Information extraction & automated knowledge graph construction

Lecture 6: IE in a specialty domain

Lauriane Aufrant (Inria Paris)

Université Paris Cité – UFRL – M2 LI – 14/12/2023

Progress

- ~~09/11 – Overview of information extraction~~
- ~~16/11 – Entities, relations...~~
- ~~23/11 – Coreference, linking...~~
- ~~30/11 – From IE to automated knowledge graph construction~~
- ~~07/12 – IE annotations~~
- 14/12 – IE in a specialty domain

Today: IE in a specialty domain

- Domain-specific NLP
- Domain adaptation methods
- Handling specialized vocabulary
- Few-shot information extraction

A glance at my research topics

Interested in an internship? a PhD?

$\rightsquigarrow$ `first.last@inria.fr`

Domain-specific NLP

Research papers, pretrained models: Wikipedia

Research papers, pretrained models: Wikipedia

Real life: NOT Wikipedia!

Research papers, pretrained models: Wikipedia

Real life: NOT Wikipedia!

Specialty domain $\Rightarrow$ specialty language

$\rightsquigarrow$ New challenges, lower performance...

Examples of domains

- Technical: biomedical / chemistry / IT / academic papers...
- Topic: geopolitics / sport / law...
- Genre: fiction / news / handbook...
- Register: formal / spoken / social media...

Domain impact: specialized vocabulary (jargon)

Terms specific to the domain:

- 2-acetoxybenzoic acid
- CC(=O)OC1=CC=CC=C1C(=O)O
- BERT
- biphasic

Domain impact: specialized vocabulary (jargon)

Terms specific to the domain:

- 2-acetoxybenzoic acid
- CC(=O)OC1=CC=CC=C1C(=O)O
- BERT
- biphasic

Senses specific to the domain:

- root

Domain impact: specialized vocabulary (jargon)

Terms specific to the domain:

- 2-acetoxybenzoic acid
- CC(=O)OC1=CC=CC=C1C(=O)O
- BERT
- biphasic

Senses specific to the domain:

- root

root PERSON

Domain impact: form differences

Different cultures = different mention forms:

- Volodymyr Oleksandrovytch Zelensky
- Mohamed Abdel Raouf Arafat al-Qoudwa al-Husseini
- George W. Bush
- General Hammond
- Saint-Rémy-lès-Chevreuse
- Krung Thep Maha Nakhon
- SARL / Inc.

Impact on mention's context: colloquial syntax, smileys, hashtags...

Domain impact: specific interests

- Entities: Vehicle, Protein...

Domain impact: specific interests

- Entities: Vehicle, Protein...
- Relations: Social[Grandparent], Frequency-range, T-melt

Domain impact: specific interests

- Entities: Vehicle, Protein...
- Relations: Social[Grandparent], Frequency-range, T-melt
- Events (N-ary relations): Gene-expression

Domain impact: specific interests

- Entities: Vehicle, Protein...
- Relations: Social[Grandparent], Frequency-range, T-melt
- Events (N-ary relations): Gene-expression

$\hookrightarrow$ Rare / unseen phenomena, new labels...

Domain/language interplay

- Arabizi (social media Arabic)

jetaime	madjid	nchalah	tkon	dima	fal3ali	wdima	mcharaf	bladna
VERB	PROPN	INTJ	VERB	ADV	NOUN	ADV	ADJ	NOUN
I love you	*Madjid*	*ichaAllah*	*may you be*	*always*	*at the top*	*and always*	*honoring of*	*our country*

- Code-switching

  Tu as pris quel train set pour tes embeddings ?

Domain adaptation methods

Domain adaptation

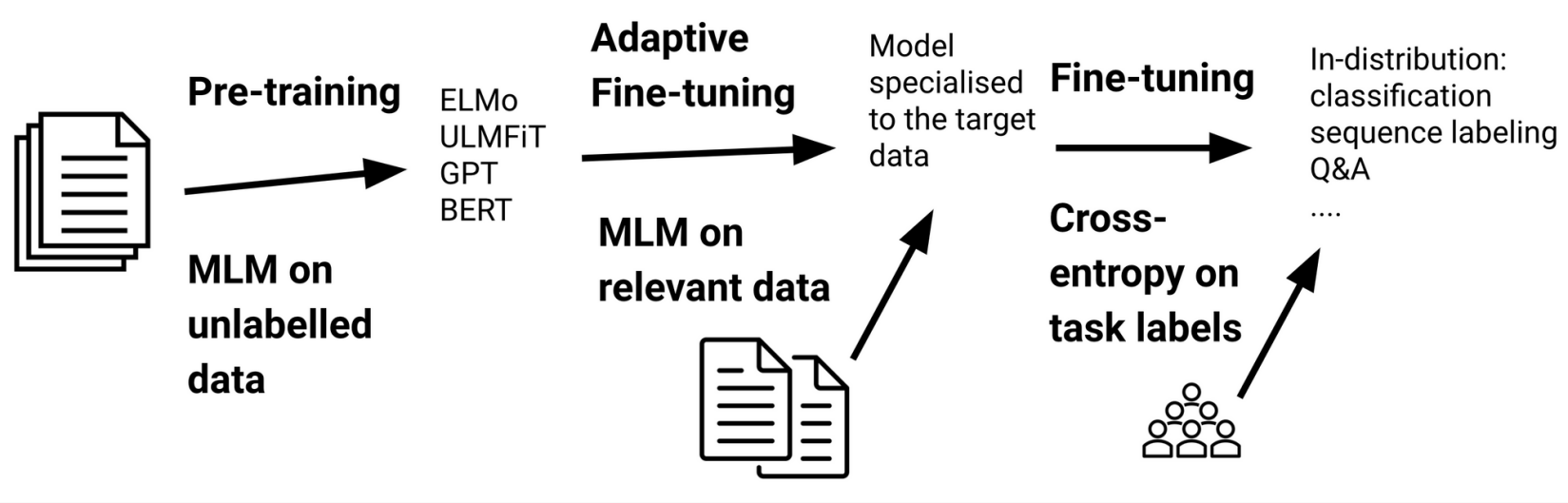
- Training in a source domain (large out-of-domain data)
- Testing in a target domain (little or no in-domain data)

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

$\hookrightarrow$ a broad range of mathematically grounded methods

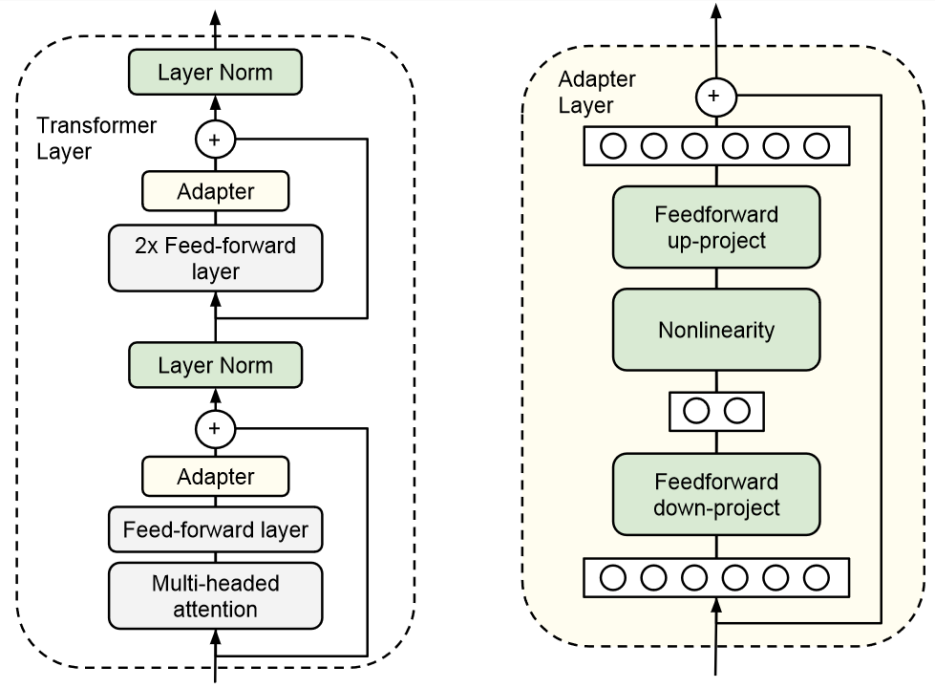
Empirical adaptation: adaptive fine-tuning



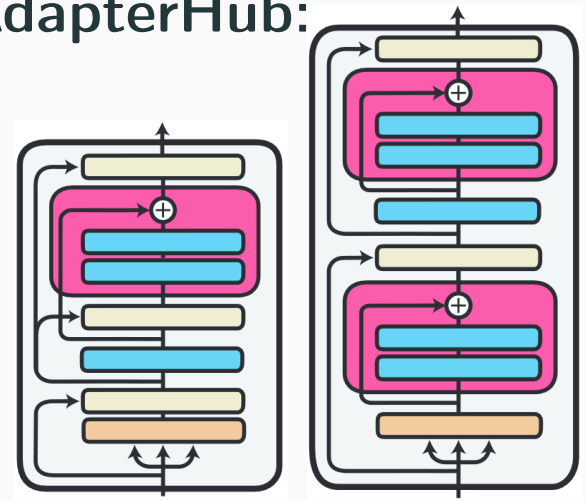
Ruder, 2021.

Parameter-efficient fine-tuning

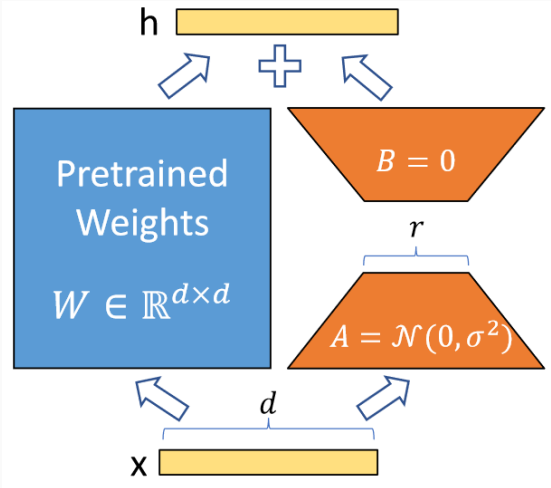
Adapters:



AdapterHub:

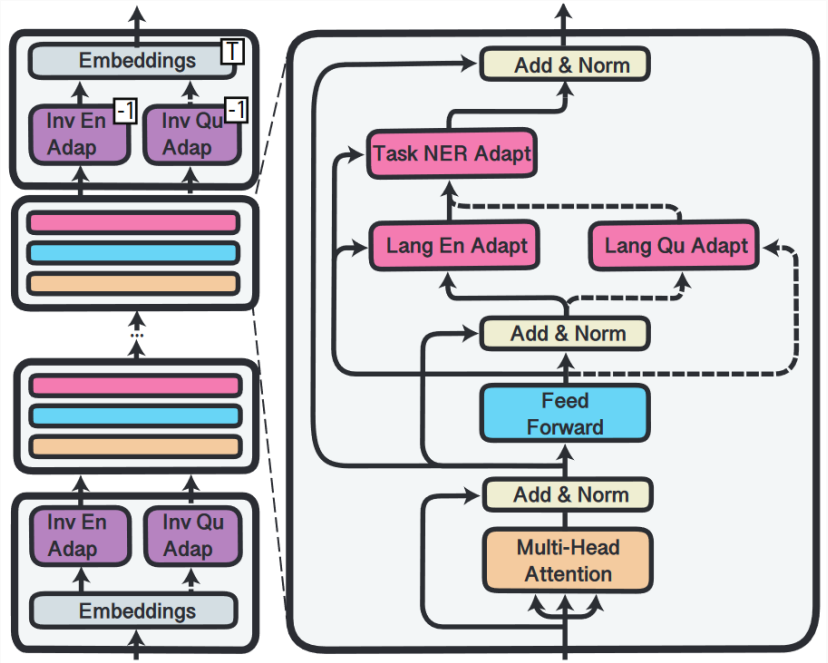


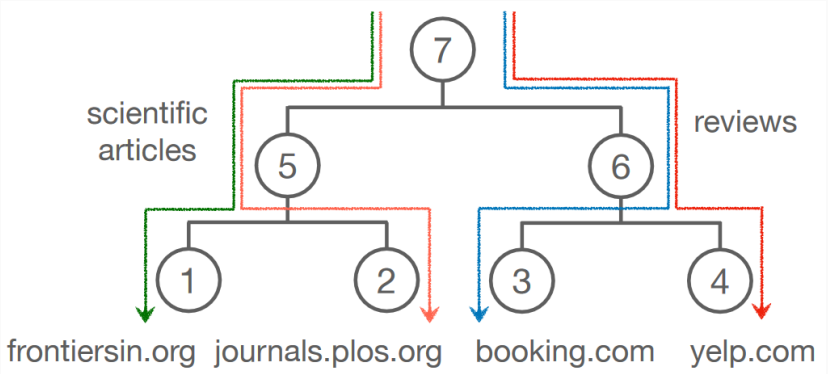
LoRA:

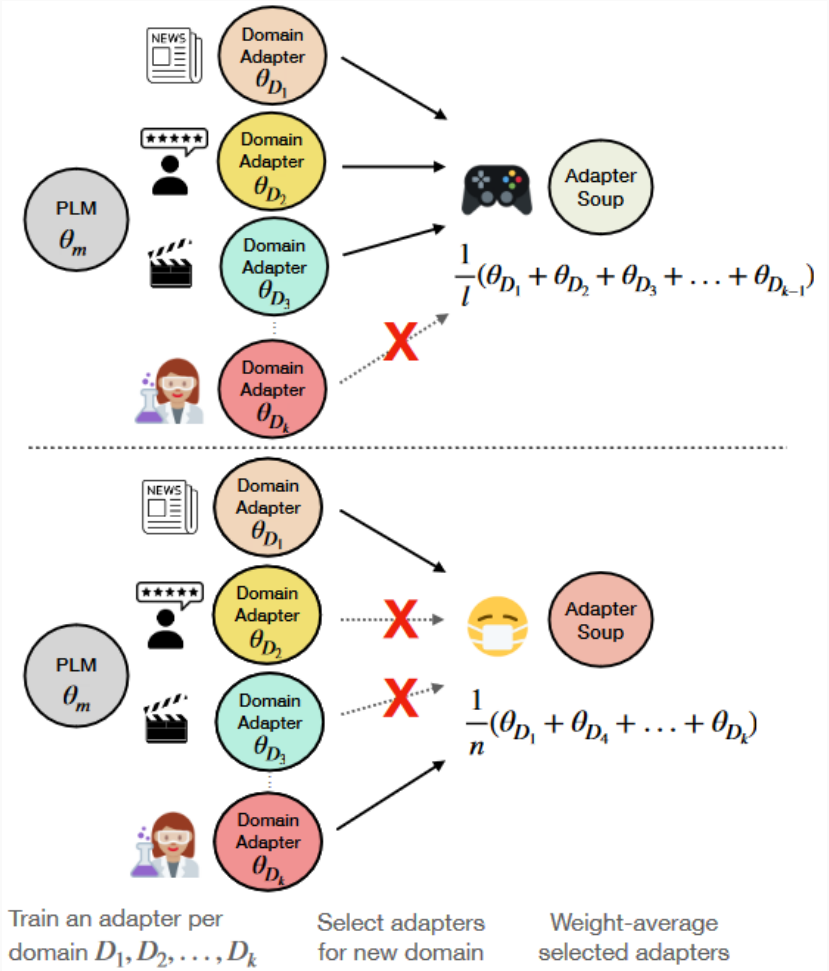


Houlsby et al., 2019. Pfeiffer et al., 2020. Hu et al., 2021.

Composing adapters







Pfeiffer et al., 2020. Chronopoulou et al., 2022. Chronopoulou et al., 2023.

Problem solved?

Handling specialized vocabulary

Subword tokenization challenges

Input	Tokenized
deconstructed	[CLS], deco, ##nst, ##ru, ##cted, [SEP]
deactivated	[CLS], dea, ##ct, ##ivated, [SEP]
unequal	[CLS], une, ##qual, [SEP]
ccabbage	[CLS], cc, ##ab, ##bag, ##e, [SEP]
cababge	[CLS], cab, ##ab, ##ge, [SEP]
cabbagee	[CLS], cabbage, ##e, [SEP]']
unsaturated	[CLS], un, ##sat, ##ura, ##ted, [SEP]
saturated	[CLS], saturated, [SEP]
pork has saturated fat	[CLS], pork, has, saturated, fat, [SEP]
pork has ##sat ##ura ##ted fat	[CLS], pork, has, ##sat, ##ura, ##ted, fat, [SEP]

Reference	Medical Vocabulary	General Vocabulary
paracetamol	[paracetamol	[para, ce, tam, ol]
choledocholithiasis	[choledoch, olithiasis]	[cho, led, och, oli, thi, asi, s]
borborygmi	[bor, bor, yg, mi]	[bo, rb, ory, gm, i]

Nayak et al., 2020. El Boukkouri et al., 2020.

Relaxing subword tokenization

Dynamic Programming Encoding for Subword Segmentation in Neural Machine Translation

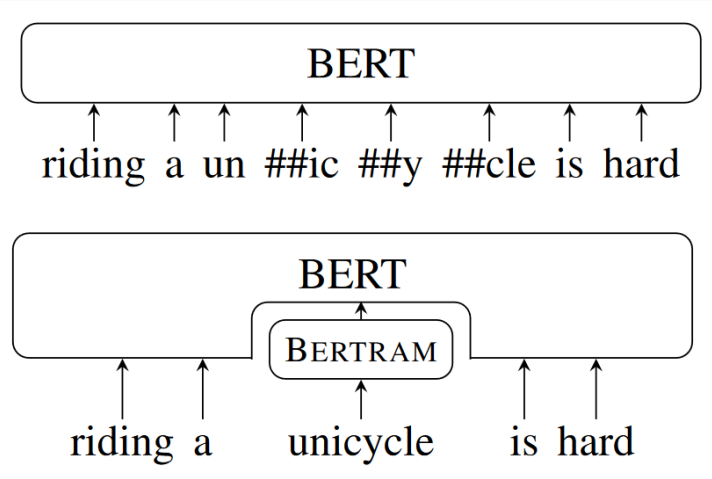
Xuanli He Monash University {xuanli.he1,gholamreza.haffari}@monash.edu

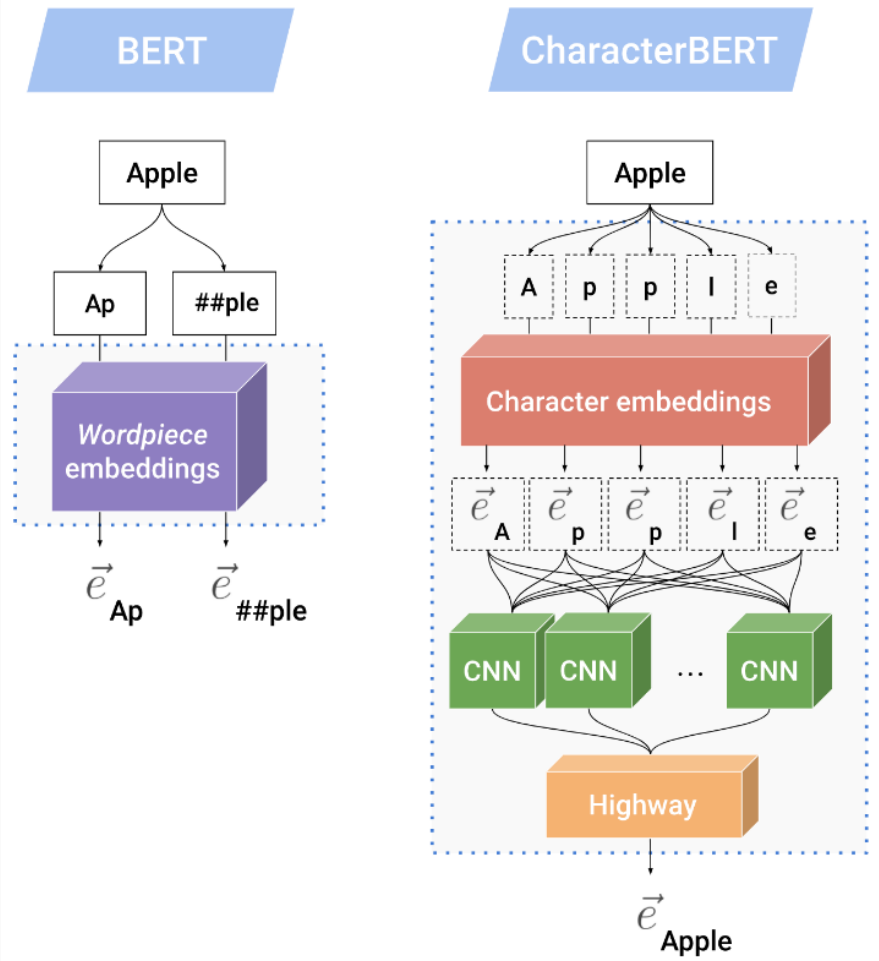
Gholamreza Haffari Monash University

Mohammad Norouzi Google Research mnorouzi@google.com

BPE-Dropout: Simple and Effective Subword Regularization

Ivan Provilkov[*1,2] **Dmitrii Emelianenko**[*1,3] **Elena Voita**[4,5]





El Boukkouri et al., 2020. Schick & Schütze, 2020.

Vocabulary transfer

- Retrain with new vocabulary vs extend vocabulary with specific terms

  $\hookrightarrow$ more challenging for contextual embeddings

Vocabulary transfer

- Retrain with new vocabulary vs extend vocabulary with specific terms

  $\hookrightarrow$ more challenging for contextual embeddings

- VIPI: *Vocabulary Initialization with Partial Inheritance* (Mosin et al., 2023)

Vocabulary transfer

- Retrain with new vocabulary vs extend vocabulary with specific terms

  $\hookrightarrow$ more challenging for contextual embeddings

- VIPI: *Vocabulary Initialization with Partial Inheritance* (Mosin et al., 2023)

The Recent Advances in Automatic Term Extraction: A survey

HANH THI HONG TRAN, Jozef Stefan Institute, Slovenia; University of La Rochelle, France, Slovenia
MATEJ MARTINC, Jozef Stefan Institute, Slovenia, Slovenia
JAYA CAPORUSSO, Jozef Stefan Institute, Slovenia, Slovenia
ANTOINE DOUCET, University of La Rochelle, France, France
SENJA POLLAK, Jozef Stefan Institute, Slovenia, Slovenia

Few-shot information extraction

Few-shot IE

Predicting a given (entity|relation|event) type based on ∼5 occurrences

Few-shot prompting

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

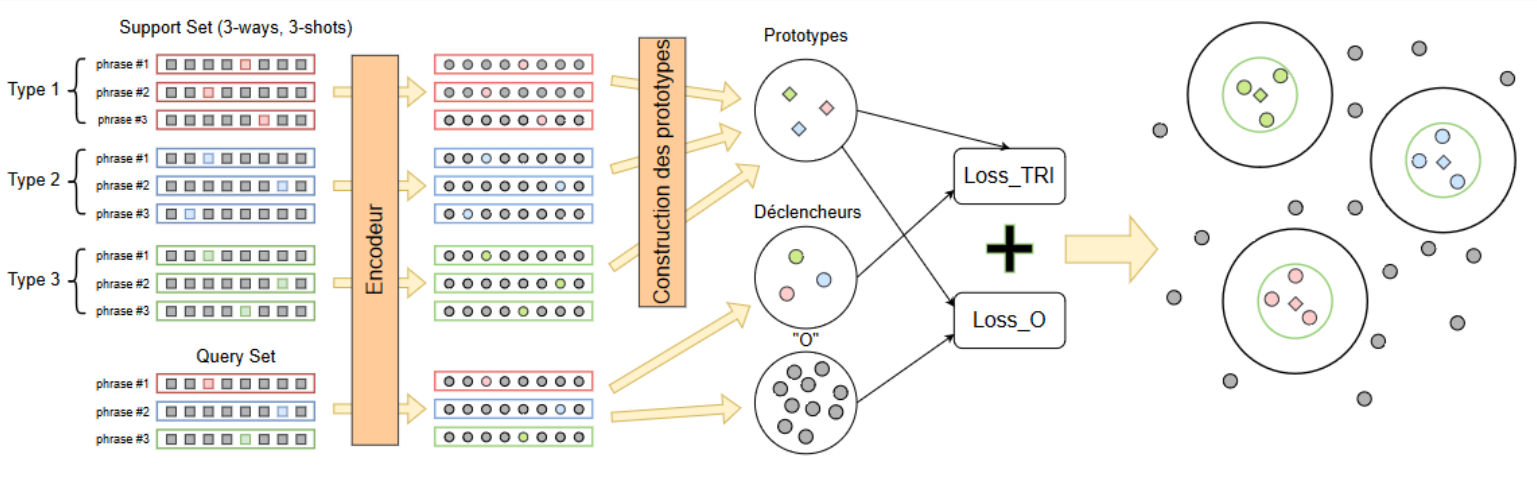
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

GPT-3 (2020)

Few-shot adaptation with prototypes

Few-shot event detection:



FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation

Xu Han[1,*] Hao Zhu[1,*] Pengfei Yu[2,*] Ziyun Wang[1,†,*]
Yuan Yao[1] Zhiyuan Liu[1,‡] Maosong Sun[1]

Tuo et al., 2023

Last questions?

Good luck with your project!

first.last@inria.fr