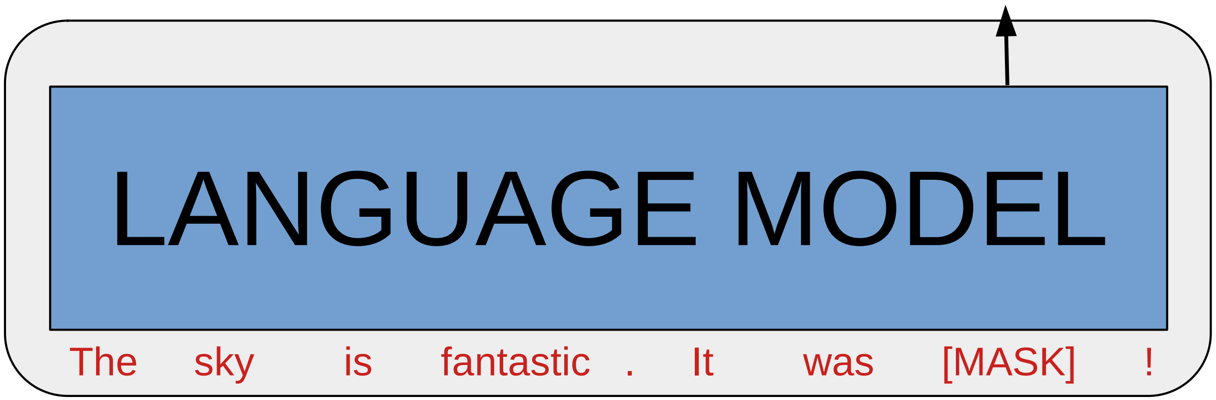


Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (zero-shot)

Frozen

MLM



P1=P(great | The sky is fantastic. It was [MASK] !)

P2=P(terrible | The sky is fantastic. It was [MASK] !)

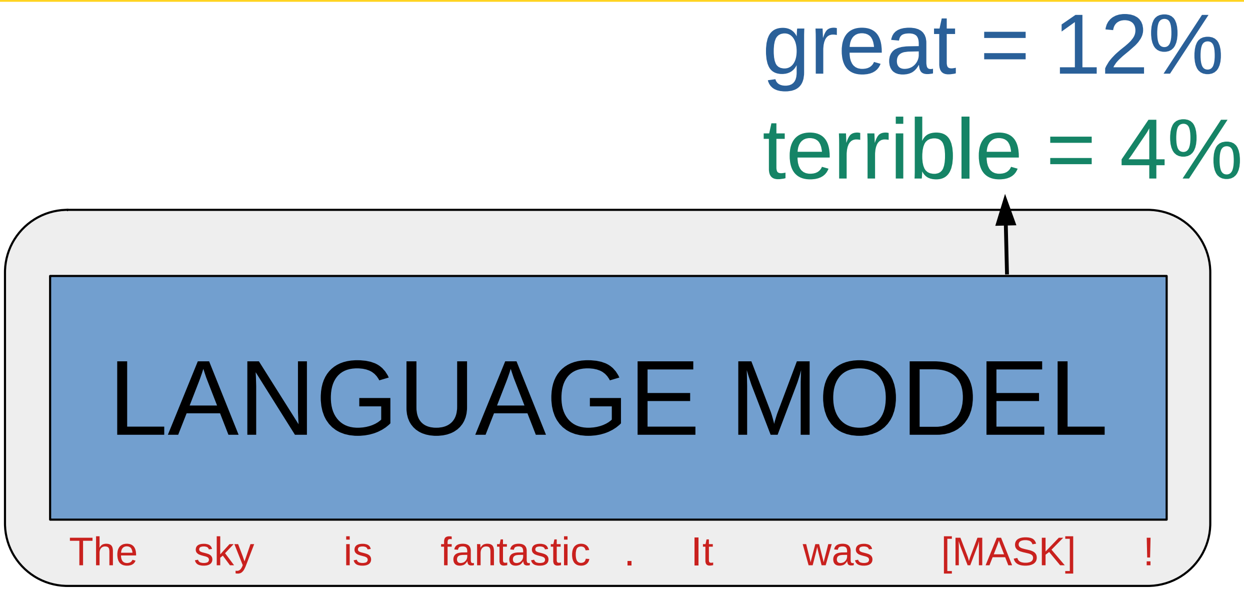
P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (zero-shot)

Frozen

MLM



P1=P(great | The sky is fantastic. It was [MASK] !)
P2=P(terrible | The sky is fantastic. It was [MASK] !)

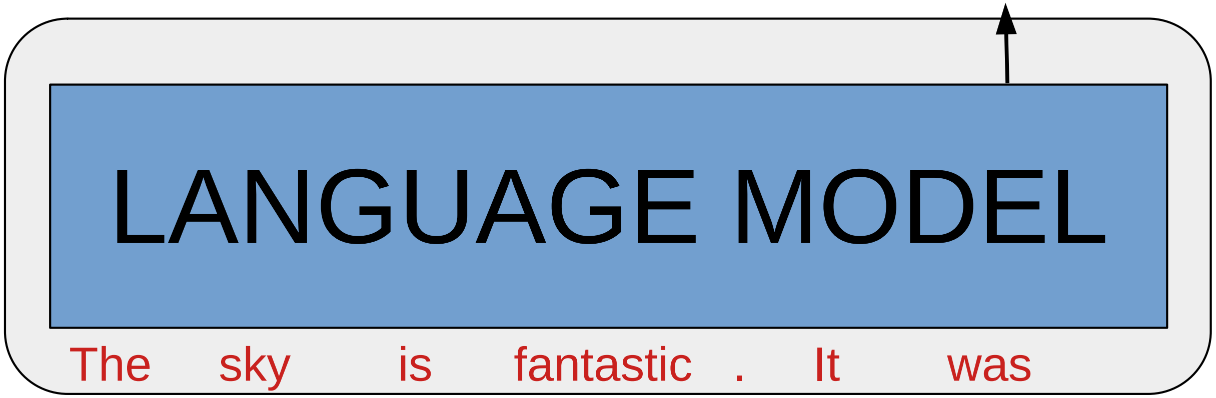
P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (zero-shot)

Frozen

LM

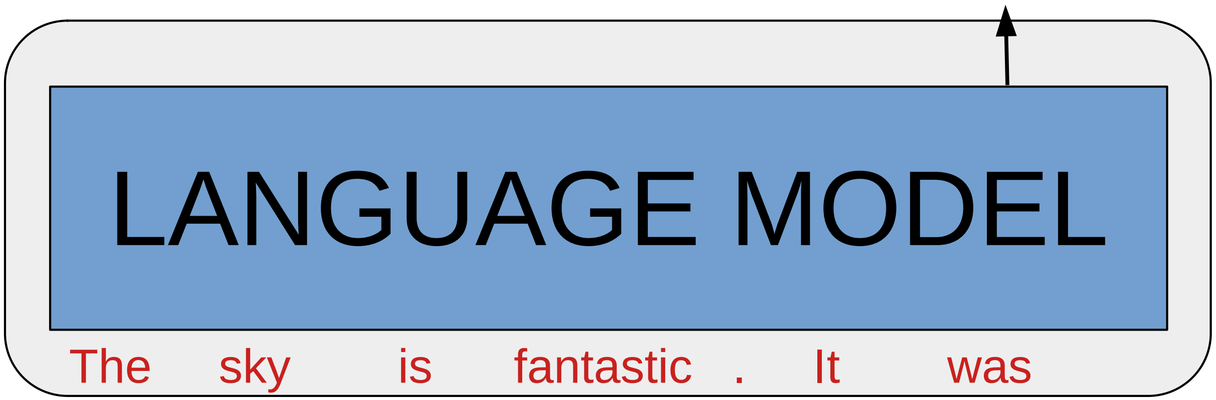


Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (zero-shot)

Frozen

LM



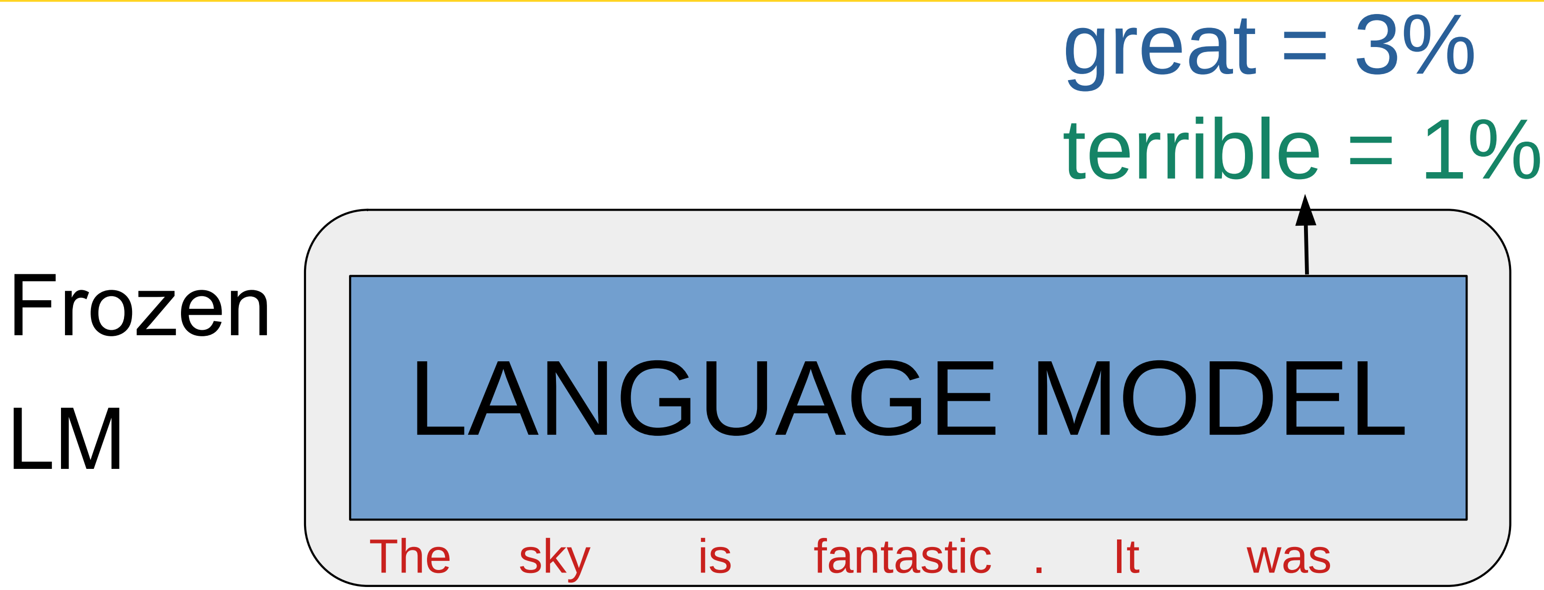
P1=P(great | The sky is fantastic. It was)

P2=P(terrible | The sky is fantastic. It was)

P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (zero-shot)



P1=P(great | The sky is fantastic. It was)
P2=P(terrible | The sky is fantastic. It was)

P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (few-shot) In-context learning

Training Data

Text: I'm not sure I like it.
Label: Negative

Text: Thank you for the amazing help.
Label: Positive

S1 = I'm not sure I like it. It was terrible!
S2 = Thank you for the amazing help. It was great!
S = The sky is fantastic. It was _____

LANGUAGE MODEL

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (few-shot) In-context learning

Training Data

Text: I'm not sure I like it.
Label: Negative

Text: Thank you for the amazing help.
Label: Positive

S1 = I'm not sure I like it. It was terrible!
S2 = Thank you for the amazing help. It was great!
S = The sky is fantastic. It was _____

LANGUAGE MODEL

P1=P(great | S1 \n S2 \n The sky is fantastic. It was)
P2=P(terrible | S1 \n S2 \n The sky is fantastic. It was)

P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (few-shot)
In-context learning

Training Data

Text: I'm not sure I like it.
Label: Negative

Text: Thank you for the amazing help.
Label: Positive

S1 = I'm not sure I like it. It was terrible!
S2 = Thank you for the amazing help. It was great!
S = The sky is fantastic. It was ____

LANGUAGE MODEL

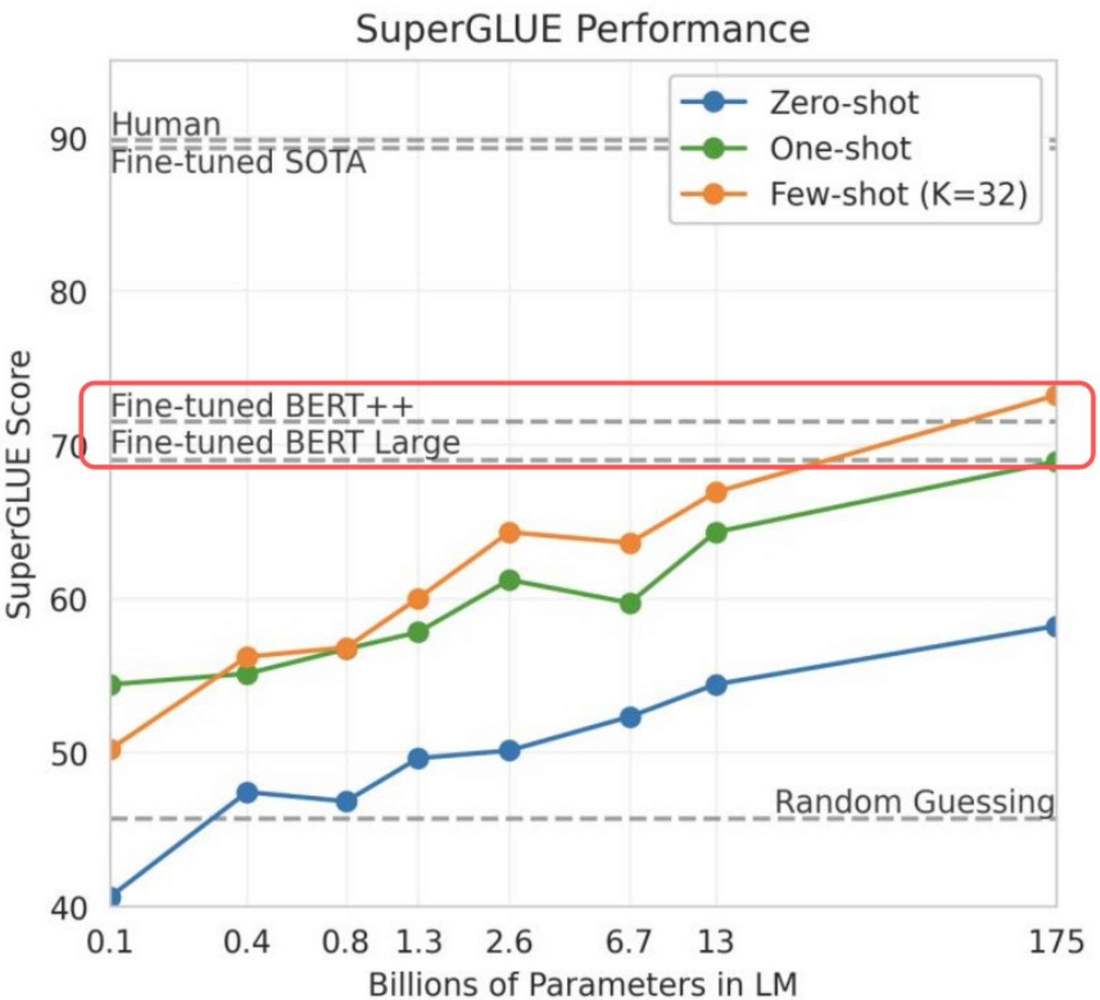
NO PARAMETER UPDATE

P1=P(great | S1 \n S2 \n The sky is fantastic. It was)
P2=P(terrible | S1 \n S2 \n The sky is fantastic. It was)

P1 > P2 then Positive

Language Models are Few-Shot Learners (Brown et al. 2020)

LM prompting (few-shot) In-context learning



(Brown et al. 2020)

Domain-experts provides templates / label map

Template: [x] It was __ !
Label map: great <=> positive

The sky is fantastic. It was ____

Template: Review: [x] Sentiment: __
Label map: positive <=> positive

Review: The sky is fantastic. Sentiment: ____

Domain-experts provide templates / label map

Template: [x] It was __ !
Label map: great <=> positive

I'm not sure I like it.	It was terrible!
Thank you for the amazing help.	It was great!
The sky is fantastic.	It was ____!

Template: Review: [x] Sentiment: __
Label map: positive <=> positive

Review: I'm not sure I like it.	Sentiment: negative
Review: Thank you for the amazing help.	Sentiment: positive
Review: The sky is fantastic.	Sentiment: ____

Zero-shot and few-shot
No parameter update

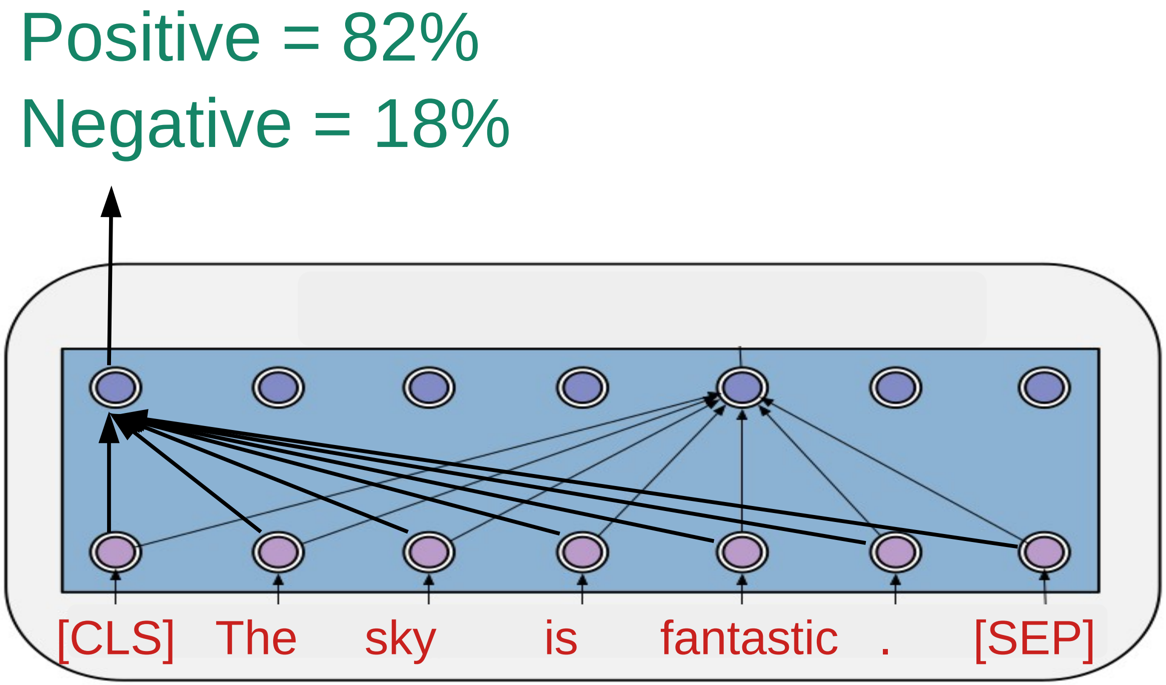
- Good results with the largest GPT-3 models (175B)
- Even if there is no parameter update
- Large variance depending on prompts (templates and label map)
- Development of prompts should be on available examples only (Perez et al. 2021)

Few-shot learning with prompts and parameter updates

Traditional fine-tuning

Training example:
(The sky is fantastic,Positive)

Fine-tuned LM



Few-shot learning with prompts and parameter updates

Traditional fine-tuning

- Low results on few-shot setting

Few-shot learning with prompts and parameter updates

Fine-tune LM using **prompted datasets**

Usually smaller LM (PET)

Training example:
(The sky is fantastic, Positive)

Prompted training example:
(The sky is fantastic. It was [MASK] !, great)

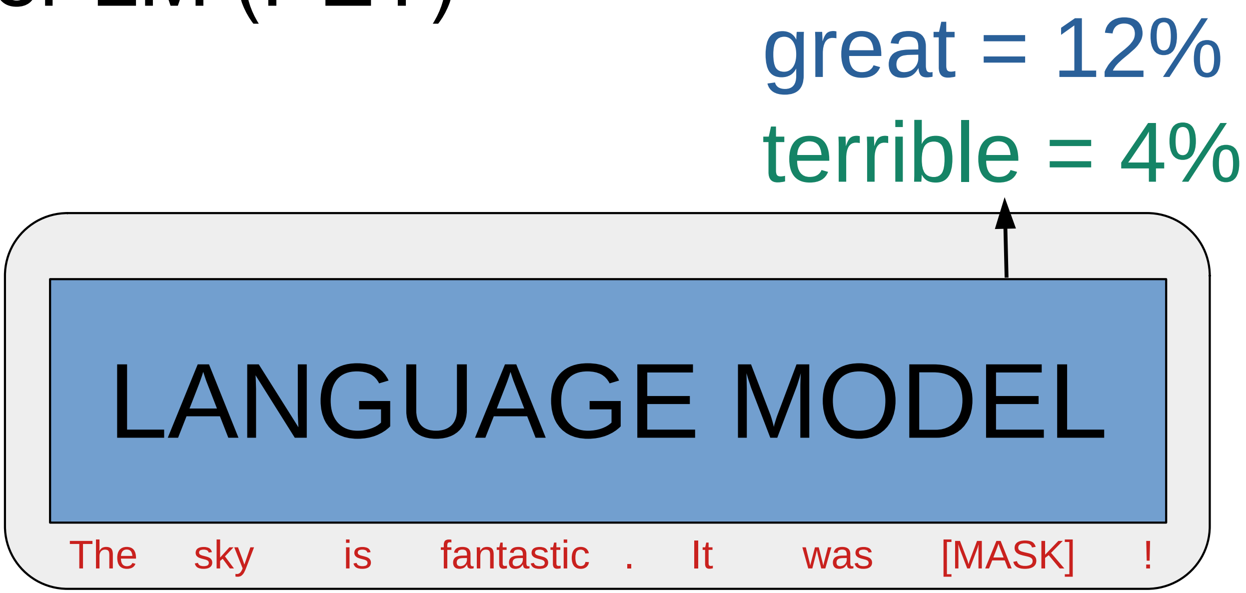
Exploiting Cloze Questions for Few Shot Text Classification and NLI (Schick and Schutze, 2020)

Few-shot learning with prompts and parameter updates

Fine-tune LM using **prompted datasets**

Usually smaller LM (PET)

Fine-tuned LM



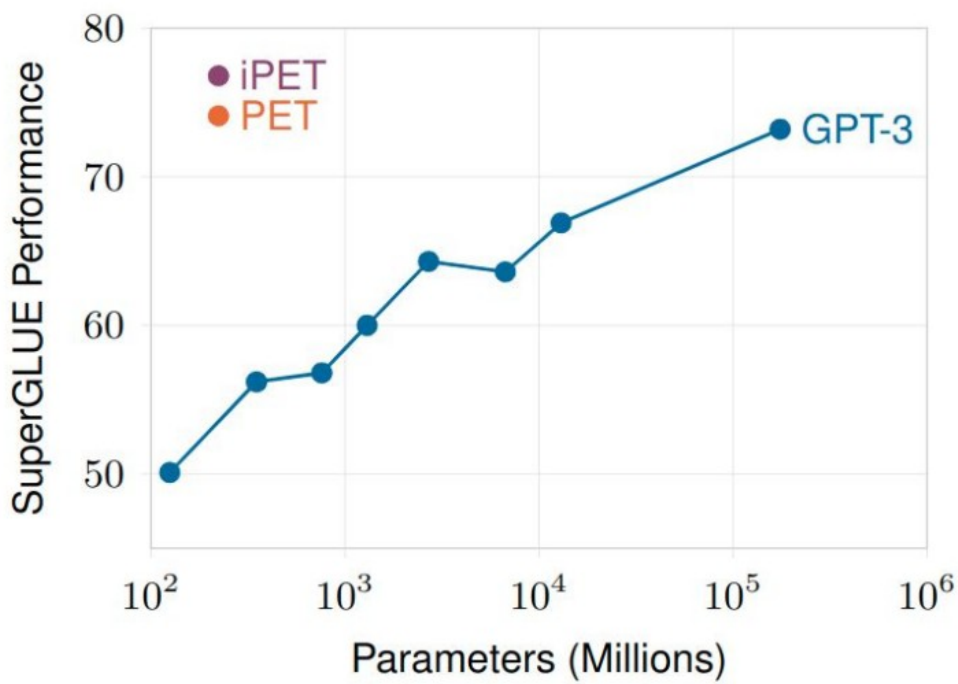
Exploiting Cloze Questions for Few Shot Text Classification and NLI (Schick and Schutze, 2020)

Few-shot learning with prompts and parameter updates

PET outperforms GPT-3 with 1000x less parameters

Ensembling

Iterations



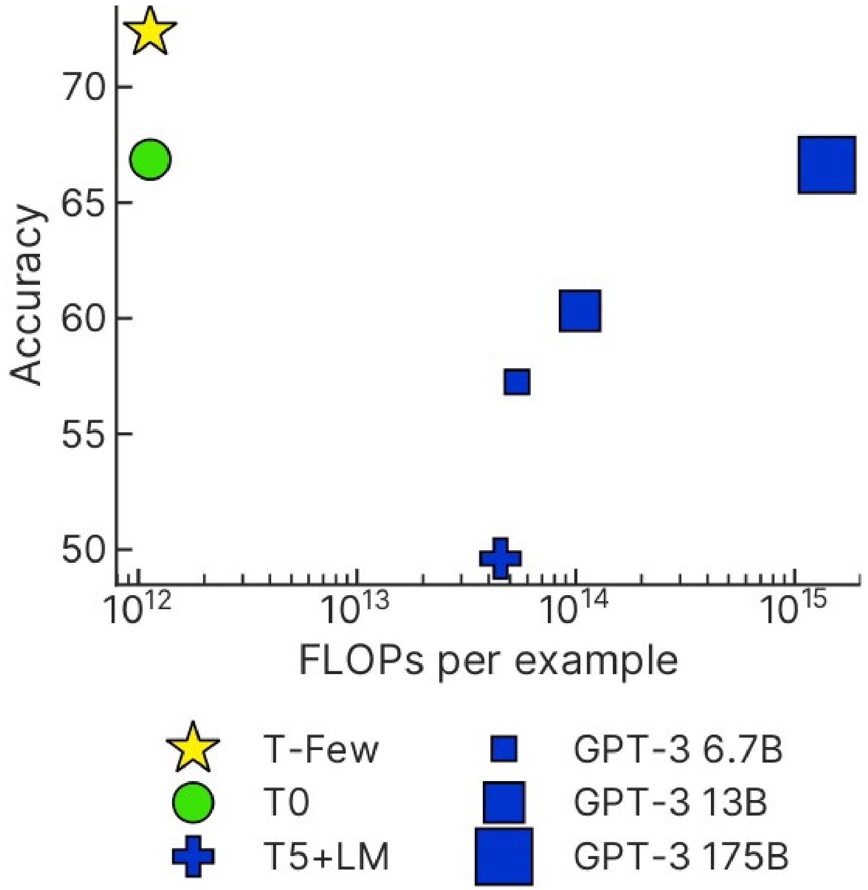
Exploiting Cloze Questions for Few Shot Text Classification and NLI (Schick and Schutze, 2020)

Few-shot learning with prompts and parameter updates

T-Few outperforms GPT-3 on held-out T0 tasks

80 times less parameters

Chart shows efficiency at inference



Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning (Liu et al. 2022)

Conclusions on prompting

- Size of models and update of parameters
  - Larger causal LM, no update: best zero-shot, strong few-shot
  - Smaller MLM, update: best few-shot (also encoder-decoder)

Conclusions on prompting

- Size of models and update of parameters
  - Larger causal LM, no update: best zero-shot, strong few-shot
  - Smaller MLM, update: best few-shot (also encoder-decoder)
- Inference ability is limited:
  - Poor results on entailment datasets (Brown et al. 2021)
  - BIG-BENCH: model performance and calibration both improve with scale, but are poor in absolute terms (Srivastava et al. 2022)
  - No wonder, LMs are capped by the phenomena needed to predict masked words, so no need to learn anything else

Conclusions on prompting

Improving inference ability is an open problem:

- PaLM: chain-of-thought fine-tuning allows to plan how to reach to result via intermediate results
- Natural-instructions: definition of the task is longer
- Combine LMs and reasoners
- Our proposal: teach inference ability via labeled entailment datasets

PaLM: Scaling Language Modeling with Pathways (Chowderhy et al. 2022)
Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks (Wang et al. 2022)

Plan for this session

- Pre-trained LM
- Prompting
- **Entailment**
- Few-shot Information Extraction

Textual Entailment (RTE), Natural Language Inference (NLI)

Dagan et al. 2005 (refined Manning et al. 2006)

- We say that Text entails Hypothesis if, typically, a human reading Text would infer that Hypothesis is most likely true.

Bowman and Zhu, NAACL 2019 tutorial

Textual Entailment (RTE), Natural Language Inference (NLI)

Dagan et al. 2005 (refined Manning et al. 2006)

- We say that Text entails Hypothesis if, typically, a human reading Text would infer that Hypothesis is most likely true.

Text (Premise): I’m not sure what the overnight low was

Hypothesis: I don't know how cold it got last night.

{entailment, contradiction, neutral}

Bowman and Zhu, NAACL 2019 tutorial

Textual Entailment (RTE), Natural Language Inference (NLI)

Dagan et al. 2005 (refined Manning et al. 2006)

- We say that Text entails Hypothesis if, typically, a human reading Text would infer that Hypothesis is most likely true.

Text (Premise): I’m not sure what the overnight low was

Hypothesis: I don't know how cold it got last night.

{**entailment**, contradiction, neutral}

Bowman and Zhu, NAACL 2019 tutorial

Textual Entailment (RTE), Natural Language Inference (NLI)

NLI datasets widely used to measure quality of models.

To perform well, NL understanding methods need to tackle several phenomena:

Textual Entailment (RTE), Natural Language Inference (NLI)

NLI datasets widely used to measure quality of models.

To perform well, NL understanding methods need to tackle several phenomena:

- Lexical entailment (cat vs. animal, cat vs. dog)
- Quantification (all, most, fewer than eight)
- Lexical ambiguity and scope ambiguity (bank, ...)
- Modality (might, should, ...)
- Common sense background knowledge
- ...

Compositional interpretation without grounding.

Textual Entailment (RTE), Natural Language Inference (NLI)

Common tasks can be cast as entailment premise-hypothesis pairs:

- **Information Extraction**: Given a text (premise), check whether it entails a relation (hypothesis)
- **Question Answering**: given a question (premise) identify a text that entails an answer (hypothesis)
- **Information Retrieval**: Given a query (hypothesis) identify texts that entail the query (premise)
- **Summarization** ...

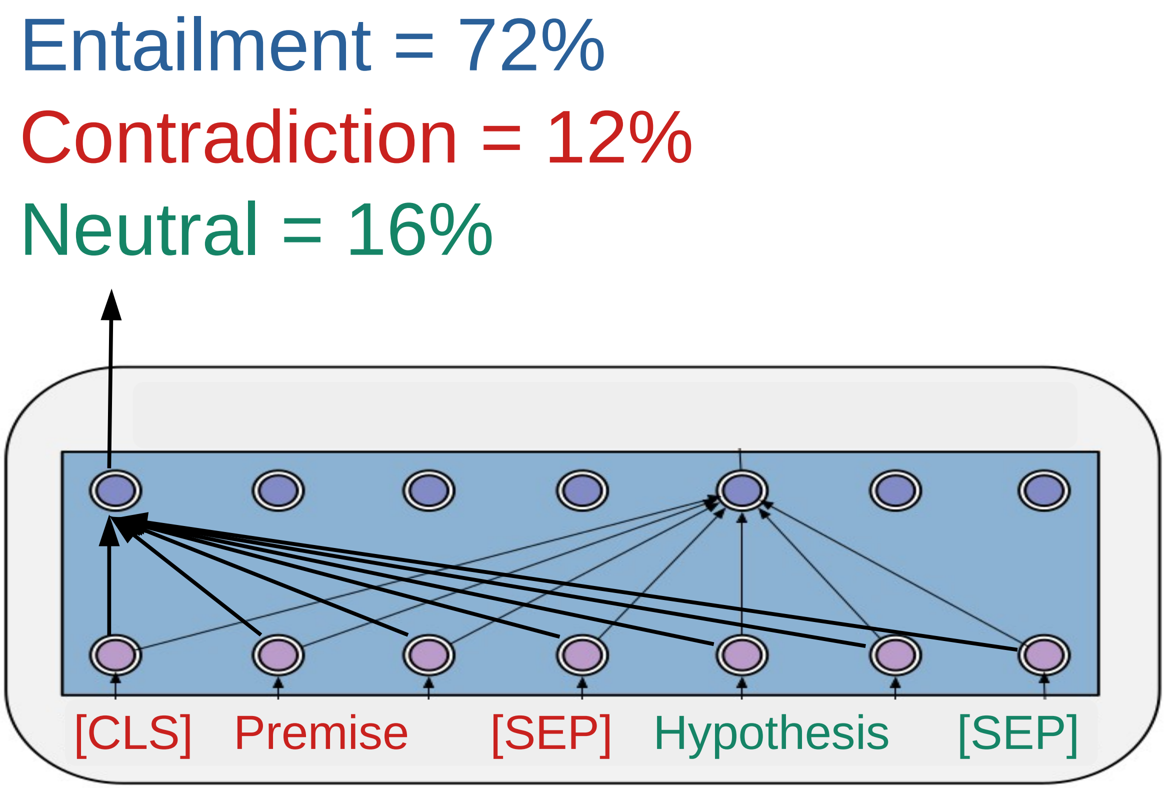
Textual Entailment (RTE), Natural Language Inference (NLI)

Datasets:

- **RTE** 1-7 (Dagan et al. 2006-2012)
  Premises (texts) drawn from naturally occurring text.
  Expert-constructed hypotheses.
  5000 examples.
- **SNLI**, **MultiNLI** (Bowman et al. 2015; Williams et al. 2017)
  Crowdsourcers provided hypothesis for captions. Then extended to other genres. 1 million examples.
  - Biases in hypotheses (Gururangan et al., 2018; Poliak et al., 2018)
  - Data generation with naïve annotators (Geva et al. 2019), artefacts
- **FEVER-NLI** (Nie et al. 2019)
  Fact verification dataset. 200,000 examples.
- **ANLI**: (Nie et al. 2012)
  Adversarially created manually. 168,000 examples.

Textual Entailment (RTE), Natural Language Inference (NLI)

Fine-tune MLM on NLI



(Devlin et al. 2019)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

	Premise
Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995.
Premise	question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False?
	answer:
Target Completion →	False

Language Models are Few-Shot Learners (Brown et al. 2020)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995. question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False? answer:
Target Completion →	False

Label

Language Models are Few-Shot Learners (Brown et al. 2020)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Billy died at his home in Tampa, Fla. on Sunday

question: **Billy died in Florida**. True or False?

answer:

Language Models are Few-Shot Learners (Brown et al. 2020)

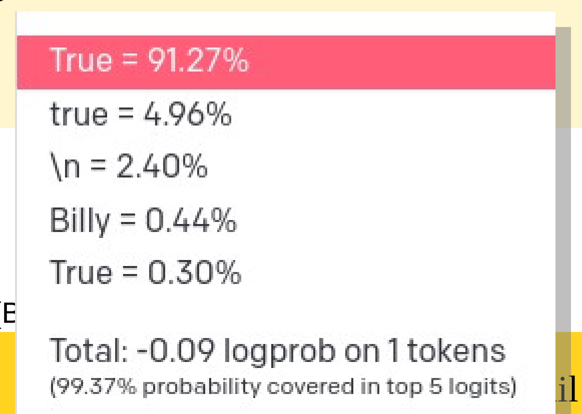
Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Billy died at his home in Tampa, Fla. on Sunday

question: **Billy died in Florida**. True or False?

answer: **True**



OpenAI Playground DaVinci

Language Models are Few-Shot Learners (E

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Billy died at his home in Tampa, Fla. on Sunday

question: **Billy died in Texas**. True or False?

answer:

OpenAI Playground DaVinci

Language Models are Few-Shot Learners (Brown et al. 2020)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Billy died at his home in Tampa, Fla. on Sunday

question: **Billy died in Texas**. True or False?

answer: **False**

False = 83.27%
false = 6.25%
\n = 5.70%
True = 3.33%
true = 0.47%

Total: -0.18 logprob on 1 tokens
(99.02% probability covered in top 5 logits)

OpenAI Playground DaVinci

Language Models are Few-Shot Learners (Brow

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts

Billy did not die **at his home in Tampa, Fla. on Sunday**

question: **Billy died in Florida**. True or False?

answer:

Language Models are Few-Shot Learners (Brown et al. 2020)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts fails

Billy did not die at his home in Tampa, Fla. on Sunday

question: **Billy died in Florida.** True or False?

answer: **True**

True = 90.07%
true = 5.29%
\n = 2.81%
Billy = 0.53%
False = 0.40%

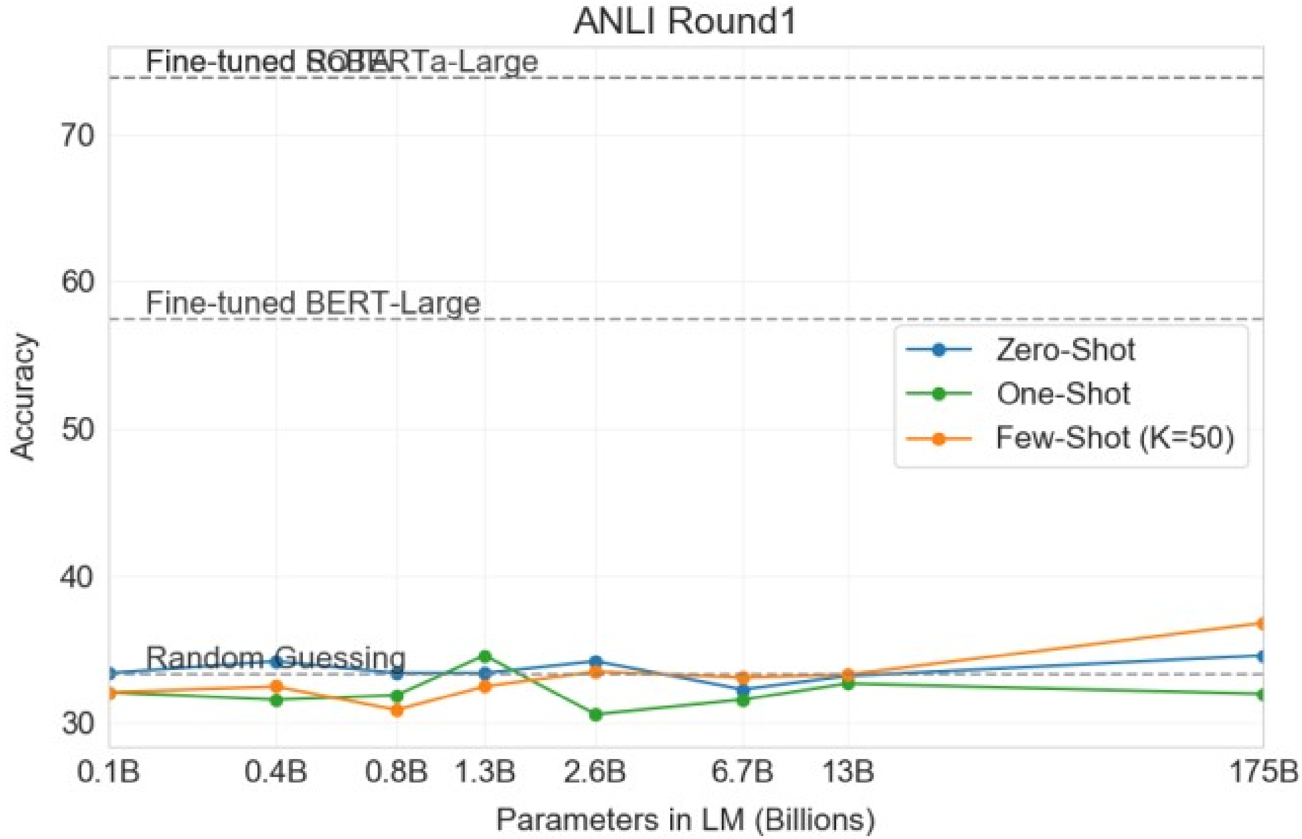
Total: -0.10 logprob on 1 tokens
(99.10% probability covered in top 5 logits)

OpenAI Playground DaVinci

Language Models are Few-Shot Learners (E

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts fails



Language Models are Few-Shot Learners (Brown et al. 2020)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts fails

"These results on both RTE and ANLI suggest that NLI is still a very difficult task for language models"

Language Models are Few-Shot Learners (Brown et al. 2020)

Also confirmed for PaLM 540B

- Results only improved when fine-tuning on NLI data

PaLM: Scaling Language Modeling with Pathways (Chowderhy et al. 2022)

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts fails

Diagnostic NLI dataset:

Tags	Sentence 1	Sentence 2	Fwd	Bwd
Lexical Entailment (Lexical Semantics), Downward Monotone (Logic)	The timing of the meeting has not been set, according to a Starbucks spokesperson.	The timing of the meeting has not been considered, according to a Starbucks spokesperson.	N	E
Universal Quantifiers (Logic)	Our deepest sympathies are with all those affected by this accident.	Our deepest sympathies are with a victim who was affected by this accident.	E	N
Quantifiers (Lexical Semantics), Double Negation (Logic)	I have never seen a hummingbird not flying.	I have never seen a hummingbird.	N	E

(Wang et al., 2019) Also used at SuperGlue leaderboard

Textual Entailment (RTE), Natural Language Inference (NLI)

GPT-3 using prompts fails

Diagnostic NLI dataset:

Double Negation: 0.0
Morphological Negation: 0.0
Anaphora/Coreference: 1.7
Nominalization: 2.6
Downward Monotone: 3.6
Conjuction: 4.0
Existential: 6.1
Disjunction: 7.4
Logic: 10.6
Negation: 11.6
Temporal: 12.4

Quantifiers: 59.5
Restrictivity: 48.5
Intersectivity: 41.4
Universal: 39.6
Active/Passive: 34.5
Knowledge: 32.0
World Knowledge: 33.0
Factivity: 31.6
Lexical Semantics: 30.0
Common Sense: 28.4

Matthew Correlation Score, from SuperGlue leaderboard

Overcoming limitations of LM

LMs fail on many inferences in NLI datasets

Our hypothesis:

Fine-tuning LMs on NLI datasets
allow LMs to learn certain inferences ...

… which the LMs will apply on target tasks

Entailment as Few-Shot Learner (Wang et al. 2021)

Plan for this session

- Pre-trained LM
- Prompting
- Entailment
- **Few-shot Information Extraction**

Few-shot Information Extraction?

Our proposal:

- Use “smaller” masked language models
- Additional pre-training with NLI datasets => Entailment Models
- Recast tasks into text:hypothesis pairs
- Run entailment model (zero-shot)
- Fine-tune entailment model (few-shot, full train)

Few-shot Information Extraction?

Our proposal:

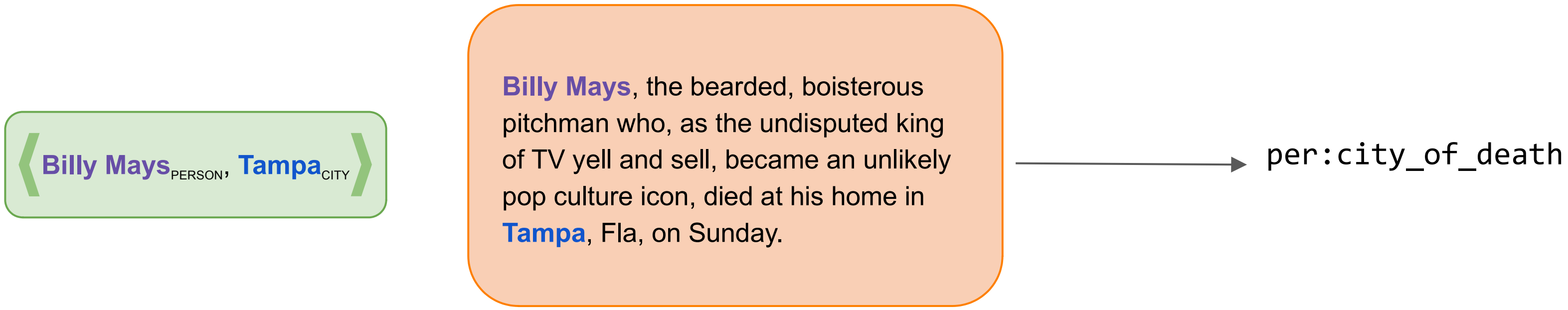
- Use “small” pre-trained language models
- Additional pre-training with NLI datasets => Entailment Models
- Recast tasks into text:hypothesis pairs
- Run entailment model (zero-shot)
- Fine-tune entailment model (few-shot, full train)

We will examine our work on:

- Relation extraction (Sainz et al 2021, EMNLP)
- Event-argument extraction (Sainz et al. 2022, NAACL findings)
- Several IE tasks (Sainz et al. 2022, NAACL demo)

Entailment for prompt-based Relation Extraction (Sainz et al 2021, EMNLP)

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



Entailment for prompt-based Relation Extraction

Given 2 entities e1 and e2 and a context c,
predict the schema 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.

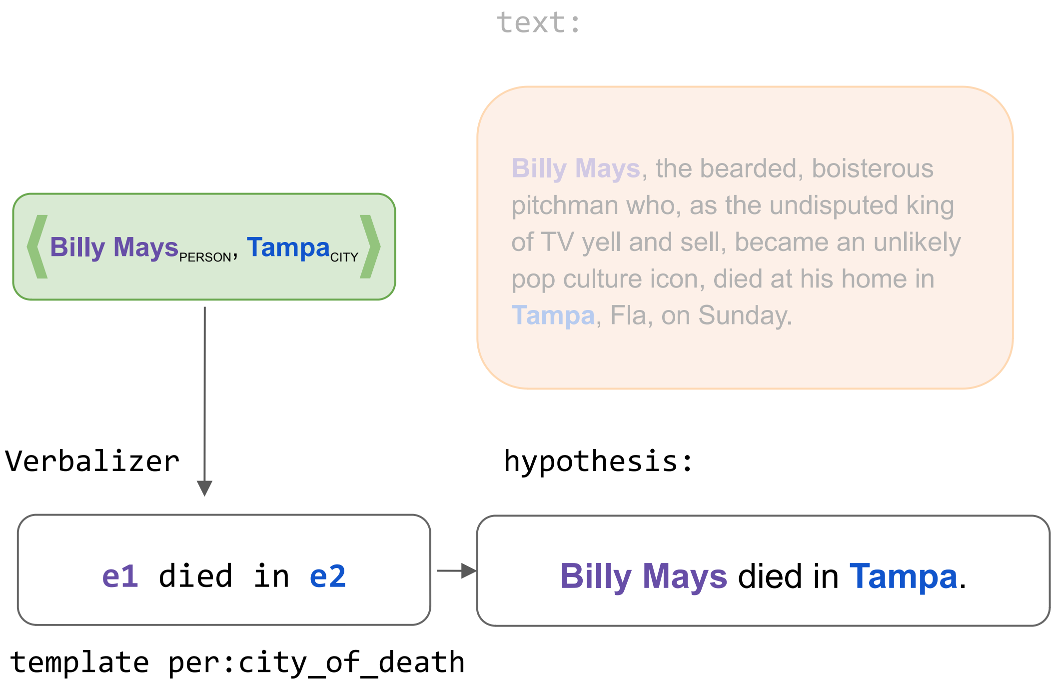
`Verbalizer`

`e1 died in e2`

`template per:city_of_death`

Entailment for prompt-based Relation Extraction

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



Entailment for prompt-based Relation Extraction

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

text:

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.

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

Verbalizer

hypothesis:

`e1 died in e2`

Billy Mays died in **Tampa**.

template per:city_of_death

Entailment for prompt-based Relation Extraction

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

text:

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.

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

Verbalizer

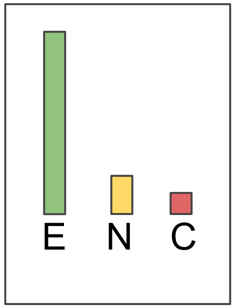
`e1 died in e2`

template per:city_of_death

hypothesis:

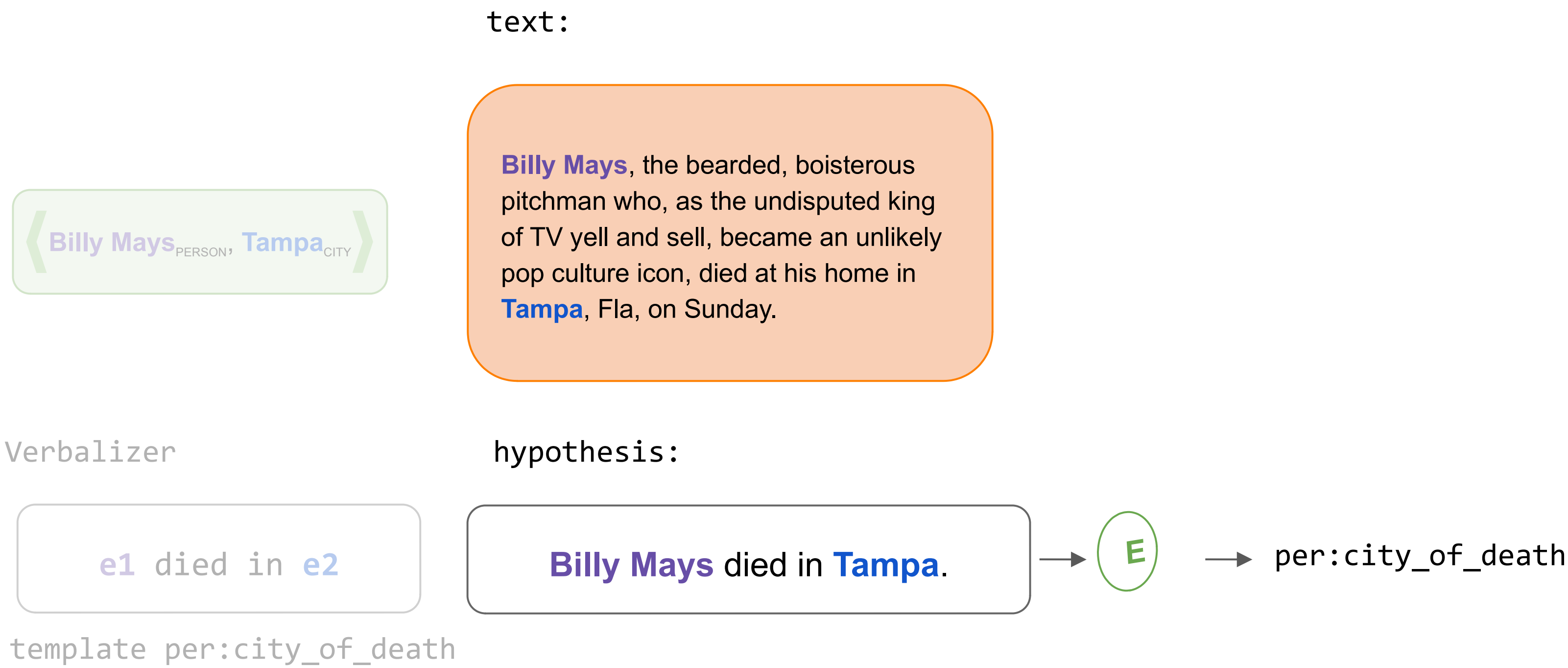
Billy Mays died in **Tampa**. →

Run fine-tuned entailment model



Entailment for prompt-based Relation Extraction

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

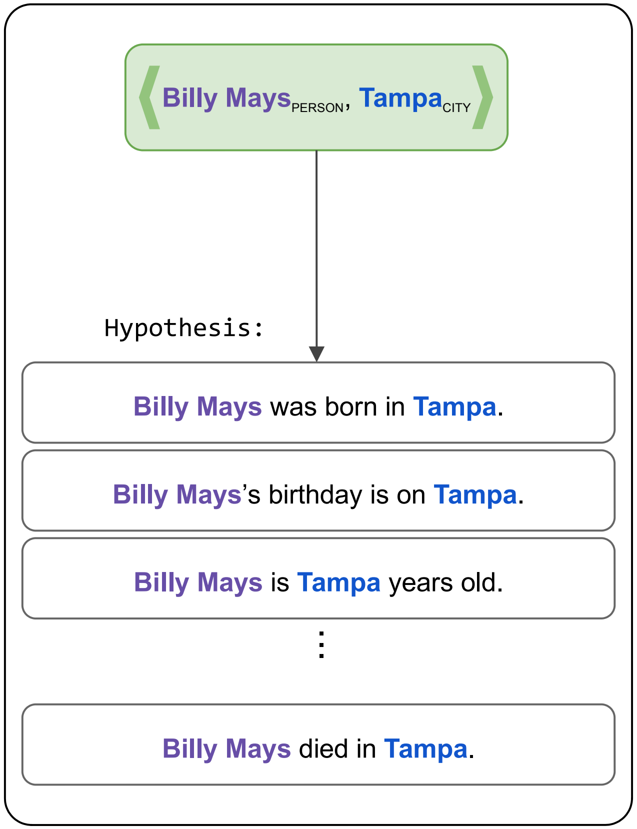


Entailment for prompt-based Relation Extraction

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 for prompt-based Relation Extraction

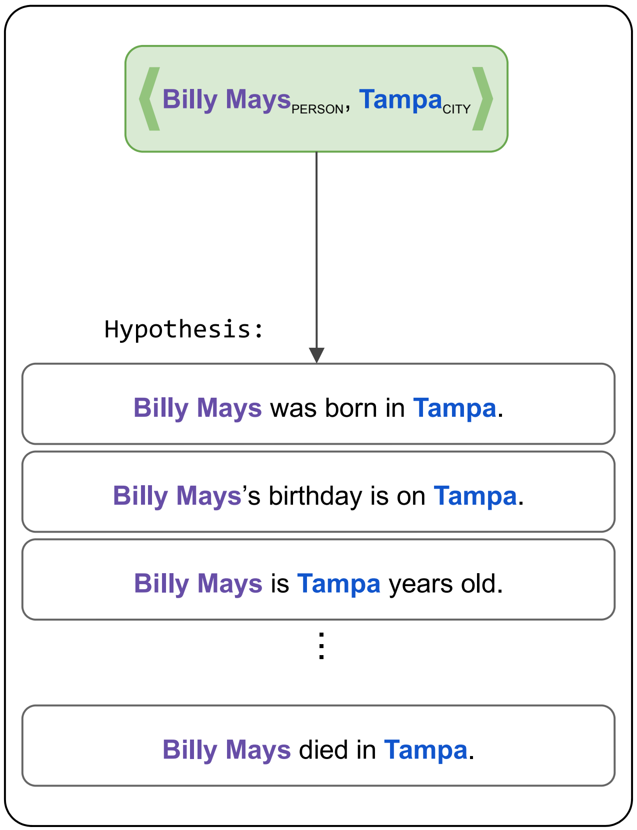
Verbalizer



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 for prompt-based Relation Extraction

Verbalizer



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

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

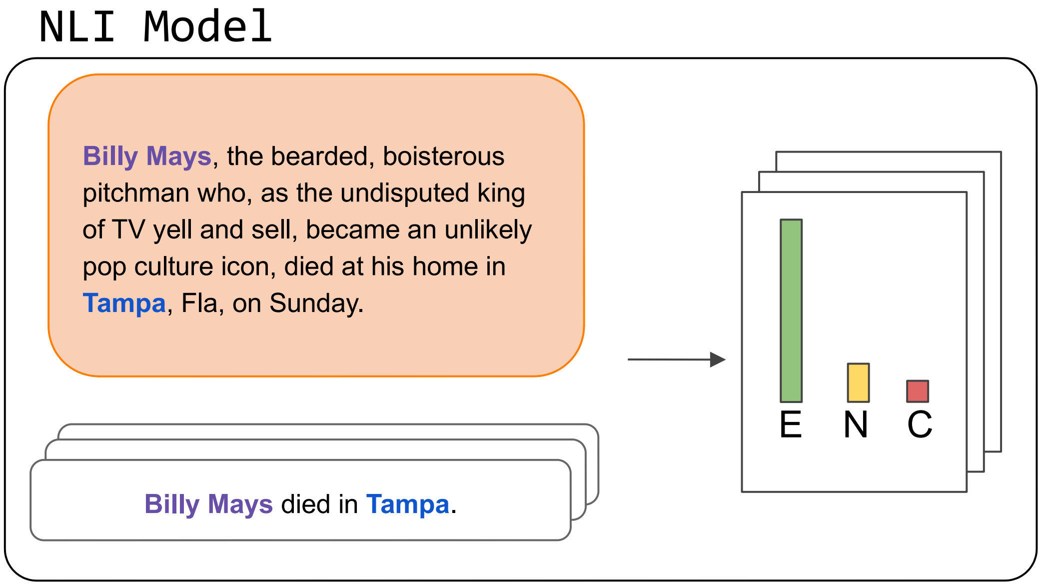
- N:M relation between templates and relations
- Also, type constraints for entities

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 for prompt-based Relation Extraction

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

$$P_{NLI}(x, hyp)$$



Entailment for prompt-based Relation Extraction

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

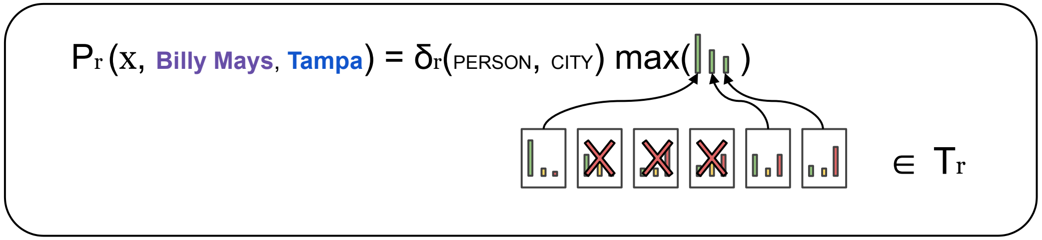
- We compute the probability of relation r based on the hypothesis probabilities and entity constraints:

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

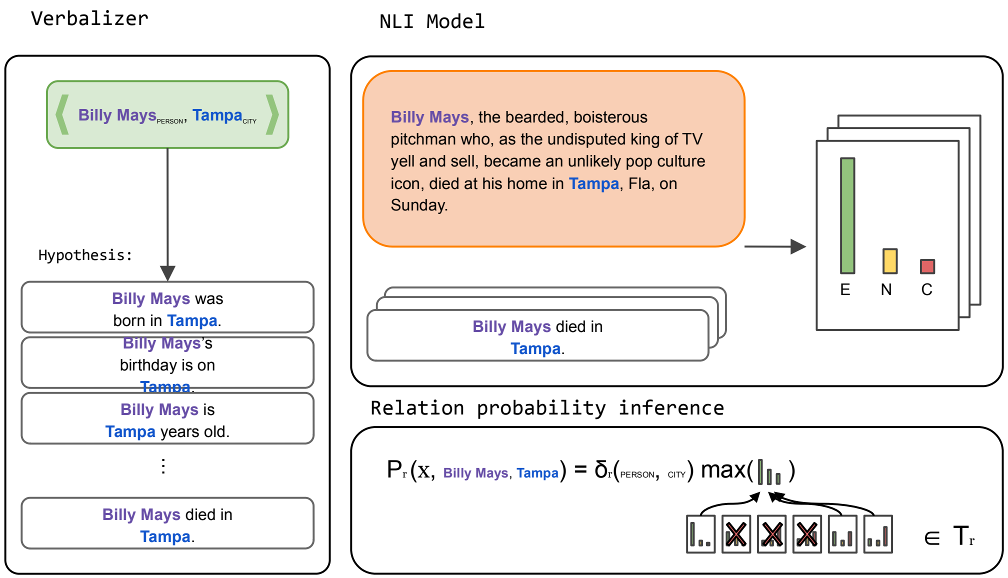
- The δ_r function describes the entity constraints of the relation r:

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

Relation probability inference



Entailment for prompt-based Relation Extraction

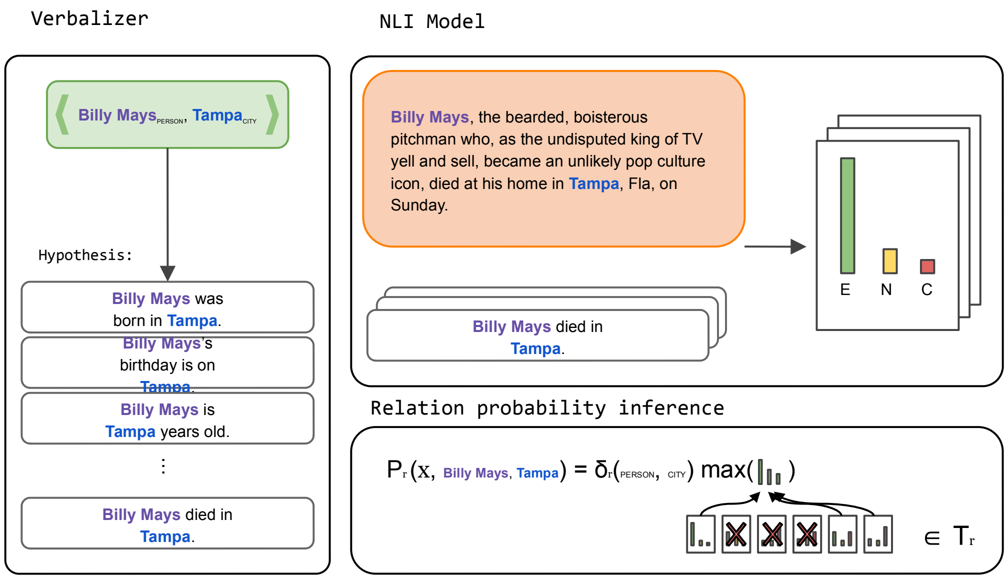


Finally, we return the relation with the highest probability:

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

If no relation is entailed, then *r* = **`no_relation`**

Entailment for prompt-based Relation Extraction



Finally, we return the relation with the highest probability:

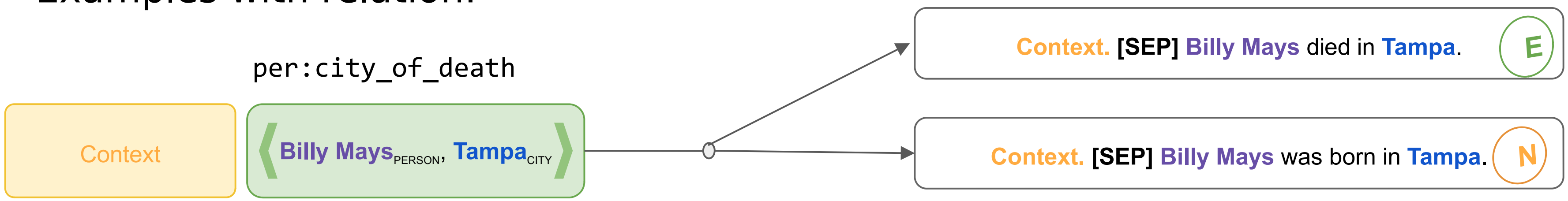
$$\hat{r} = \arg\max_{r \in R} \mathsf{P}_r(x, x_{e1}, x_{e2})$$

If no relation is entailed, then r = **`no_relation`**

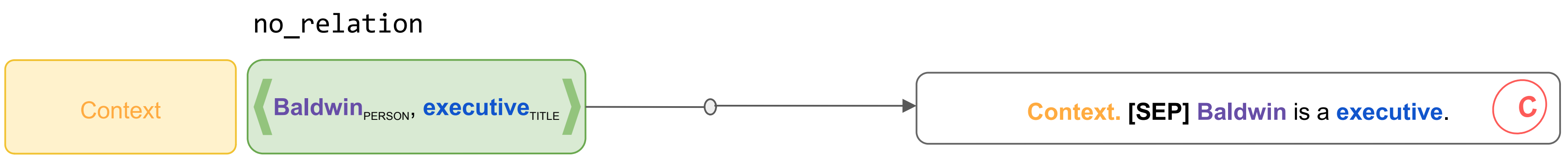
ZERO-SHOT

Fine-tuning with prompted Relation Extraction dataset

Examples with relation:

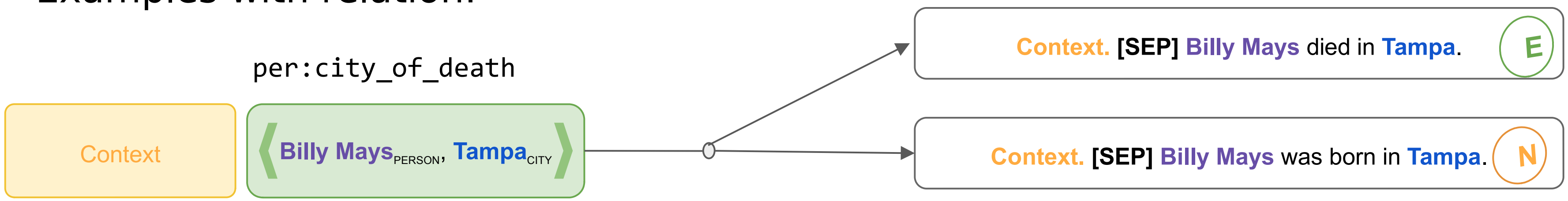


Examples with no relation:

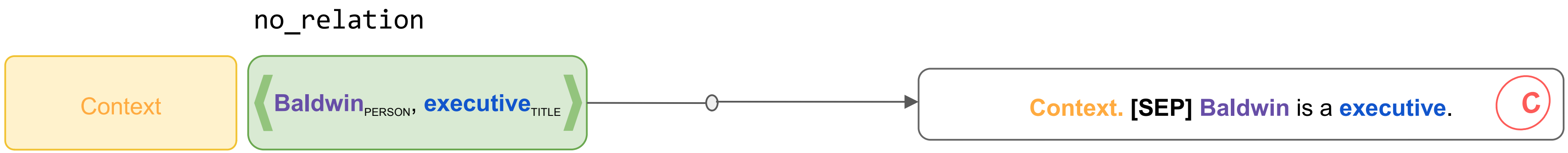


Fine-tuning with prompted Relation Extraction dataset

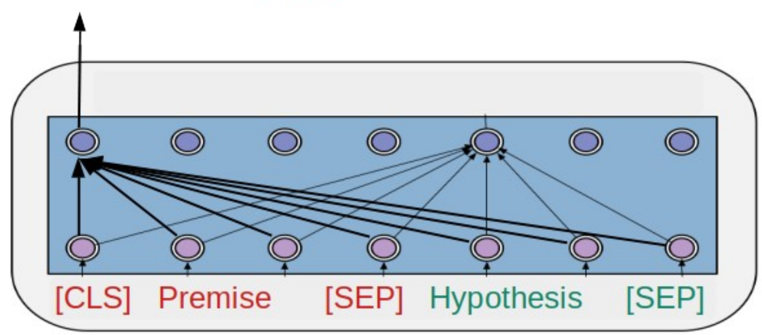
Examples with relation:



Examples with no relation:



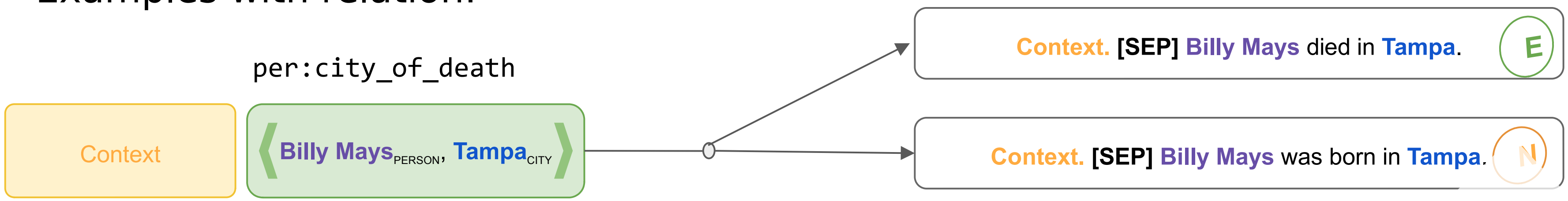
Fine-tune MLM with prompted examples



Fine-tuning with prompted Relation Extraction dataset

Examples with relation:

per:city_of_death



Examples with no relation:

no_relation

Fine-tune MLM with prompted examples

Evaluation dataset

TACRED (Zhang et al., 2017), based on TAC

41 relation labels (positive), no relation (negative).

Training:

- Zero-shot: 0 examples
- Few-shot:
  - 5 examples per class (1%)
  - 23 examples per class (5%)
  - 46 examples per class (10%)
- Full-train: 460 examples per class

Evaluation: 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

Evaluation: zero-shot

NLI Model	# Param.	MNLI Acc.	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:

- Best results with DeBERTa

Evaluation: zero-shot

NLI Model	# Param.	MNLI Acc.	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:

- Best results with DeBERTa
- Note that minor variations in MNLI (±2) produce large variations in F1.

Evaluation: 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.

Evaluation: 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
$NLI_{RoBERTa}$ (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
$NLI_{DeBERTa}$ (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.
- Our systems large improvements over SOTA systems. **1% > 10%**
- DeBERTa model score the best.

Entailment for prompt-based Event Argument Extraction (Sainz et al. 2022, NAACL)

Given the success on Relation Extraction, we extended the work:

- Check Event Argument Extraction
- Transfer knowledge across event schemas (ACE, Wikievents)
- Measure effect of different NLI datasets
- Measure domain-expert hours

Entailment for prompt-based Event Argument Extraction

Given **event** $\mathbf{e}$ and **argument candidate** $\mathbf{a}$ and a **context** $\mathbf{c}$, predict the **argument relation** (if any) holding between the event and candidate in the context.

⟨hired$_{\text{START-POSITION}}$, John D. Idol$_{\text{PERSON}}$⟩

In 1997, the company **hired** **John D. Idol** to take over as chief executive.

→ `Start-Position:Person`

Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any) holding between the event and candidate in the context.

text:

In 1997, the company **hired** **John D. Idol** to take over as chief executive.

⟨hired$_{\text{START-POSITION}}$, John D. Idol$_{\text{PERSON}}$⟩

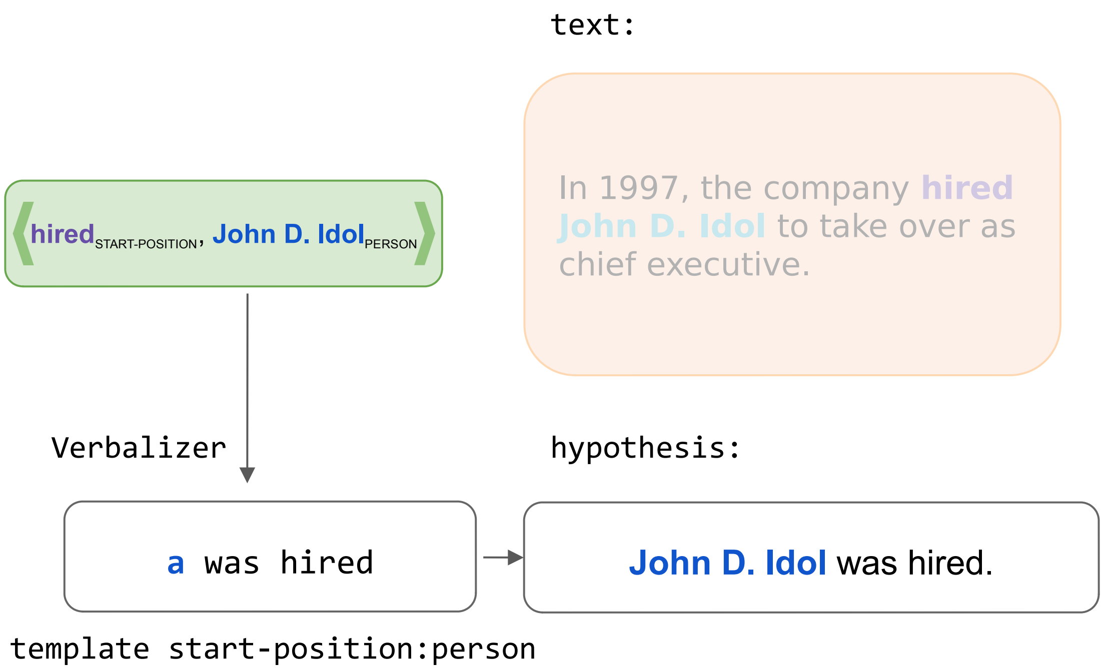
Verbalizer

```
a was hired
```

template start-position:person

Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any) holding between the event and candidate in the context.



Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any) holding between the event and candidate in the context.

hired$_{\text{START-POSITION}}$, John D. Idol$_{\text{PERSON}}$

text:

In 1997, the company **hired** **John D. Idol** to take over as chief executive.

Verbalizer

a was hired

template start-position:person

hypothesis:

John D. Idol was hired.

Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any) holding between the event and candidate in the context.

⟨hired$_{\text{START-POSITION}}$, John D. Idol$_{\text{PERSON}}$⟩

`text:`

In 1997, the company **hired** **John D. Idol** to take over as chief executive.

`Verbalizer`

`a was hired`

`template start-position:person`

`hypothesis:`

John D. Idol was hired. → Run fine-tuned entailment model

Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any) holding between the event and candidate in the context.

text:

In 1997, the company **hired** **John D. Idol** to take over as chief executive.

⟨hired START-POSITION, John D. Idol PERSON⟩

Verbalizer

a was hired

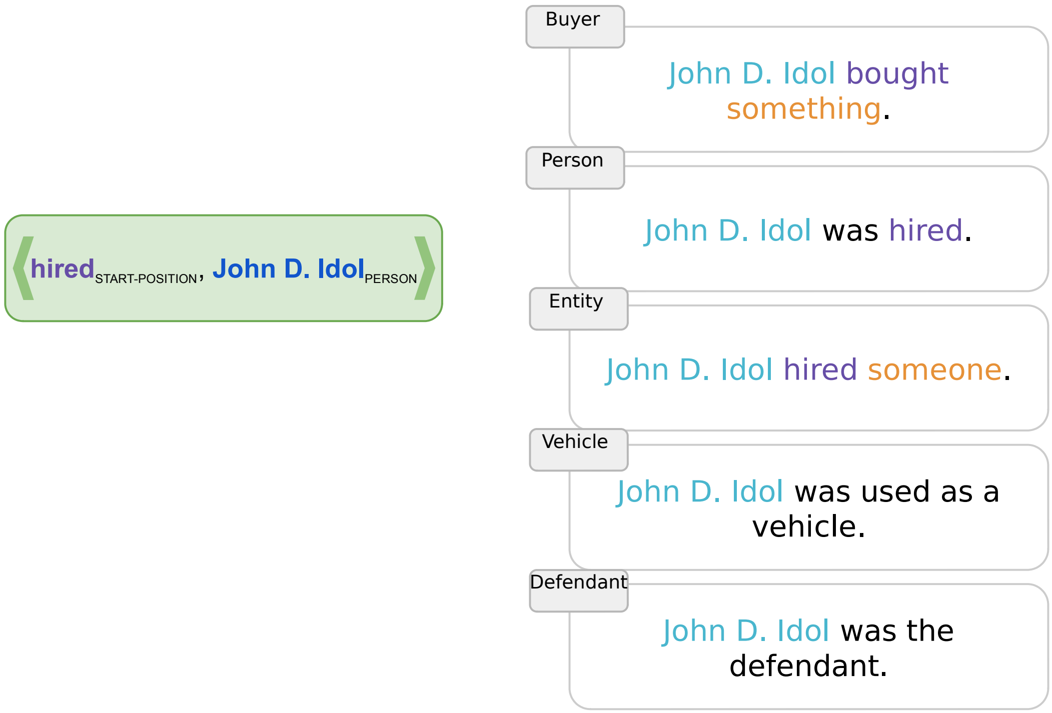
template start-position:person

hypothesis:

John D. Idol was hired. → E → start-position:person

Entailment for prompt-based Event Argument Extraction

Given **event** e and **argument candidate** a and a **context** c, predict the **argument relation** (if any)
holding between the event and candidate in the context.



Evaluation datasets

ACE (Walker et al., 2006). 22 argument types.

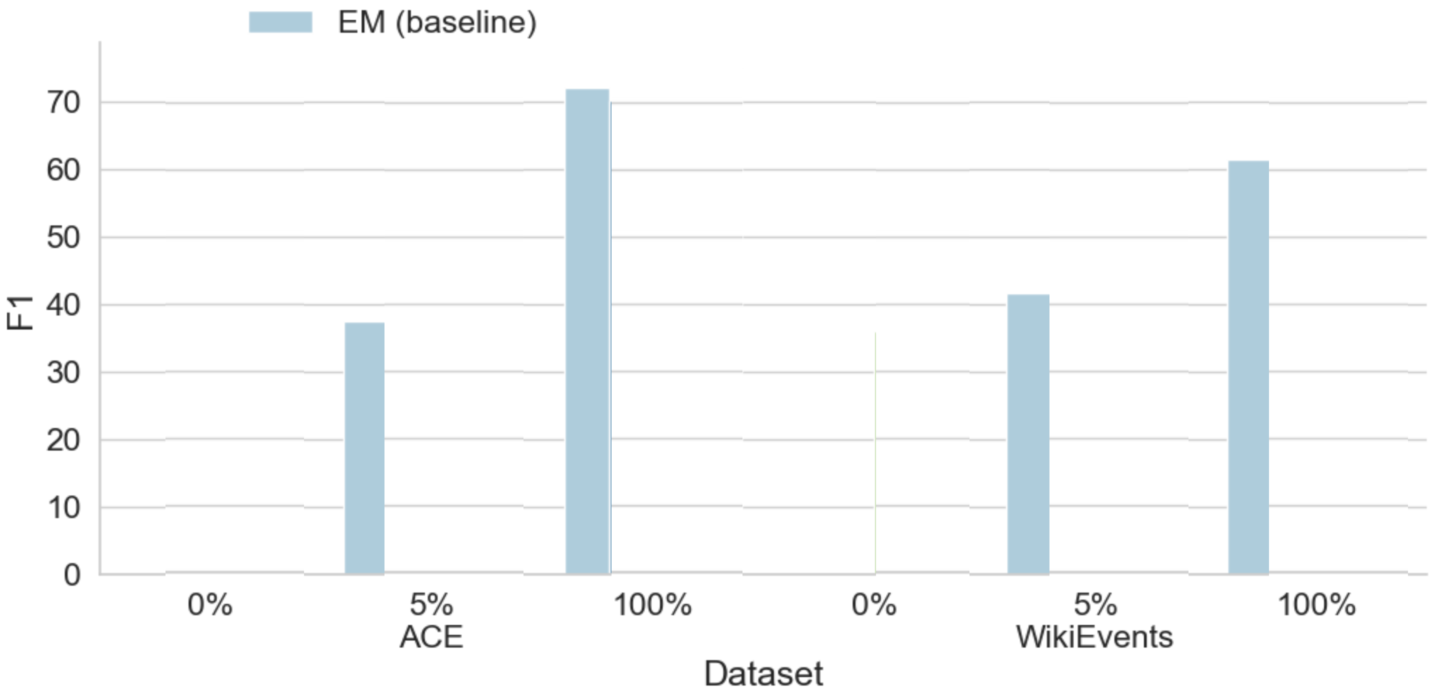
WikiEvents (Li et al., 2021). 59 argument types.

Training (ACE / Wikievents):

- Zero-shot: 0 examples
- Few-shot: 11 / 4 examples per class (5%)
- Full-train: 220 / 80 examples per class (100%)

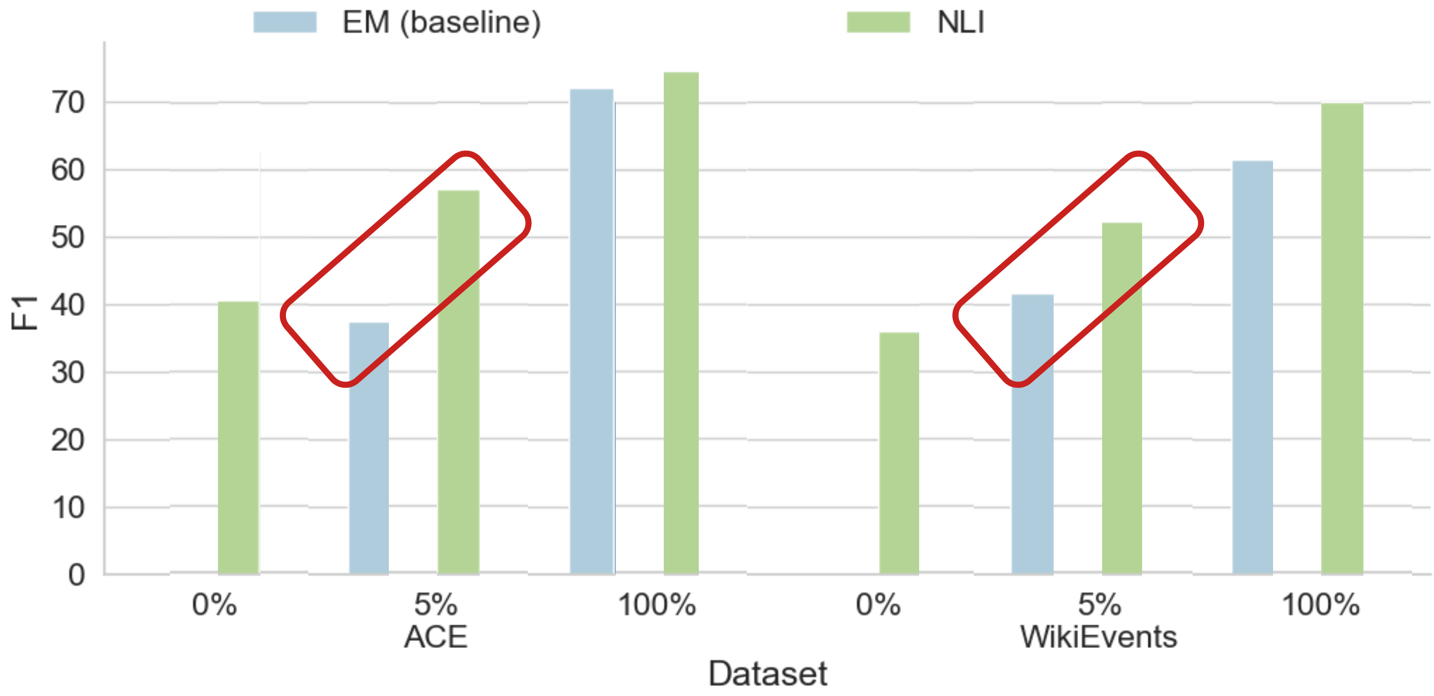
Evaluation: ACE and Wikievents

- EM is a fine-tuned RoBERTa (baseline)



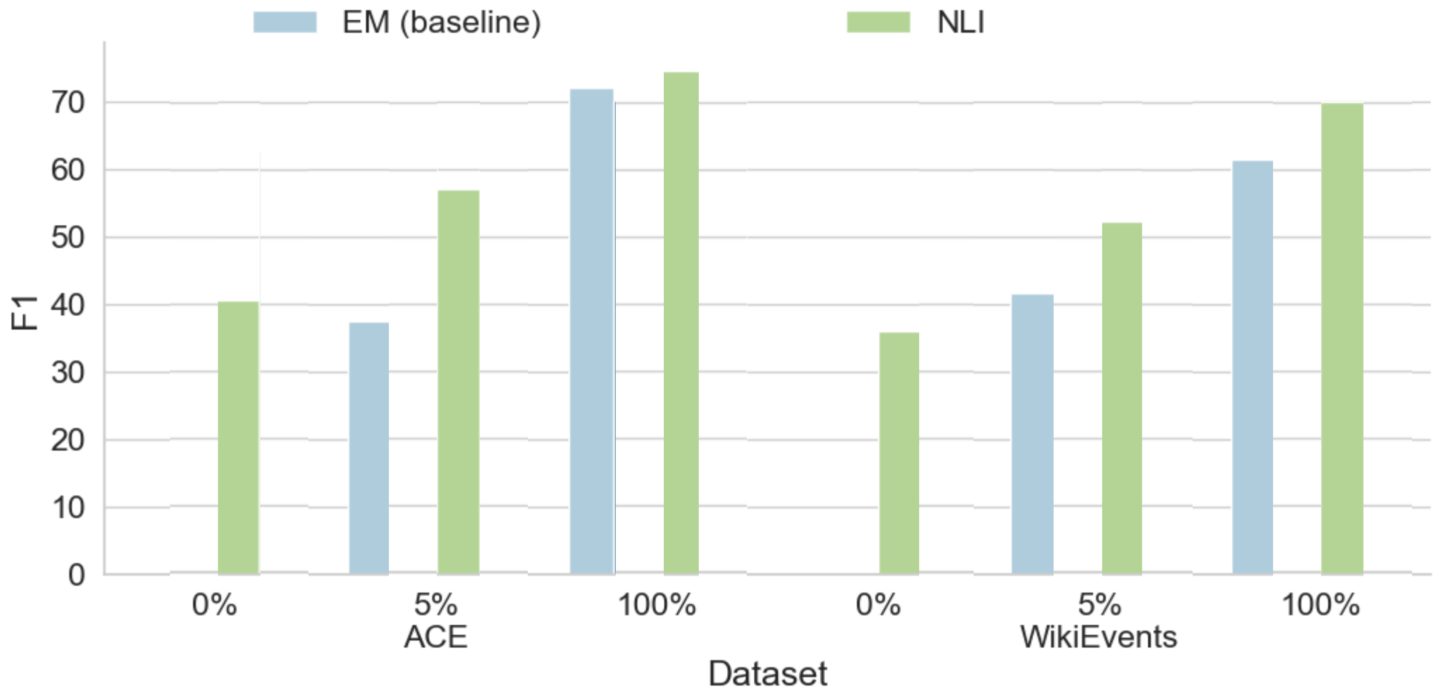
Evaluation: ACE and Wikievents

- NLI is our entailment-based system (RoBERTa)



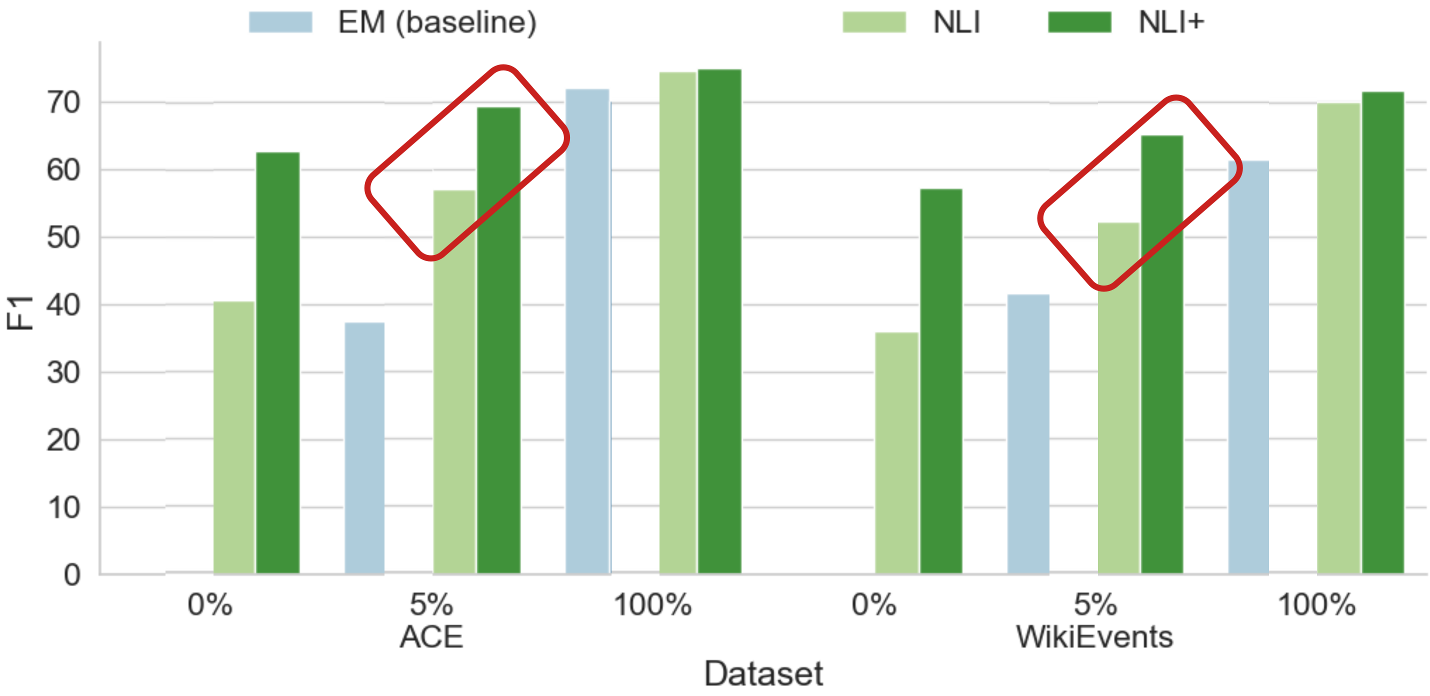
Can we transfer between schemas (ACE <=> WikiEvents)

- NLI+: pre-train on other schema (Wikievents or ACE respectively)



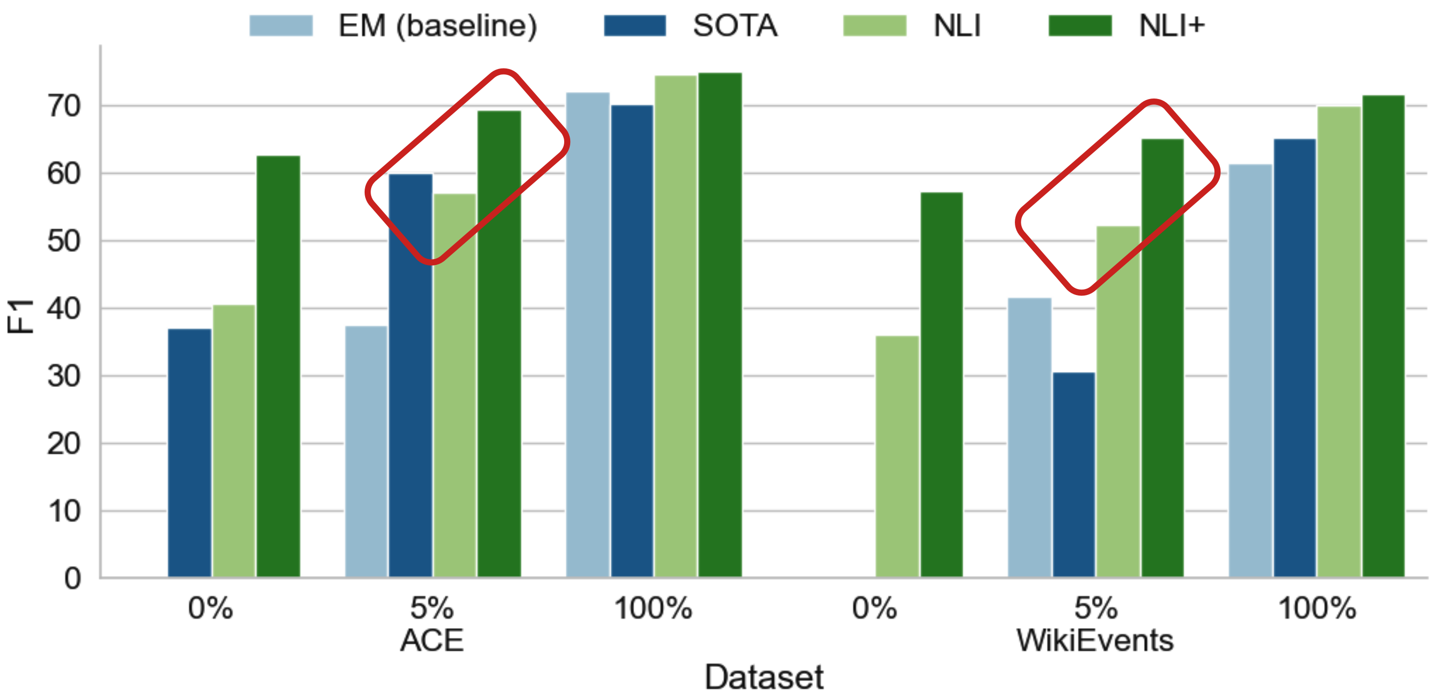
Can we transfer between schemas (ACE <=> WikiEvents)

- NLI+: pre-train on other schema (Wikievents or ACE respectively)



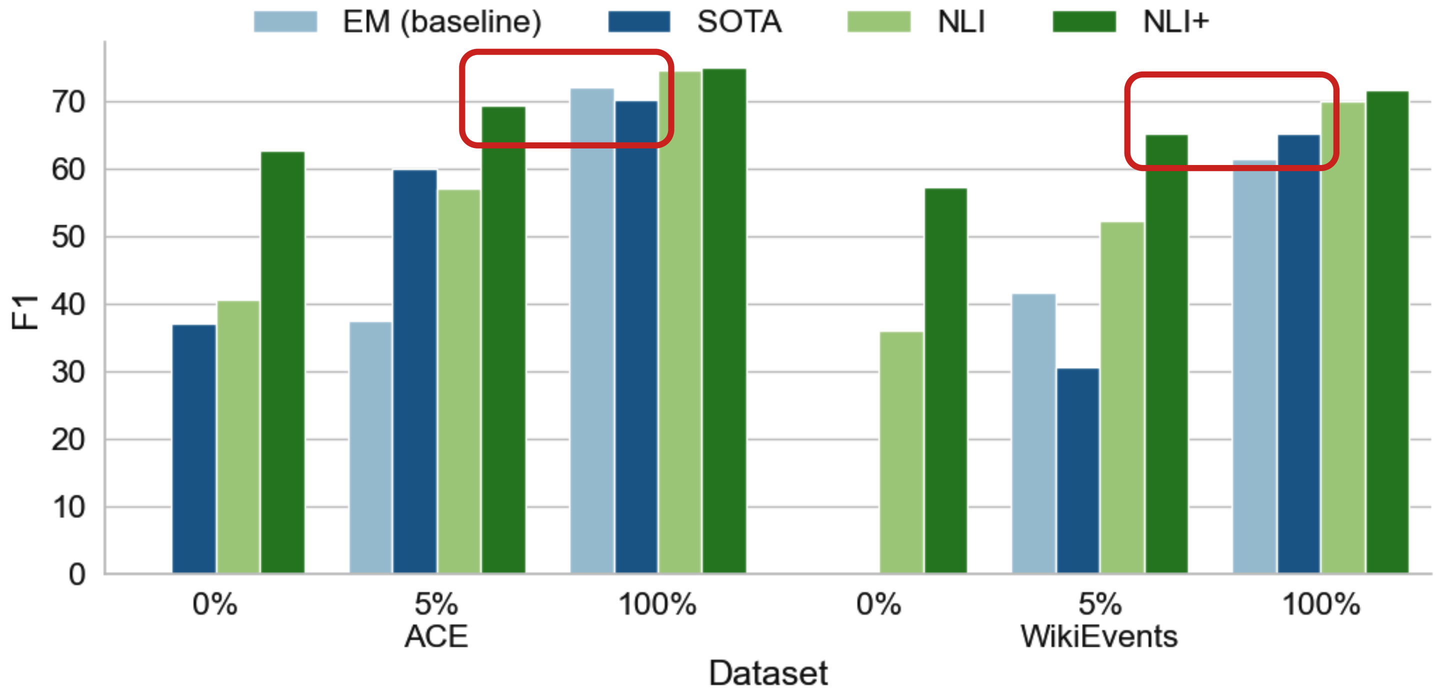
Evaluation: ACE and Wikievents

- We beat SOTA, thanks to entailment, schema transfer

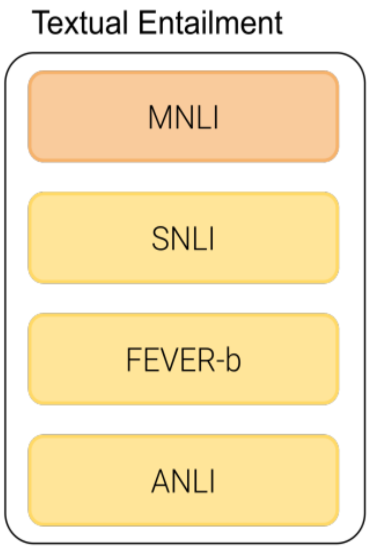


Evaluation: ACE and Wikievents

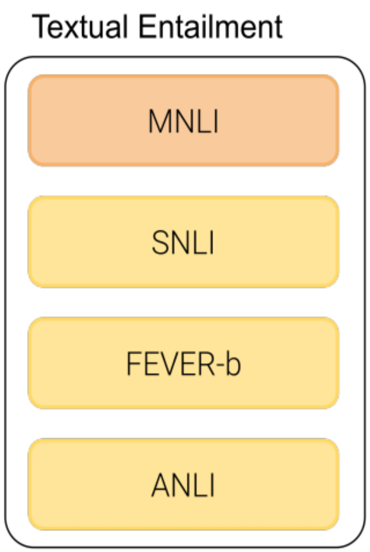
- We beat SOTA, thanks to entailment, schema transfer
- Reach full-train with only 5% of the annotations

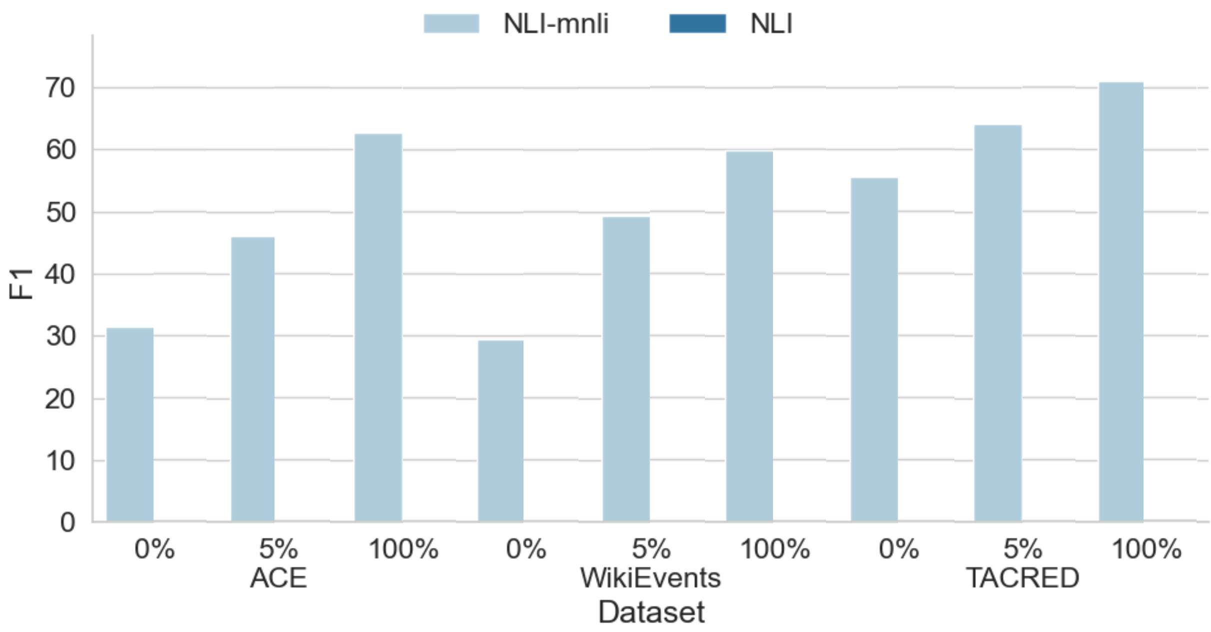


The more NLI pre-training the better



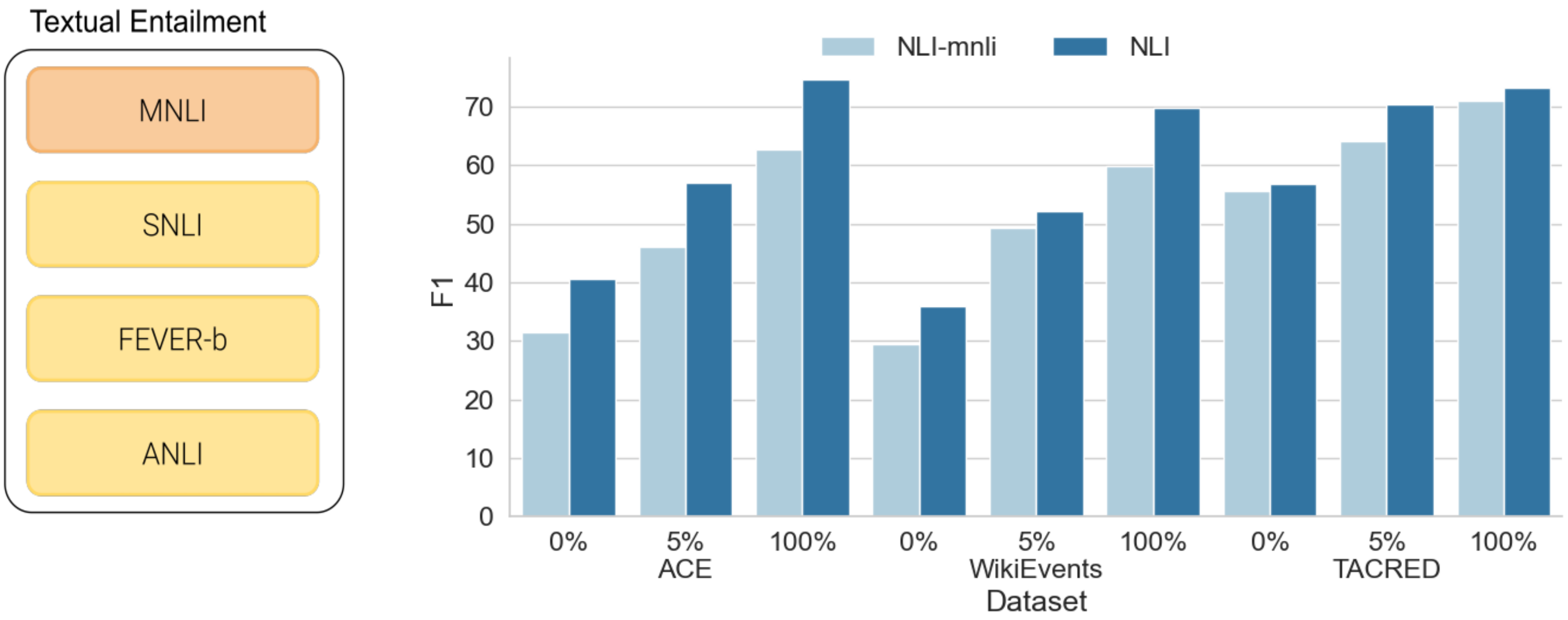
The more NLI pre-training the better





Combining several NLI training data helps (also in TACRED)

The more NLI pre-training the better



Combining several NLI training data helps (also in TACRED)

Is all dependent on the domain-expert?

- We gave the task to a **computational linguist** PhD
  - Very similar results across all training regimes
  - Replicable, robust to variations in prompts

- She found prompt writing friendly:

  "Writing templates is more natural and rewarding than annotating examples, which is more repetitive, stressful and tiresome."

  "When writing templates, I was thinking in an abstract manner, trying to find generalizations. When doing annotation I was paying attention to concrete cases."

Is all dependent on the domain-expert?

- We gave the task to a **computational linguist** PhD
  - Very similar results across all training regimes
  - Replicable, robust to variations in prompts

- She found prompt writing friendly:

  "Writing templates is more natural and rewarding than annotating examples, which is more repetitive, stressful and tiresome."

  "When writing templates, I was thinking in an abstract manner, trying to find generalizations. When doing annotation I was paying attention to concrete cases."

What is the manual cost compared to annotation

- Time devoted by domain-expert in template writing:
  - Max. 15 minutes per argument
  - **ACE: 5 hours** for 22 argument types
  - WikiEvents: 12 hours for 59 argument types

- Estimate of time by domain-expert for annotation (under-estimation, no quality control, speed):
  - **ACE: 180 hours** for whole dataset (16,500 examples)

What is the manual cost compared to annotation

Two frameworks, **9 hours of domain-expert** effort (ACE):

1) Define, annotate, train: annotate 850 ex. (5%)

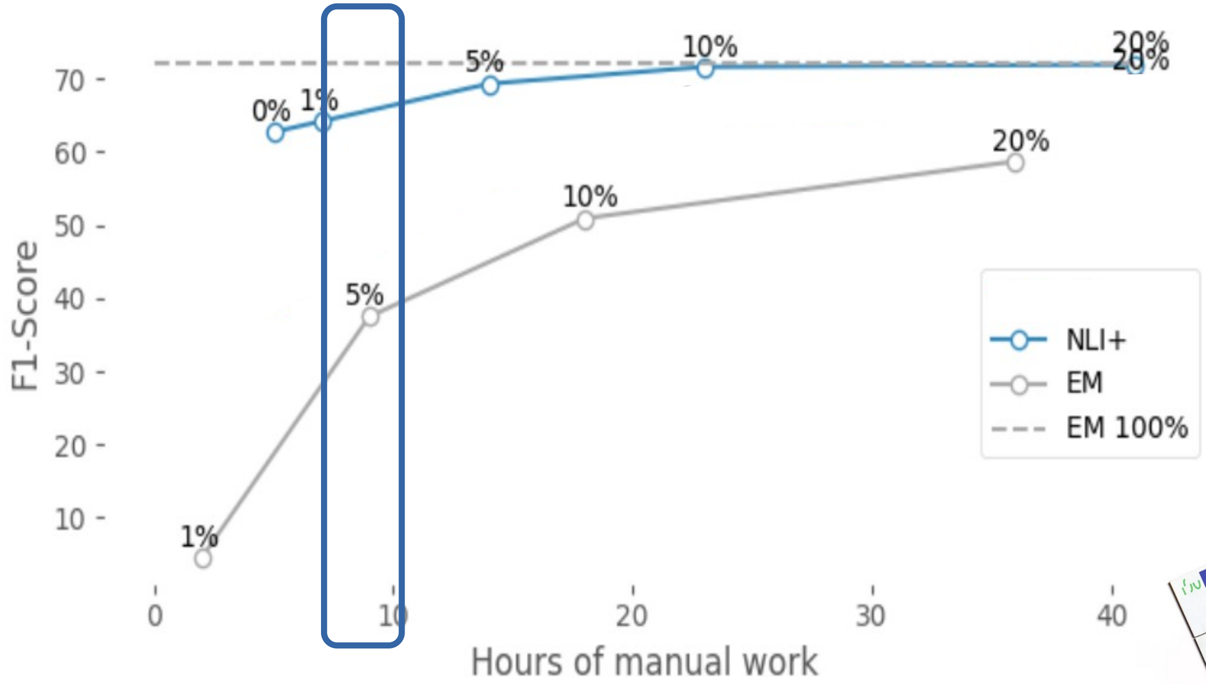
2) Verbalize while defining: prompts (5h), annotate 350 ex. (4h)

What is the manual cost compared to annotation

Two frameworks, **9 hours of domain-expert** effort (ACE):

1) Define, annotate, train: annotate 850 ex. (5%)

2) Verbalize while defining: prompts (5h), annotate 350 ex. (4h)



What is the manual cost compared to annotation

Two frameworks, **23 hours of domain-expert** effort (ACE):

1) Define, annotate, train: annotate 13%

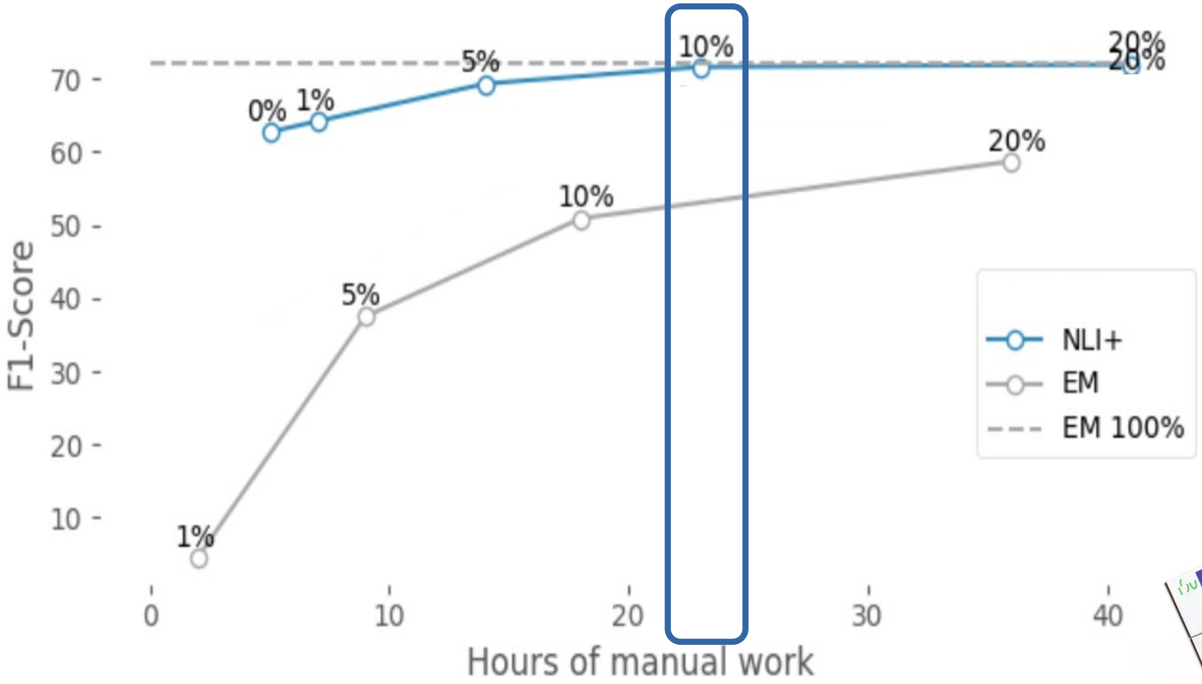
2) Verbalize while defining: prompts (5h), ann. 10% (18h)

What is the manual cost compared to annotation

Two frameworks, **23 hours of domain-expert** effort (ACE):

1) Define, annotate, train: annotate 13%

2) Verbalize while defining: prompts (5h), ann. 10% (18h)



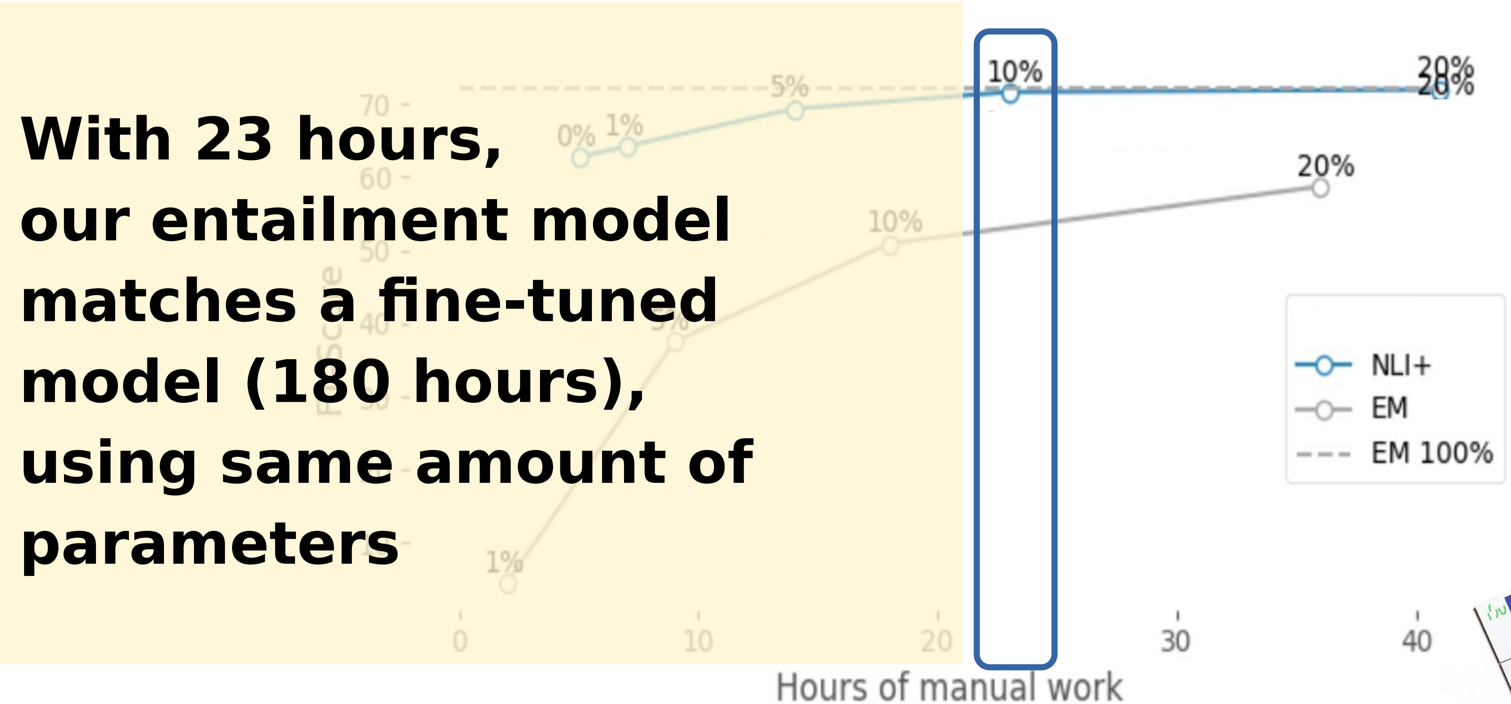
What is the manual cost compared to annotation

Two frameworks, **23 hours of domain-expert** effort (ACE):

1) Define, annotate, train: annotate 13%

2) Verbalize while defining: prompts (5h), ann. 10% (18h)

With 23 hours, our entailment model matches a fine-tuned model (180 hours), using same amount of parameters



Conclusions for prompt-based extraction using NLI

- Very effective for zero- and few-shot IE
- Allows for transfer across schemas (for the first time)
- 8 times less hours from domain-expert
- It is now feasible to build an IE system from scratch with limited effort
  - Develop schema and verbalization at the same time.
  - Verbalize then annotate a few examples

Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

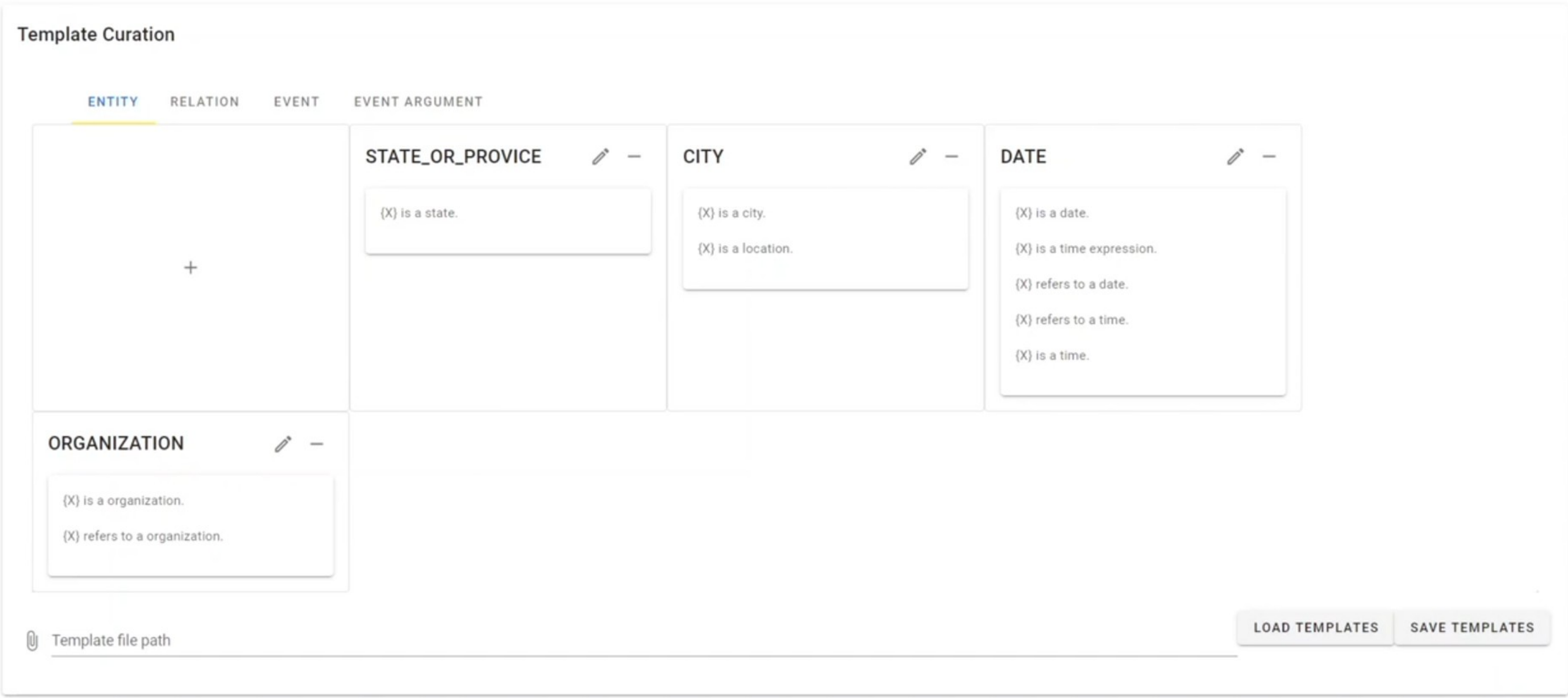
1) Domain expert defines entities and relations in English

2) Runs the definitions on examples

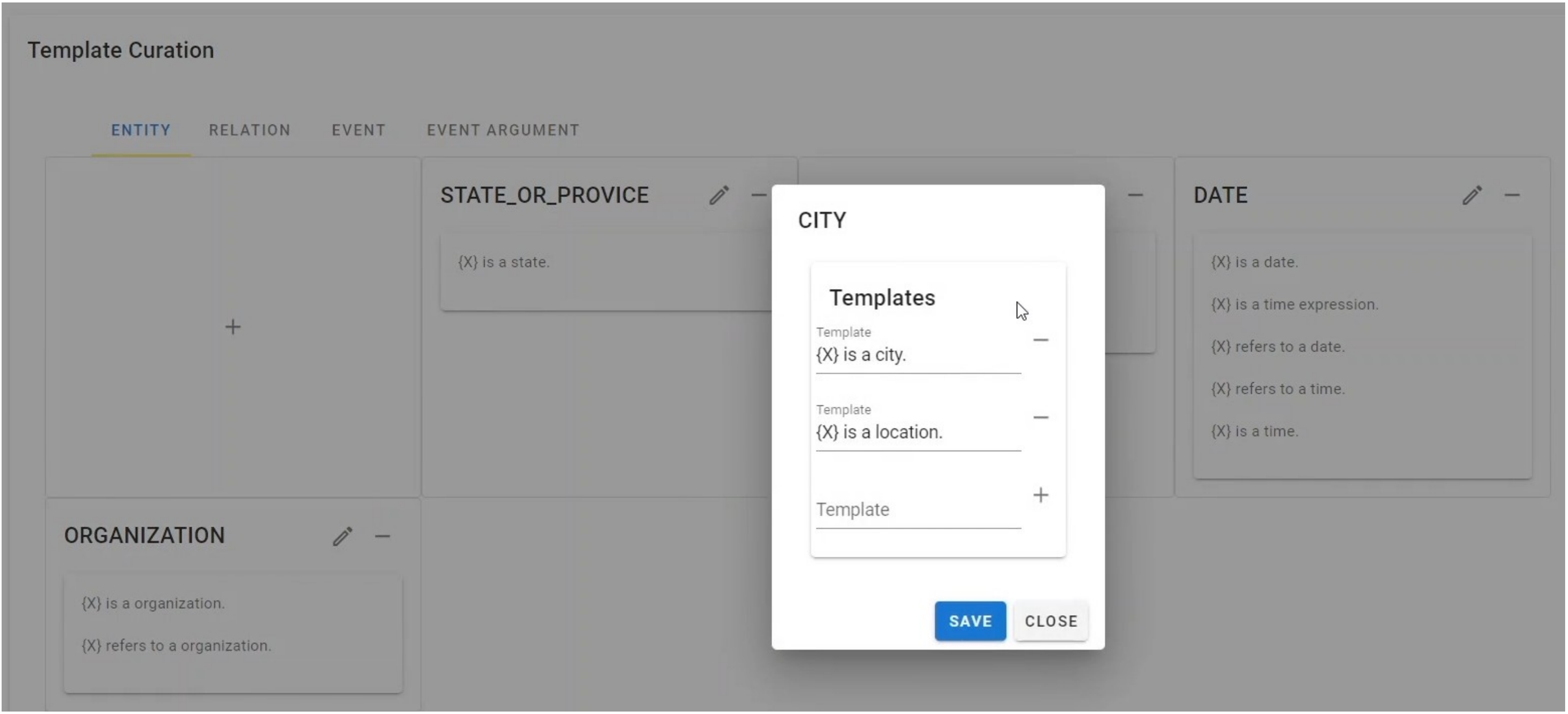
3) Annotates a handful of incorrect examples, iterates

- User interface for NERC, RE, EE, EAE
- 2 minute video

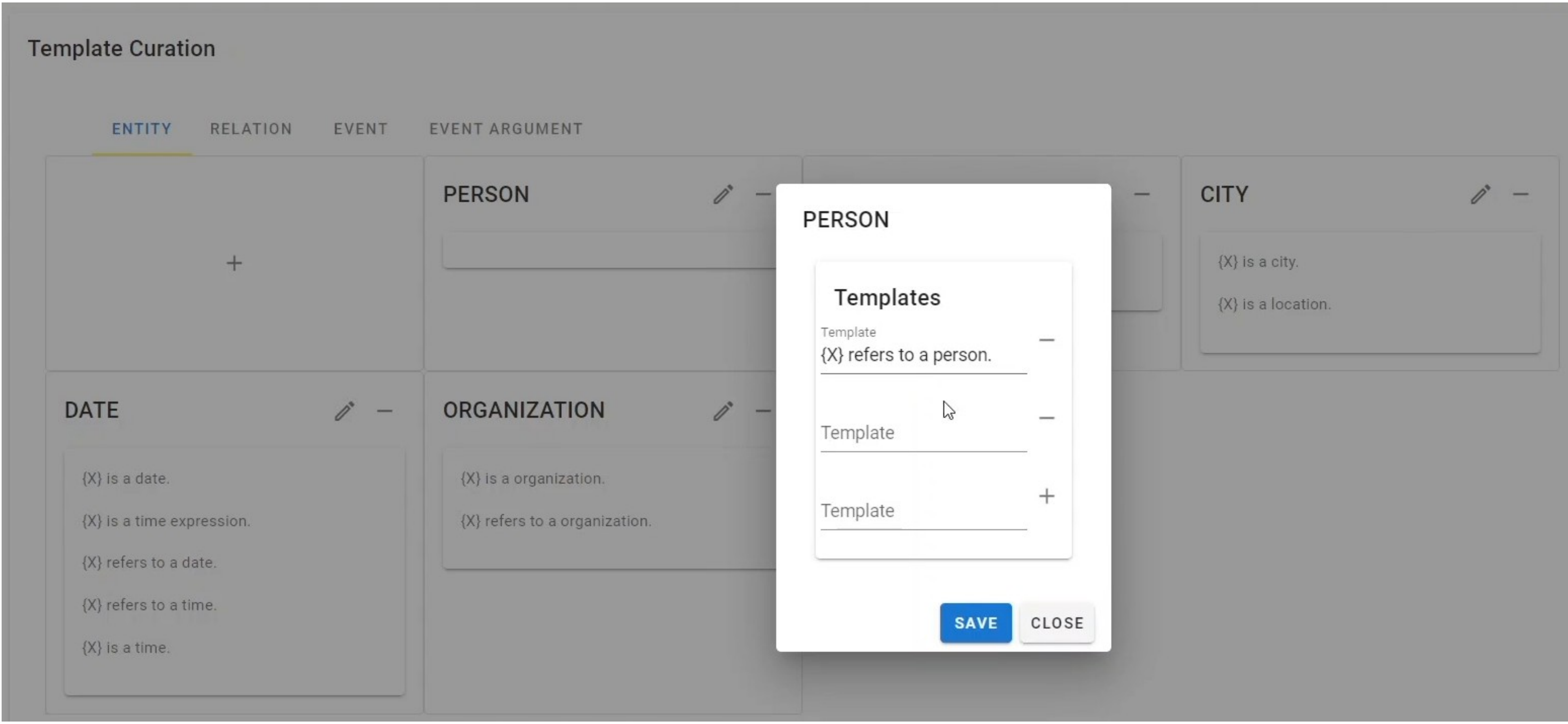
Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



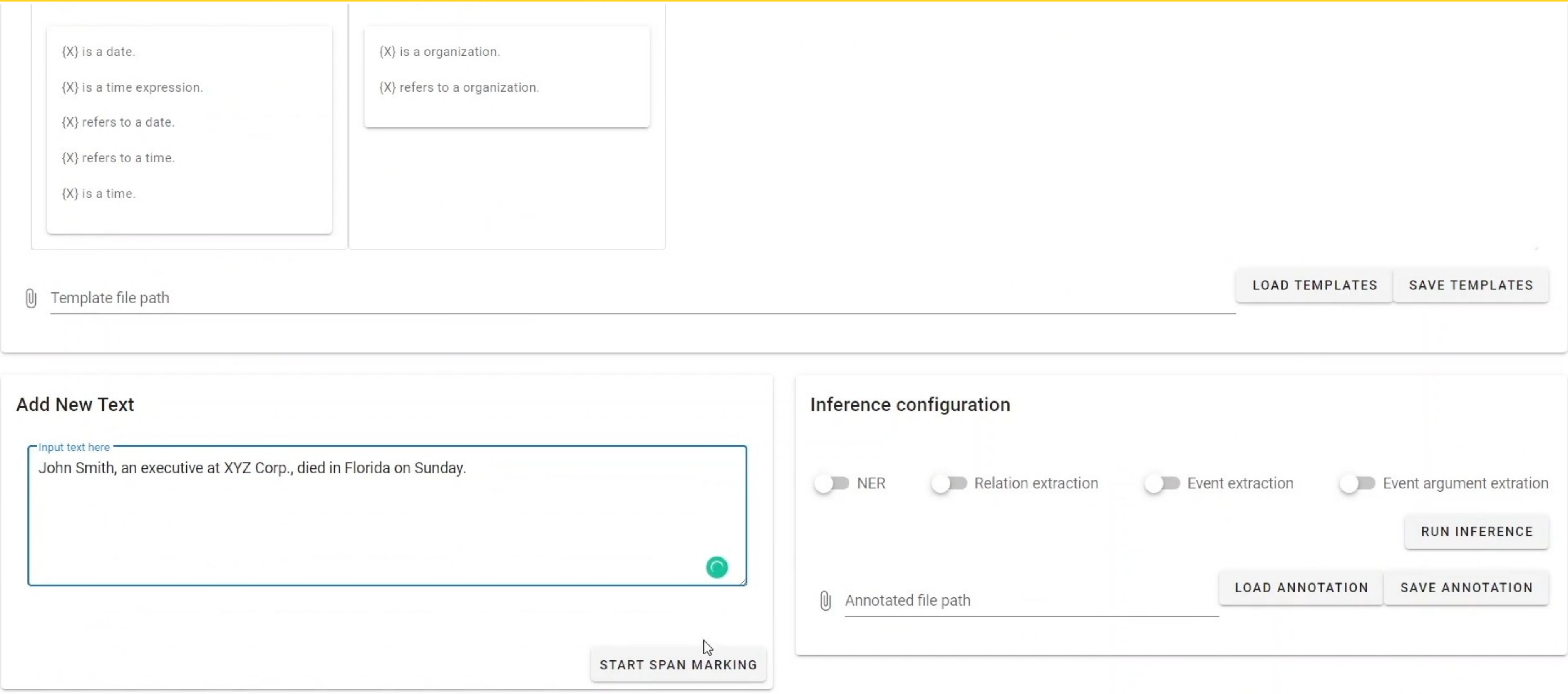
Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

NER

John Smith , an executive at XYZ Corp. , died in Florida on Sunday .

John Smith is a/an PERSON

Type	Template	Score
PERSON	{X} is a person.	0.991
ORGANIZATION	{X} refers to a organization.	0.955
PERSON	{X} refers to a person.	0.883

× − +

Sunday is a/an DATE

Type	Template	Score
DATE	{X} refers to a date.	0.867
DATE	{X} is a time expression.	0.733
DATE	{X} refers to a time.	0.721
PERSON	{X} refers to a person	0.665

Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

ORGANIZATION	{X} refers to a organization.	0.575

× − +

XYZ Corp. is a/an ORGANIZATION

Type	Template	Score
ORGANIZATION	{X} is a organization.	0.882
ORGANIZATION	{X} refers to a organization.	0.861

× − +

Florida is a/an CITY

Type	Template	Score
CITY	{X} is a location.	0.970
STATE_OR_PROVICE	{X} is a state.	0.636

× − +

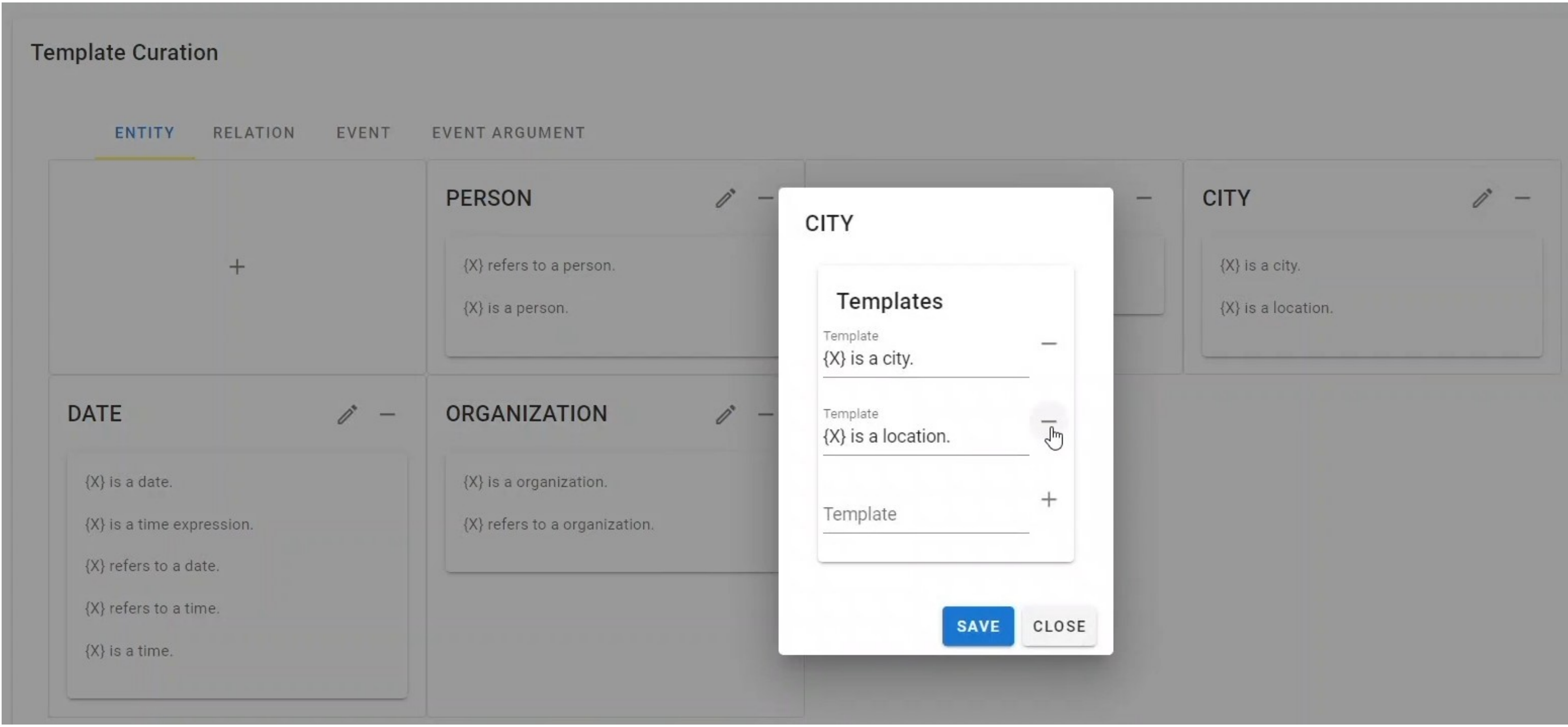
Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

Type	Total	Correct	Incorrect
PERSON	1	1 (1.000)	0 (0.000)
DATE	1	1 (1.000)	0 (0.000)
ORGANIZATION	1	1 (1.000)	0 (0.000)
CITY	1	0 (0.000)	1 (1.000)

Rows per page: 10 1-4 of 4

Type	Total	Correct	Incorrect
{X} is a person.	1	1 (1.000)	0 (0.000)
{X} refers to a organization.	3	3 (1.000)	0 (0.000)
{X} refers to a person.	2	2 (1.000)	0 (0.000)

Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

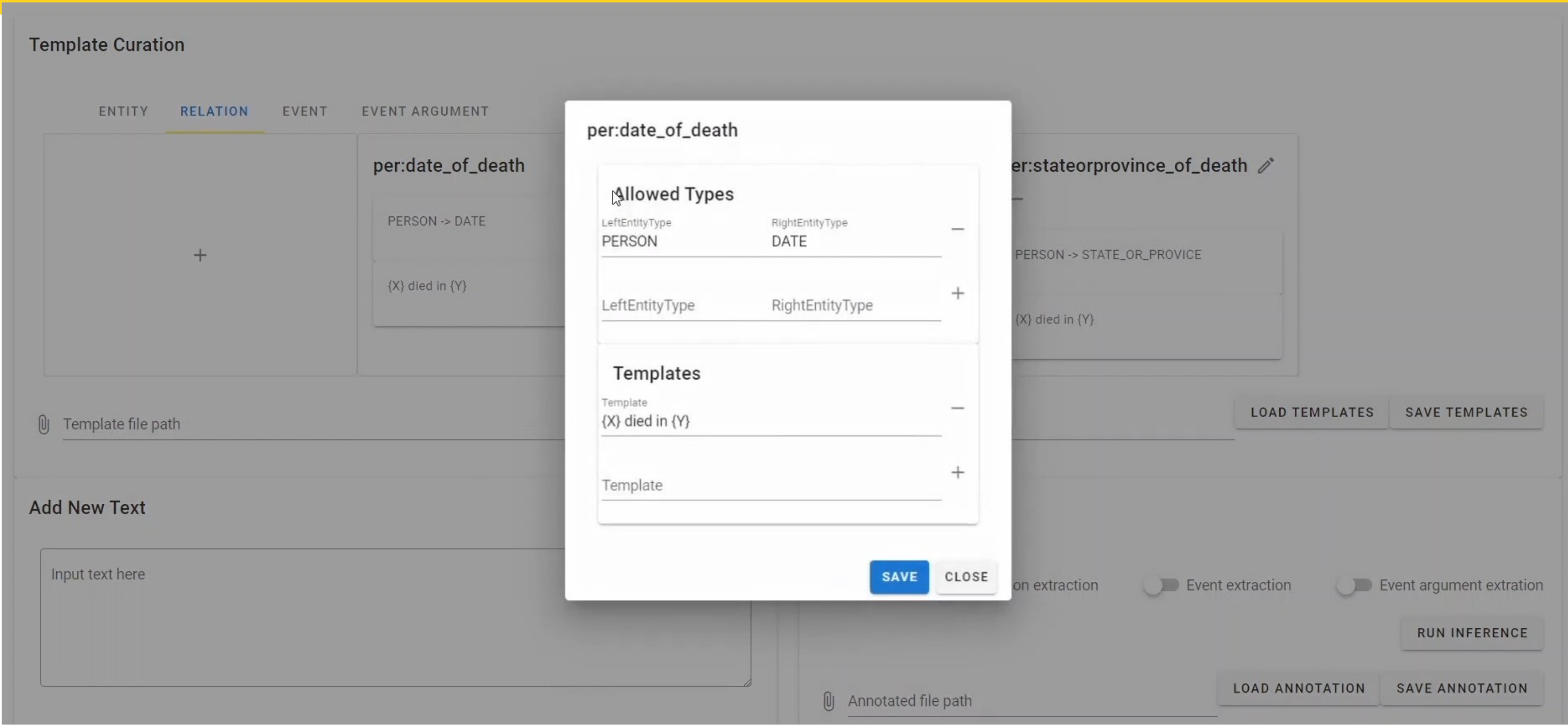
NER Score

Task	Total	Correct	Incorrect
NER	3	3 (1.000)	0 (0.000)

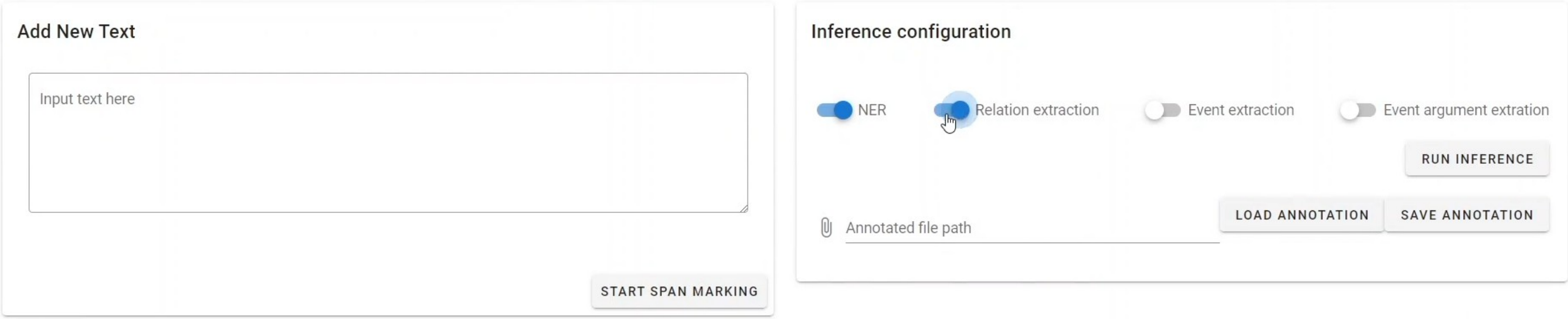
Rows per page: 10 1-1 of 1

Type	Total	Correct	Incorrect
PERSON	1	1 (1.000)	0 (0.000)
DATE	1	1 (1.000)	0 (0.000)
ORGANIZATION	1	1 (1.000)	0 (0.000)

Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)



Verbalize while defining, interactive workflow (Sainz et al. 2022, NAACL demo)

Relation extraction

John Smith, an executive at XYZ Corp., died in Florida on Sunday.

John Smith per:date_of_death Sunday

Type	Template	Score
per:date_of_death	{X} died in {Y}	0.988

John Smith per:employee_of XYZ Corp.

Type	Template	Score
per:employee_of	{X} is an employee of {Y}	0.976
per:employee_of	{X} is member of {Y}	0.933

Plan for this session

- Pre-trained LM
- Prompting
- Entailment
- Few-shot Information Extraction
- **Conclusions**

Conclusions

- Pre-train, prompt and entail works
  - Using “smaller” MLMs
- Few-shot Information Extraction is here
- Verbalize while defining, interactive workflow
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates

Conclusions

- Pre-train, prompt and entail works
  - Using “smaller” MLMs
- Few-shot Information Extraction is here
- Verbalize while defining, interactive workflow
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates
- Lower cost for building IE applications
- Friendlier to domain-experts

Conclusions

- Pre-train, prompt and entail works
  - Using “smaller” MLMs
- Few-shot Information Extraction is here
- Verbalize while defining, interactive workflow
  - Domain expert defines entities and relations in English
  - Runs the definitions on examples
  - Annotates a handful of incorrect examples, iterates
- Lower cost for building IE applications
- Friendlier to domain-experts
- Slides in my website, code at:

  https://github.com/osainz59/Ask2Transformers

Future work

- Verbalize while defining, interactive workflow
  - Check real use-cases
- Pre-train, prompt and entail works
  - Check tasks beyond IE
  - Compare head-to-head to plain LM (PET) and QA
  - Understand the role of contradictions
  - Identify useful inferences
  - Entailment as a method to teach inference to LM
- DL – reasoning research

Future work

- Verbalize while defining, interactive workflow
  - Check real use-cases
- Pre-train, prompt and entail works
  - Check tasks beyond IE
  - Compare head-to-head to plain LM (PET) and QA
  - Understand the role of contradictions
  - Identify useful inferences
  - Entailment as a method to teach inference to LM
- DL – reasoning research

We are hiring!

Few-shot Information Extraction
Pre-train, Prompt, Entail

THANKS!

Eneko Agirre
Director of HiTZ
Basque Center for Language Technology
(UPV/EHU)
@eagirre

http://hitz.eus/eneko/

https://github.com/osainz59/Ask2Transformers
Relation extraction (Sainz et al 2021, EMNLP)
Event-argument extraction (Sainz et al. 2022, NAACL findings)
Several IE tasks (Sainz et al. 2022, NAACL demo)