# Prompt generalization across Language Models

Nathanaël Carraz Rakotonirina

upf. Universitat Pompeu Fabra Barcelona

Computational Linguistics and Linguistic Theory

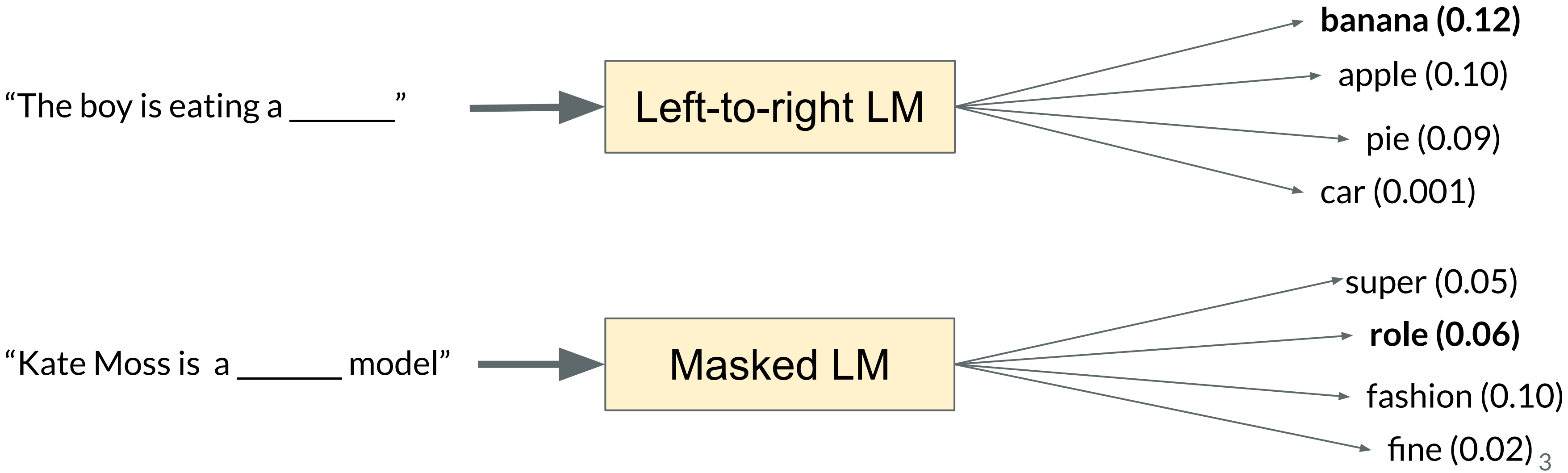
ALiEN

# Research interests

- Prompt engineering
- Emergent communication in pre-trained language models
- Representation learning

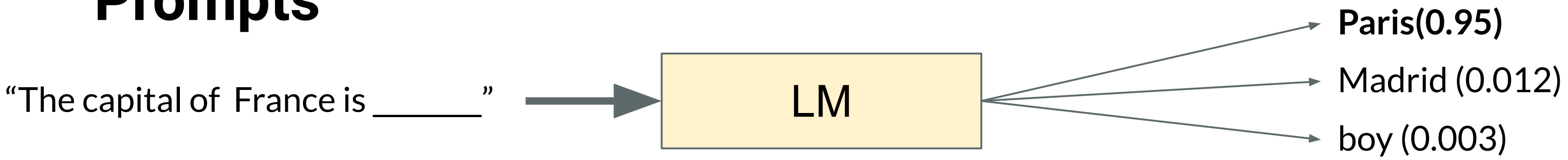
# Language models

- A **language model (LM)** assigns a probability to a sentence.
- A LM can be used to predict words in a sentence:



# Prompting for knowledge extraction

- We can extract information from existing LMs using only **Prompts**




- The dataset is composed of triples <*subj,rel,obj*> like <*France,capital,Paris*>
- Prompts can be categorized into **manual/automated** and **discrete/continuous**

(Petroni, Fabio, et al. "Language models as knowledge bases?.")

# Automated Prompting

**Automated** prompting methods learn to generate prompts using a **LM** and a **dataset**

LM

Prompting method

Dataset

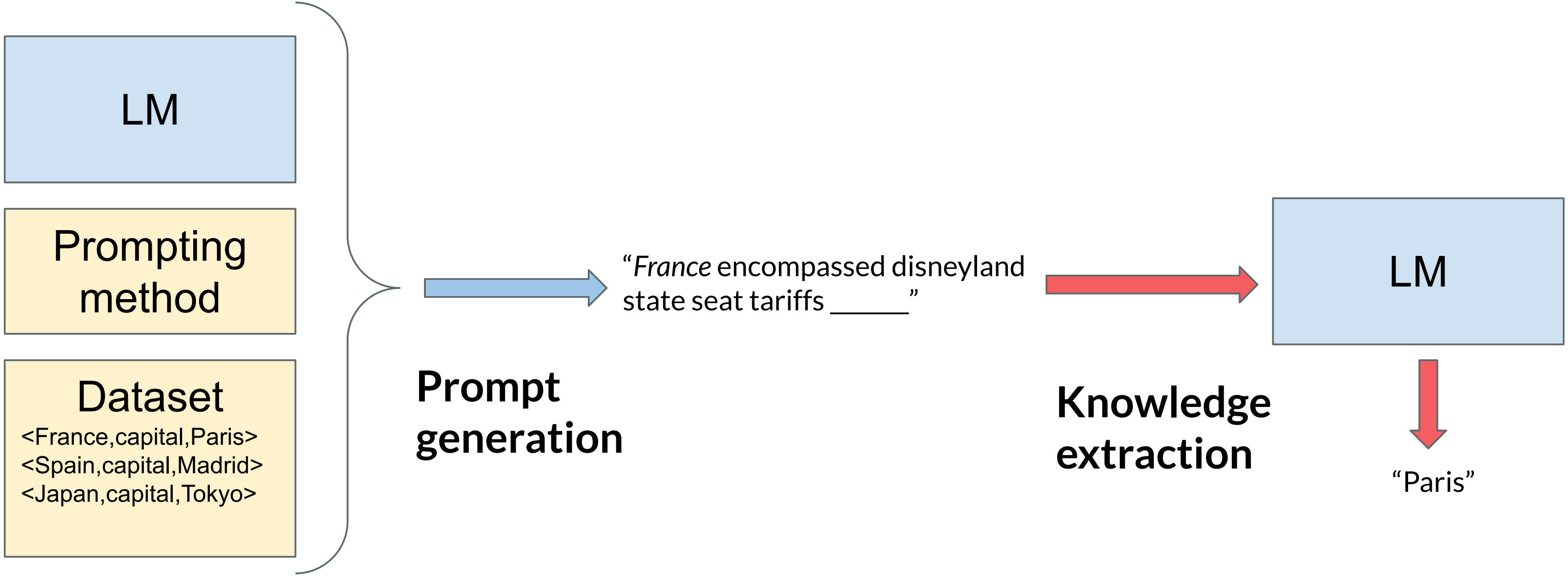
<France,capital,Paris>
<Spain,capital,Madrid>
<Japan,capital,Tokyo>

**Prompt generation**

"*France* encompassed disneyland state seat tariffs ______"

# Automated Prompting

**Automated** prompting methods learn to generate prompts using a **LM** and a **dataset**



# Automated Prompting

Examples of prompts induced using Autoprompt

| Relation | Manual prompts | Autoprompt (BERT) |
|---|---|---|
| Place of birth | [X] was born in [Y]. | [X] who flightstial cyclist \u00a1 [Y]. |
| Instrument | [X] plays [Y]. | [X] playingdrum concertoative electric [Y]. |
| Capital | The capital of [X] is [Y]. | [X] includesiidae geologic countryside near [Y]. |

(Shin, Taylor, et al. "Autoprompt: Eliciting knowledge from language models with automatically generated prompts.")

# Setup

## Prompting methods

- Manual
- Semi-manual (LPAQA)
- Discrete (AutoPrompt)
- Continuous (OptiPrompt)

## Dataset

LAMA TReX (Slot-filling task)

## Metric

Accuracy (P@1)

## Language models

- Left-to-right (GPT2)
- Masked (BERT, RoBERTa)
- Sequence-to-sequence (T5, BART)

# How do prompting methods perform?

- We induce and evaluate prompts with the **same** LM
- Trends across LMs

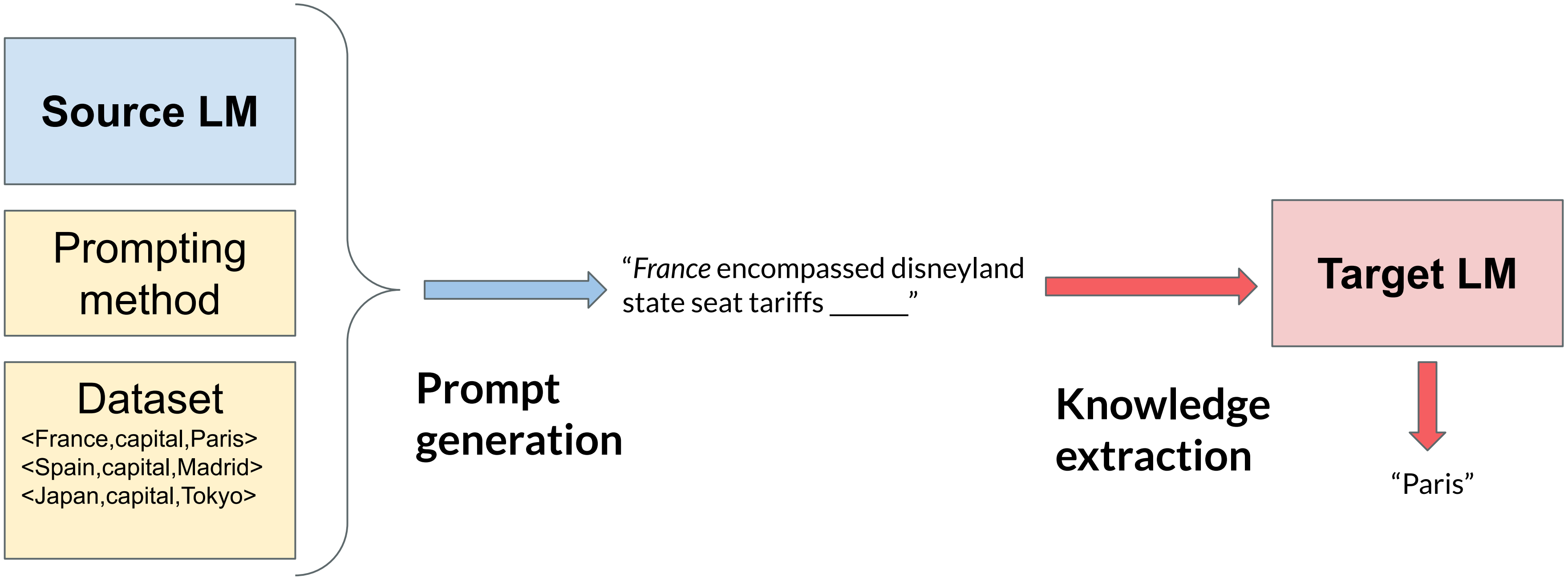
**continuous > discrete > semi-manual > manual**

- Trends across prompting methods

**masked > seq-2-seq > left-to-right**

# Prompt generalization

We induce prompts with a **source** LM and evaluate them with a **target** LM



# Do prompts generalize across LM?

- We induce prompts with a **source** LM and evaluate with a **target** LM
- We only focus on a **discrete prompting method**

- Induced prompts do not generalize to LM they were not trained on
- Gap gets bigger as we transfer across different LM types

# How to induce prompts that generalize better?

- We use **two** LMs instead of one during training
- The **generator** LM proposes candidates that the **evaluator** LM evaluate

- Induced prompts generalize better
- There are limitations to mixing

# What are the properties of general prompts?

Compared to regular prompts, **mixed prompts** are/have:

- Higher semantic overlap with English
- More word-like
- More robust to token shuffling
- More robust to token deletion

# What's next?

- Input-specific prompts (instead of relation-specific)
- Machine-to-machine communication