College of Computing
UM6P

Building efficient Natural Language Processing systems for low-resource languages

Abdellah EL MEKKI
College of Computing, Mohammed VI Polytechnic University

رياضيات المغرب
MATH MAROC

Who am I?

Abdellah EL MEKKI

[2018 - 2023] PhD. Mohammed VI Polytechnic University, Morocco

- Presented my work at several international conferences (NAACL, COLING, CAAI).
- Ranked Top-3 in various NLP challenges.

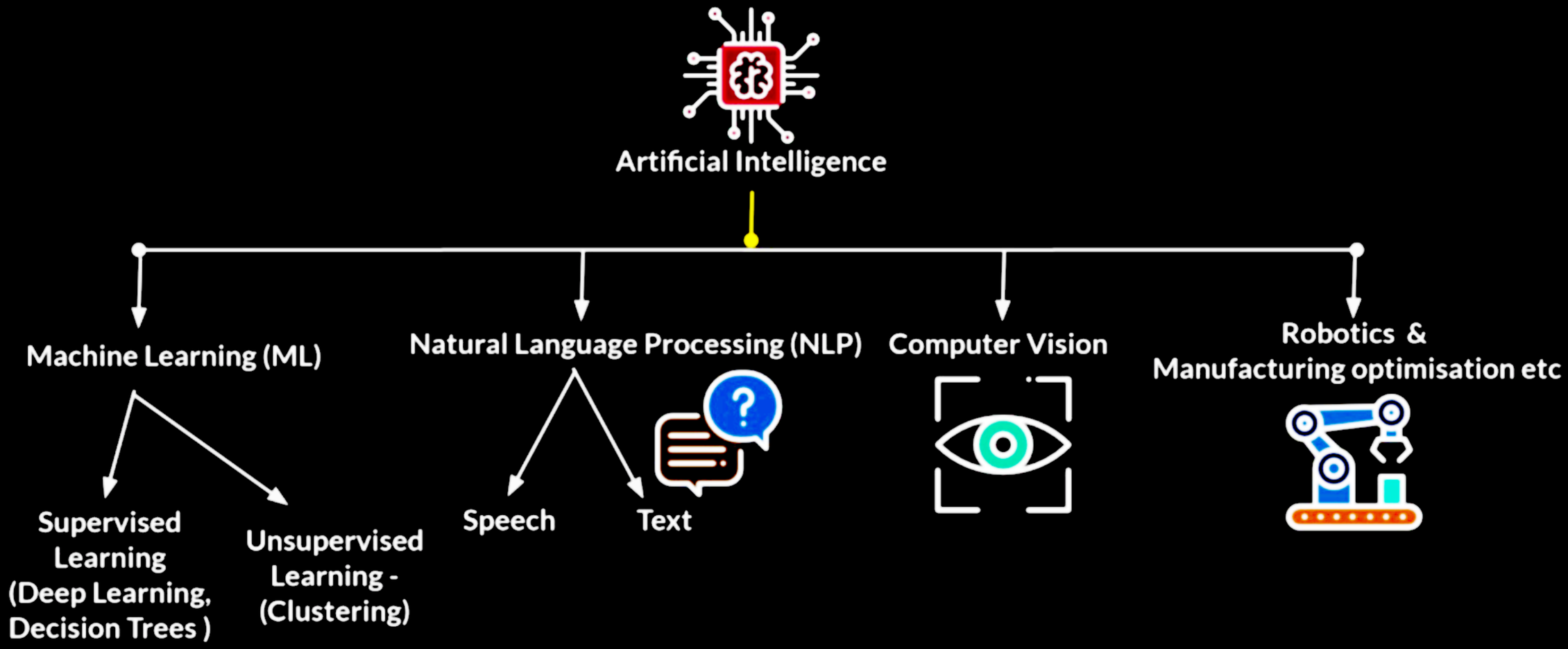
[2022 - 2023] Visiting Researcher University of British Columbia, Canada

Who am I?

Topics

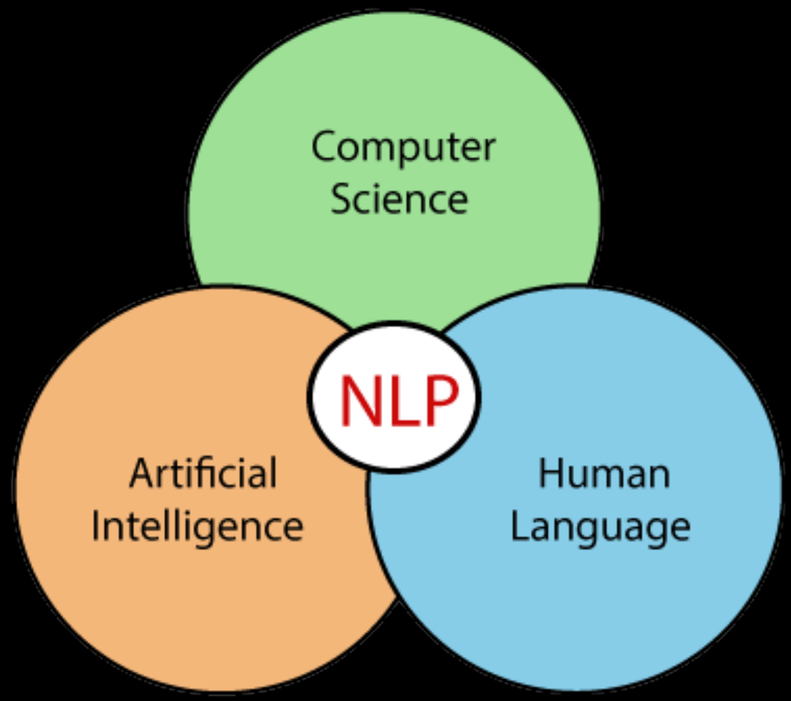
- NLP for low-resource languages.
- Multilingual NLP.
- Transfer Learning in NLP.
- Speech Recognition.
- And now … Large Language Models.

NLP and AI


Artificial Intelligence
Machine Learning (ML)
Supervised Learning (Deep Learning, Decision Trees)
Unsupervised Learning - (Clustering)
Natural Language Processing (NLP)
Speech
Text
Computer Vision
Robotics & Manufacturing optimisation etc

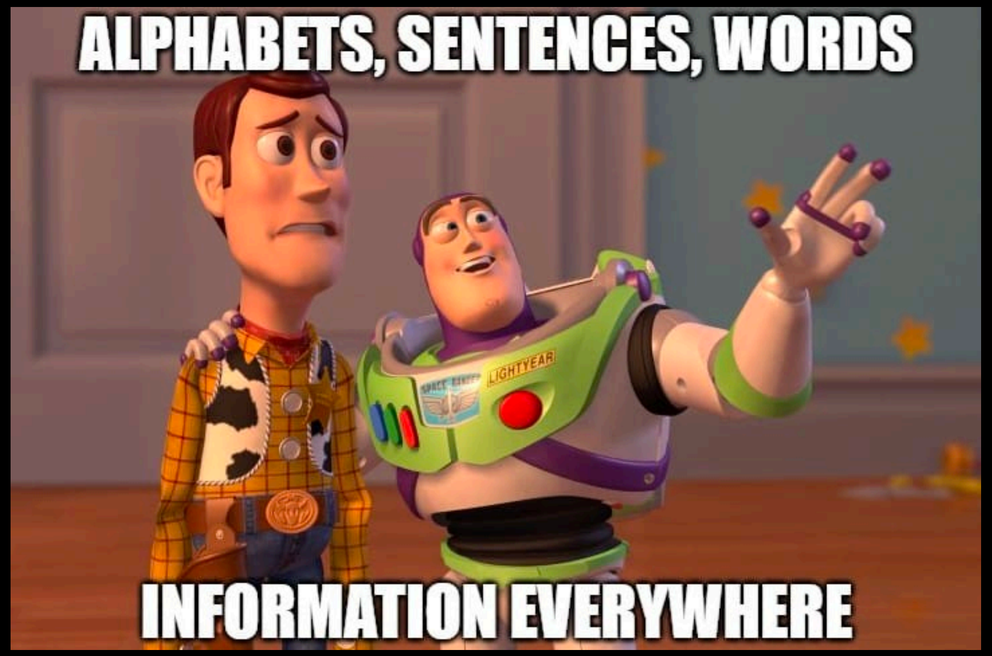
What is NLP?

★**Wiki: Natural language processing (NLP)** is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human **(natural)** languages.

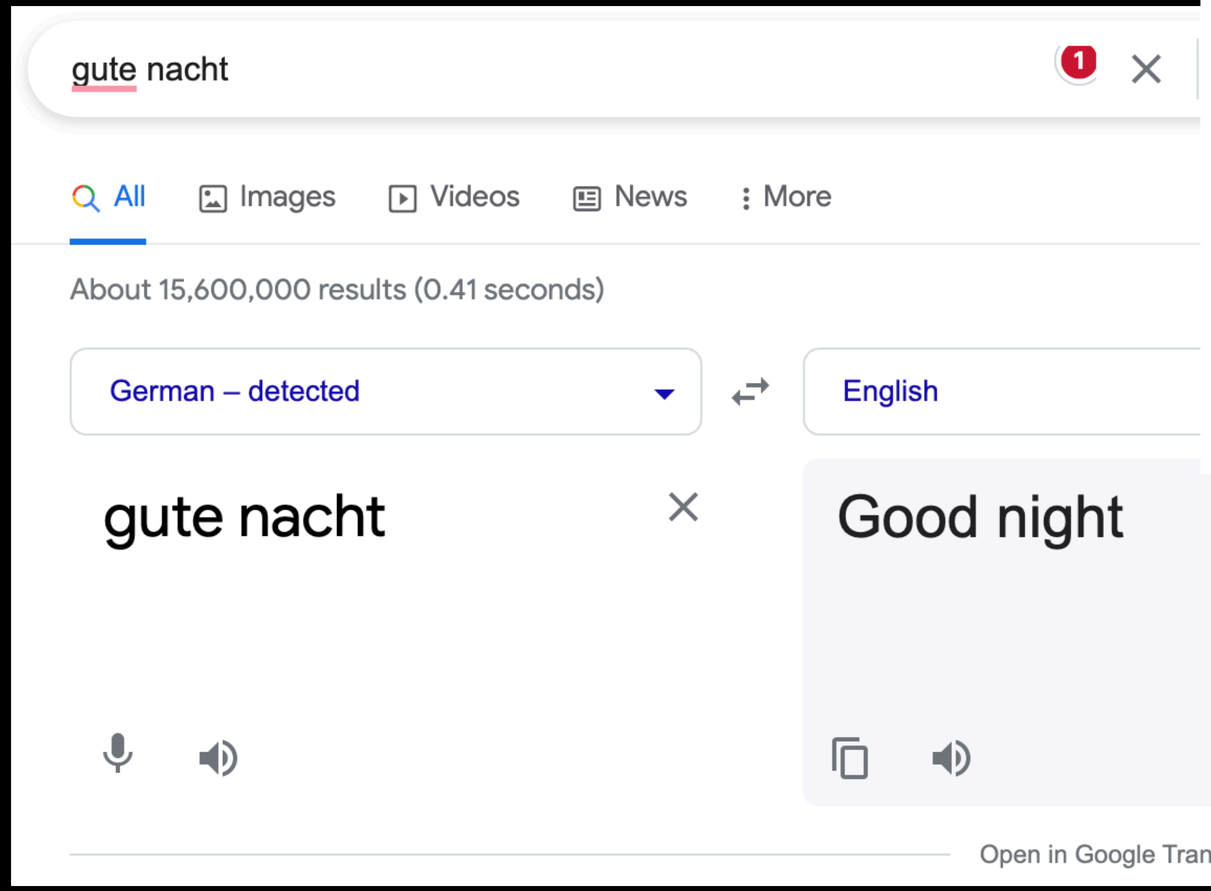


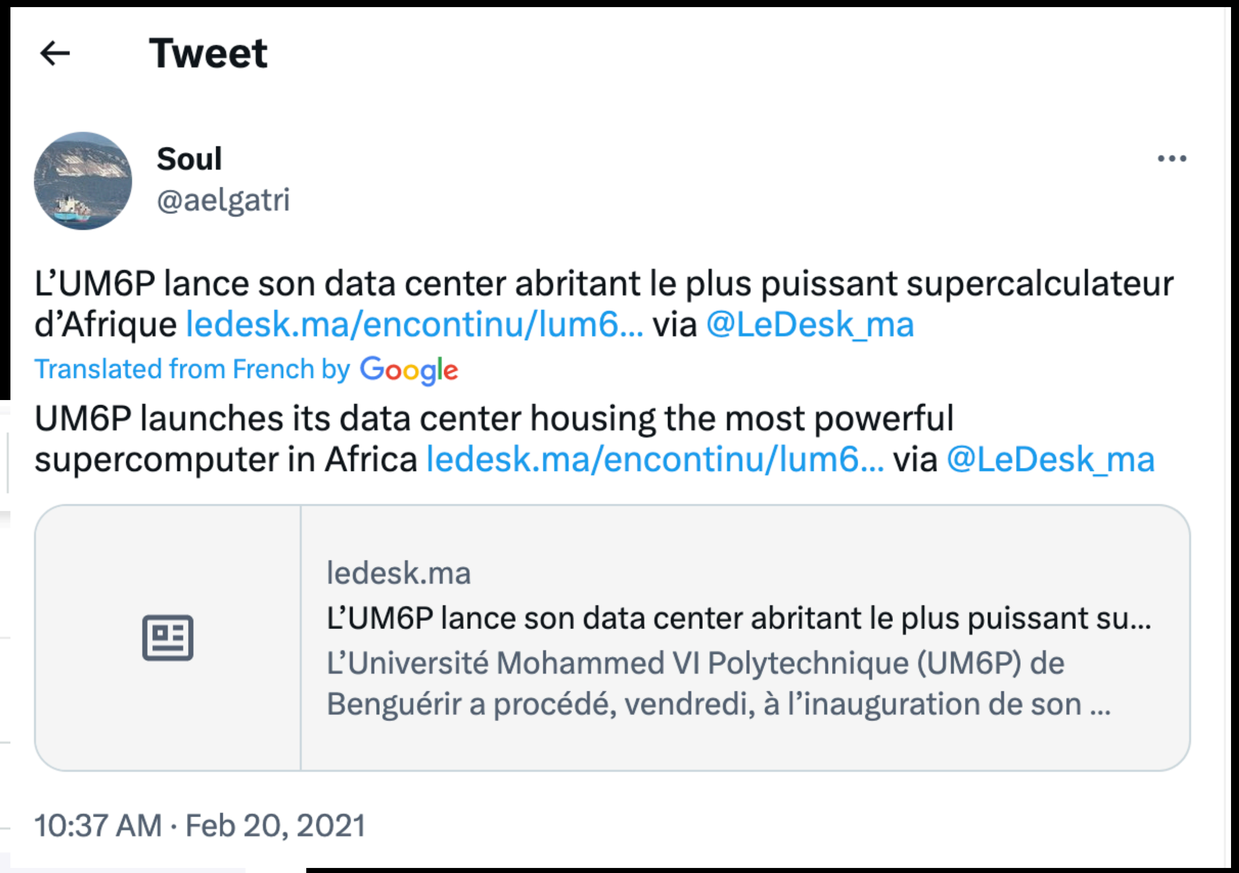
What is NLP?

- Identify the structure and meaning of words, sentences, texts and conversations.
- Deep understanding of broad language.
- NLP is all around us.



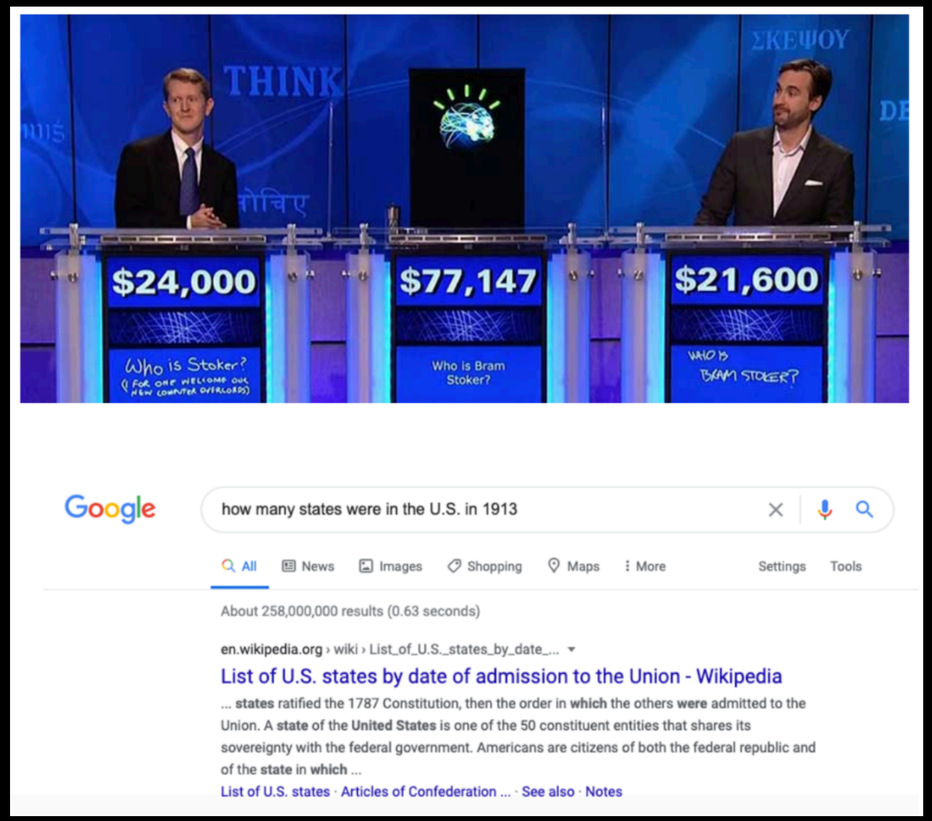
Machine Translation





Question answering

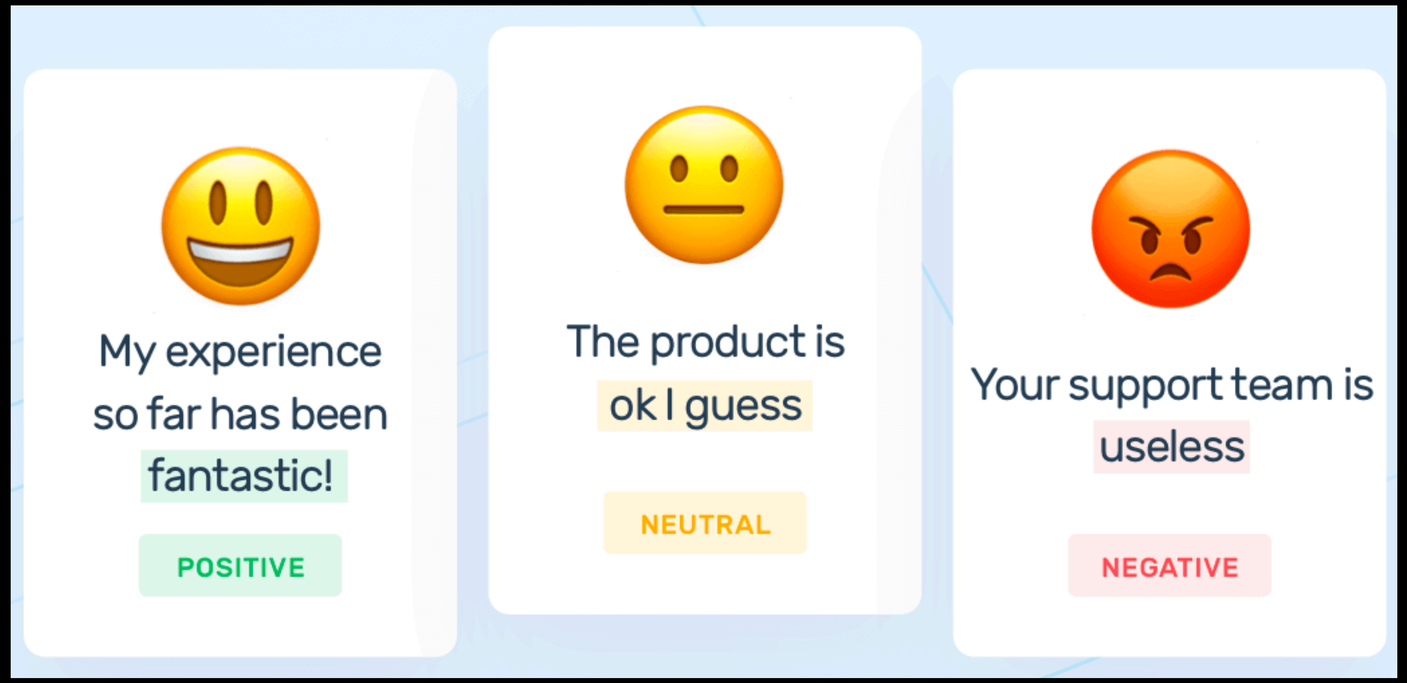
- What does “divergent” mean?
- What year was Abraham Lincoln born?
- How many states were in the United States that year?
- How much Chinese silk was exported to England at the end of the 18th century?
- What do scientists think about the ethics of human cloning?



Smart voice assistants

- **Siri contains:**
  - Speech recognition
  - Language analysis
  - Dialog processing
  - Text to speech

Sentiment/opinion analysis



How to build an NLP system?

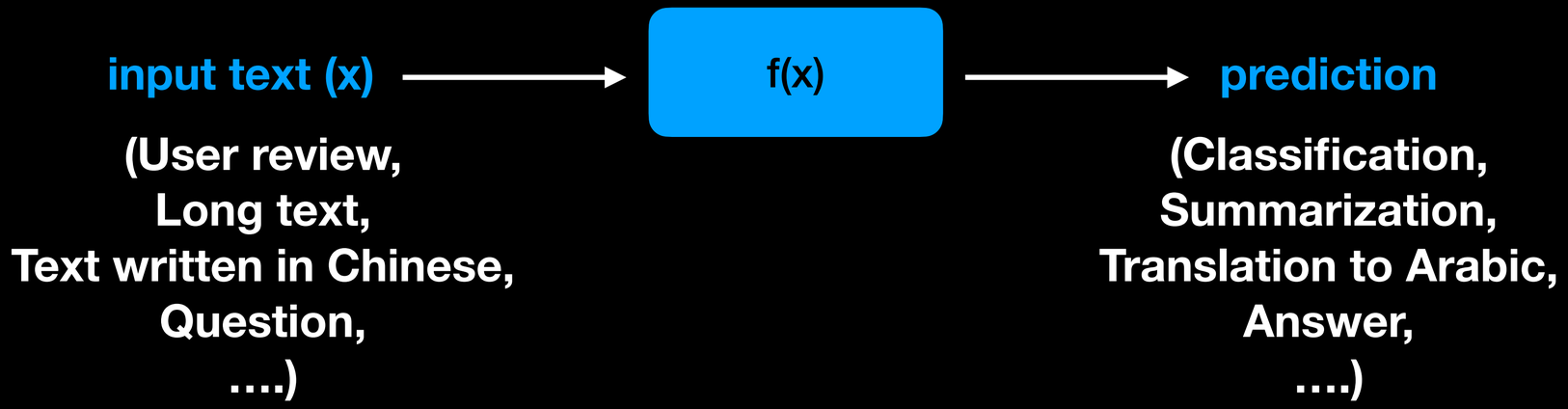
input text (x)

(User review,
Long text,
Text written in Chinese,
Question,
….)

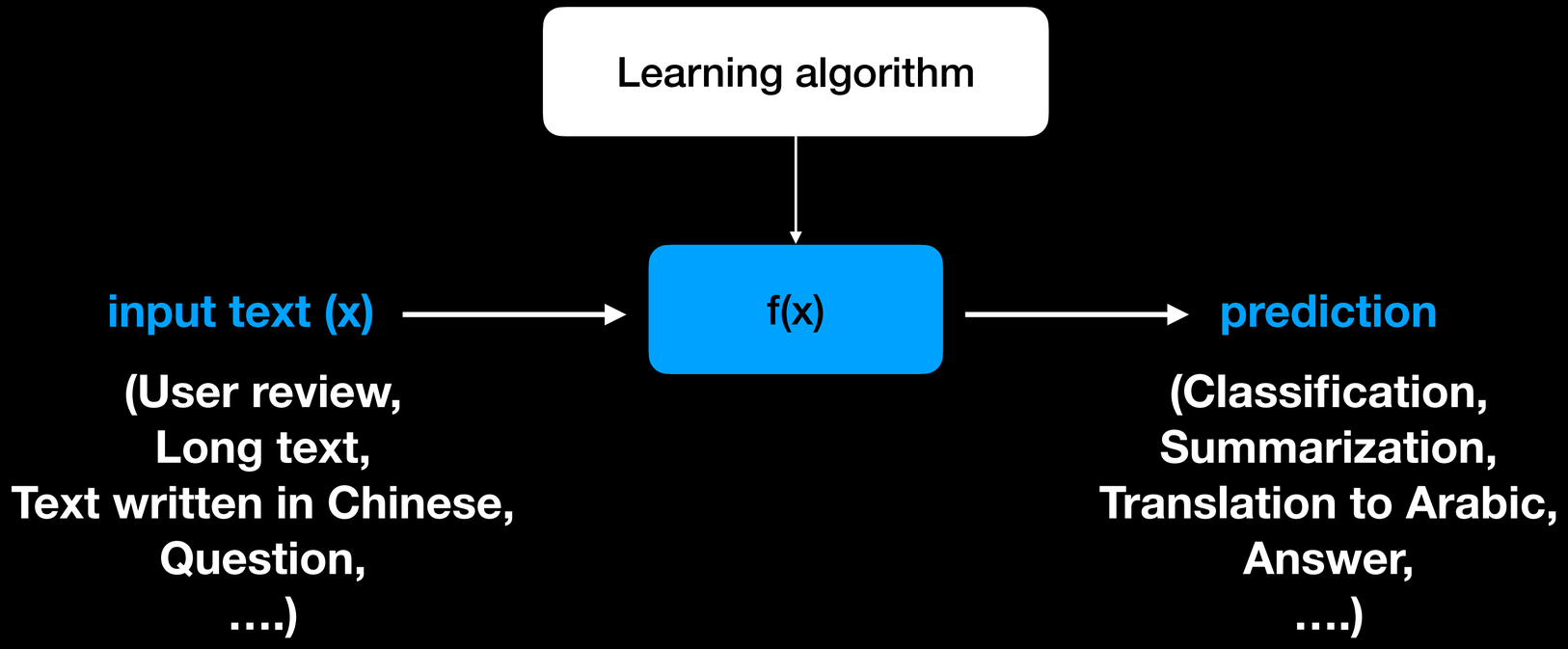
prediction

(Classification,
Summarization,
Translation to Arabic,
Answer,
….)

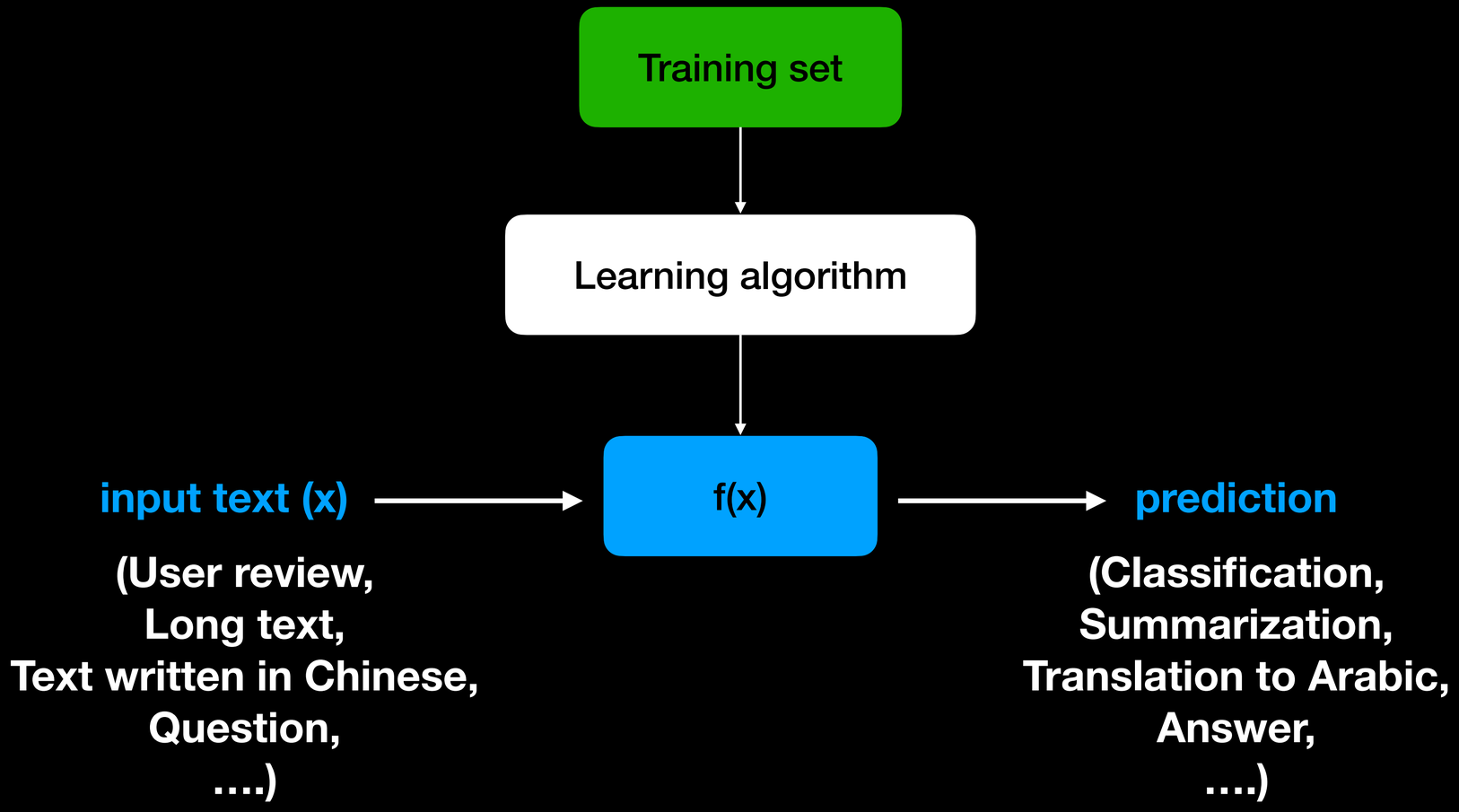
How to build an NLP system?



How to build an NLP system?



How to build an NLP system?



History of NLP

- **Hand-crafted Systems:** Knowledge Engineering **[1950s-]**
  - Rules written by hand; adjusted by error analysis.
  - Require experts who understand both the systems and domain.
  - Iterative guess-test-tweak-repeat cycle.
- **Automatic Trainable (Machine Learning)** Systems with engineered features **[1985s-2012]**
  - The tasks are modeled in a statistical way.
  - More robust techniques based on rich annotations.
  - Perform better than rules (Parsing 90% vs. 75% accuracy).
- **Automatic Trainable Neural architectures** with no/limited engineered features **[2012-]**

Components of NLP

- Natural Language Understanding
  - ✦Mapping the given input in the natural language into a useful representation.
  - ✦Different level of analysis required: morphological, syntactic, semantic.
- Natural Language Generation
  - ✦Producing output in the natural language from some internal representation.
  - ✦Different level of synthesis required:
    - ◉ Deep planning (what to say)
    - ◉ Syntactic generation

What is NLP?

- NL ∈ {English, Arabic, French, Hindi, …, Mandarin}
- Automation of NLs:
  - ✦Analysis (NL → Representation)
  - ✦Generation (Representation → NL)
  - ✦Acquisition of Representation from knowledge and data.

Why NLP is HARD?

- Richness: there are many ways to express the same meaning, and immeasurably many meaning to express.
- Linguistic diversity across languages, dialects, genres styles.

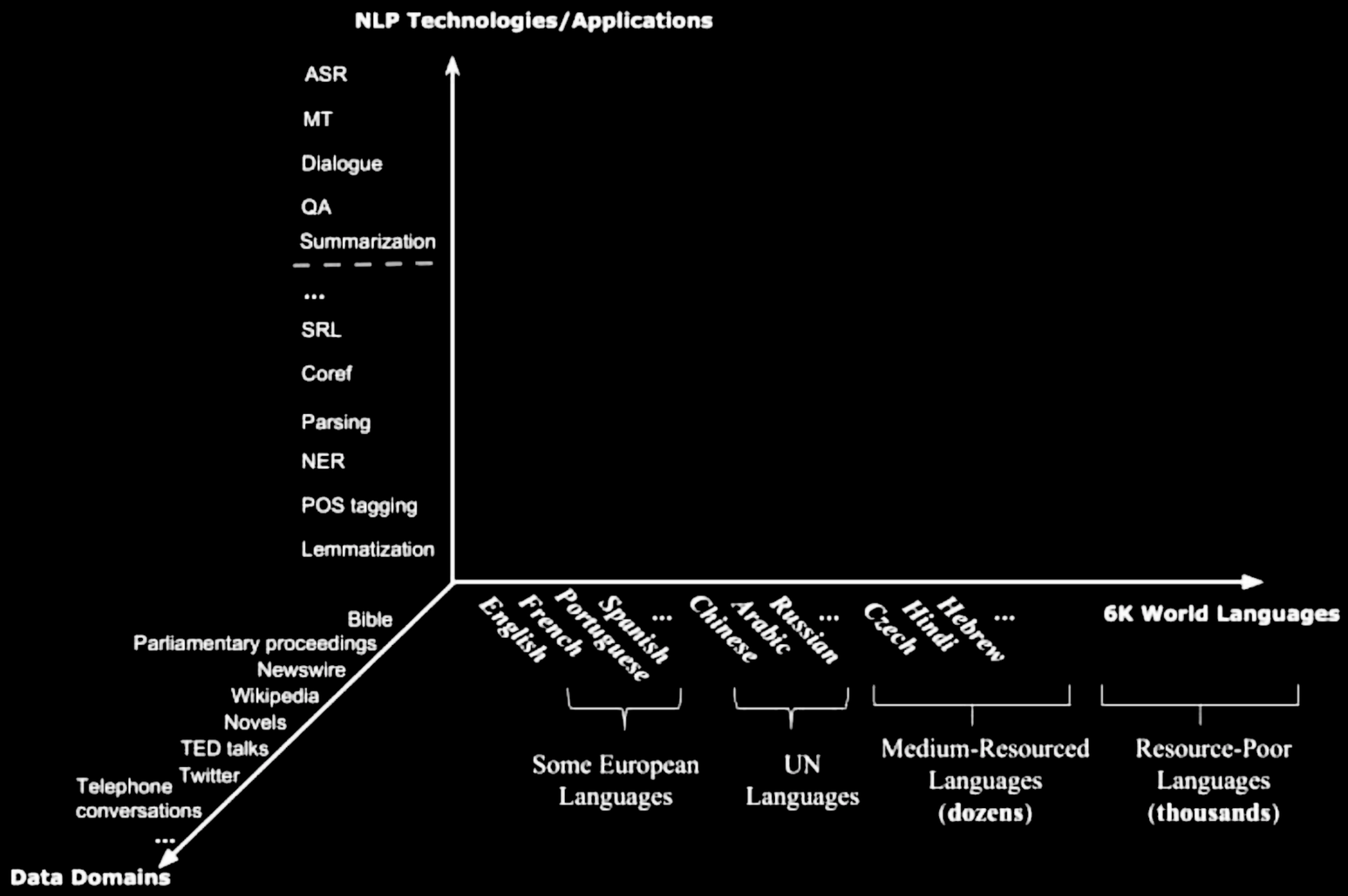
1. **Ambiguity**
2. **Sparsity**
3. **Variation**
4. **Expressivity**

Why NLP is HARD?

Ambiguity

- Ambiguity at multiple levels:
  - Word senses: **bank** (finance or river?)
  - Part of speech: **chair** (noun or verb?)
  - Syntactic structure: **I can see a man with a telescope**
  - Multiple: **I saw her duck**

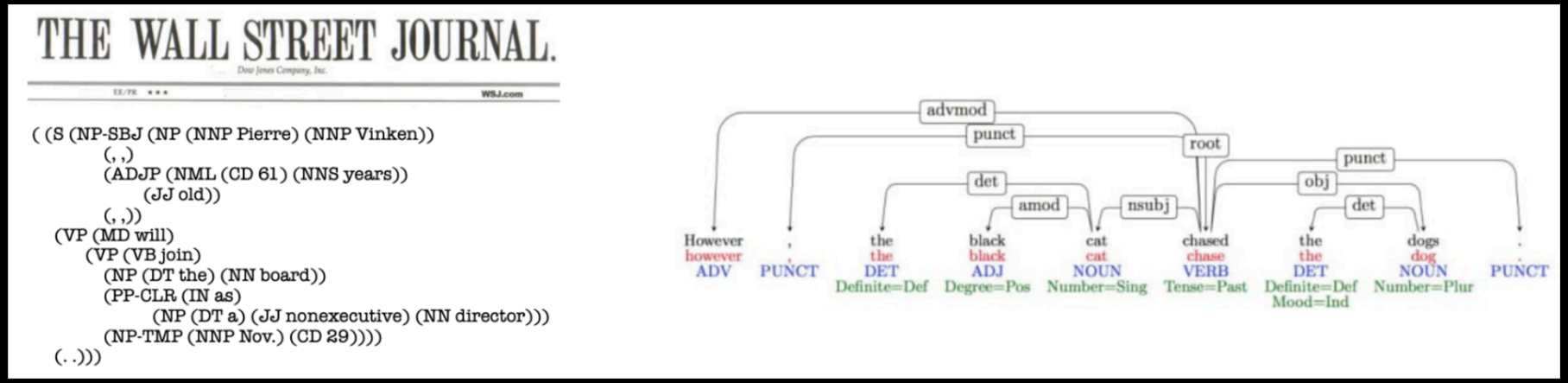
Why NLP is HARD?



Why NLP HARD

Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal…



- What will happen if we try to use this tagger/parser on social media?

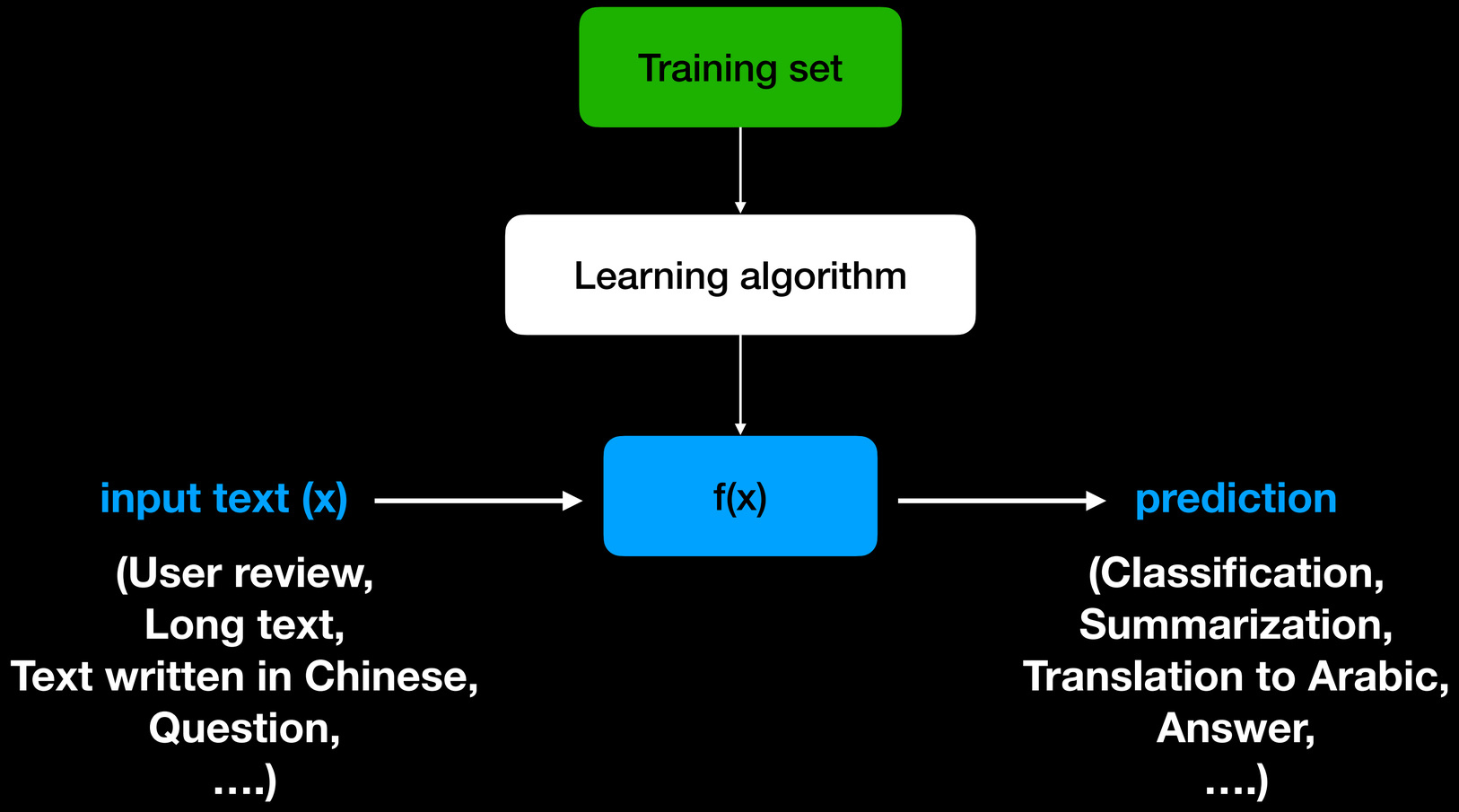
Why NLP is HARD?

Expressivity

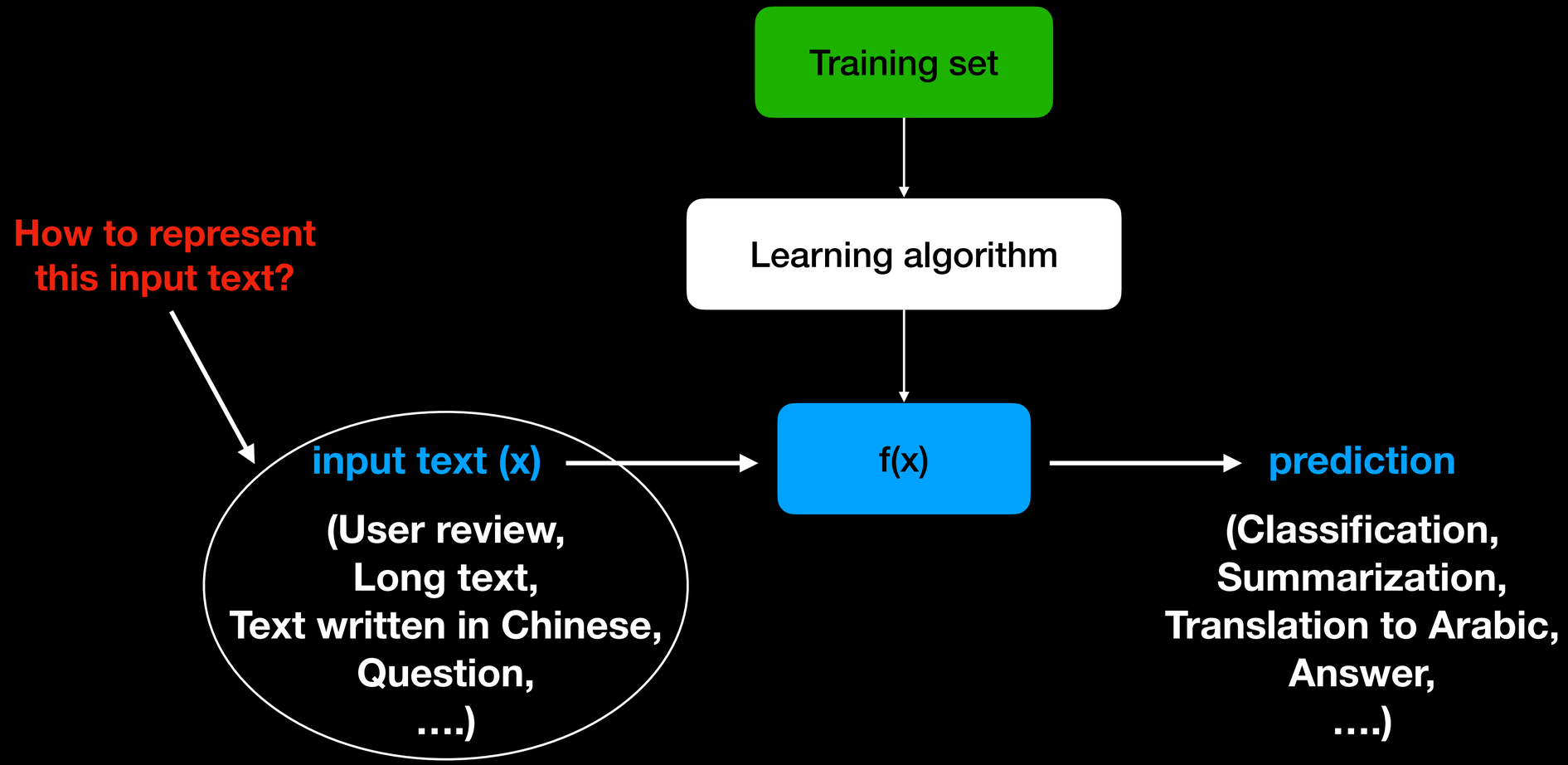
- Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

She gave the book to Deni.	vs.	She gave Deni the book.
Some kids popped by.	vs.	A few children visited.
Is that window still open?	vs.	Please close the window.

How to build an NLP system?

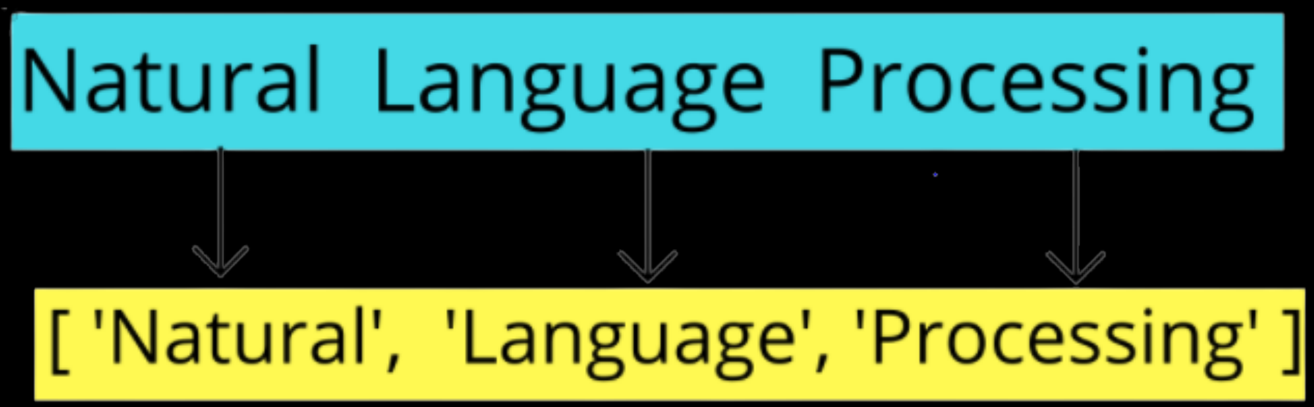


How to build an NLP system?



Tokenization

- Break a complex sentence into words.
- Understand the importance of each of the words with respect to the sentence.
- Produce a structural description of an input sentence.



One-hot encoding

- We've now converted our text to tokens. How can we convert a single token a feature vector?
  - Use one-hot encoding.
  - Length of feature vector = #distinct tokens (so can be pretty large, e.g. 10-100K).
  - Map token i to a feature vector x with entry i equal to 1 and other entries 0.
  - E.g. movie might map to feature vector [0, 1, 0, 0, 0, 0].

good	movie	not	a	did	like
0	1	0	0	0	0

Bag-of-Words

- How can we map a **sequence** of tokens (sentence, paragraph, document) to a feature vector.
  - Count number of occurrences of each token in our text -> later we'll then try to identify marker words like excellent or disappointed.
  - Length of feature vector = #distinct tokens
  - Entry i of vector is set equal to number of times token i appears in text. E.g.

A good good movie
Not a good movie
Did not like

good	movie	not	a	did	like
2	1	0	1	0	0
1	1	1	1	0	0
0	0	1	0	1	1

- Information about the word order is removed, that's why it's called a "bag of words".

Vector-based similarity

Train	A good good movie
	Not a good movie
	Did not like
Test	**A great movie**

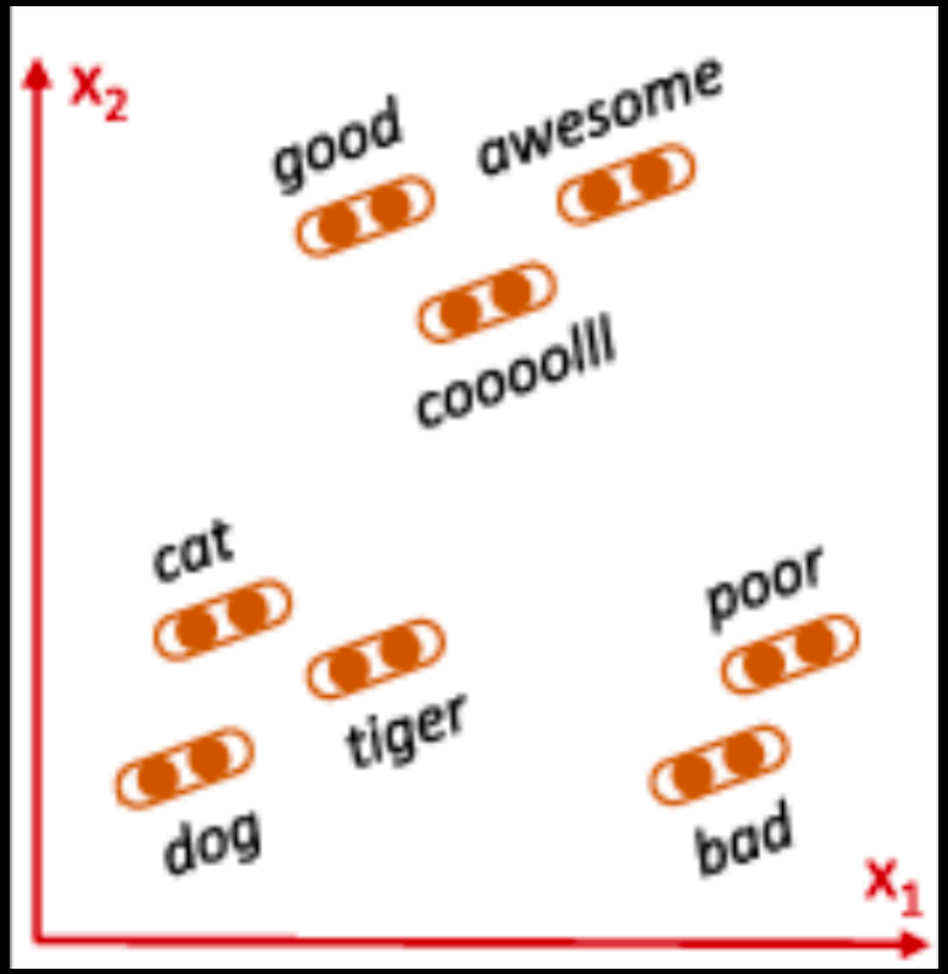
good	movie	not	a	did	like	great	*Label*
2	1	0	1	0	0	0	Pos
1	1	1	1	0	0	0	Neg
0	0	1	0	1	1	0	Neg
0	**1**	**0**	**1**	**0**	**0**	**1**	**?**

- What would an ML model predict for the last sentence?

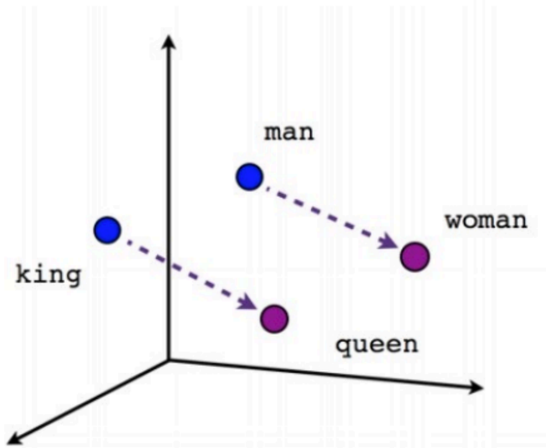
Cons of bag-of-words

- Large vocabulary <=> Large dimensions.
- All word vectors are orthogonal (no similarity).

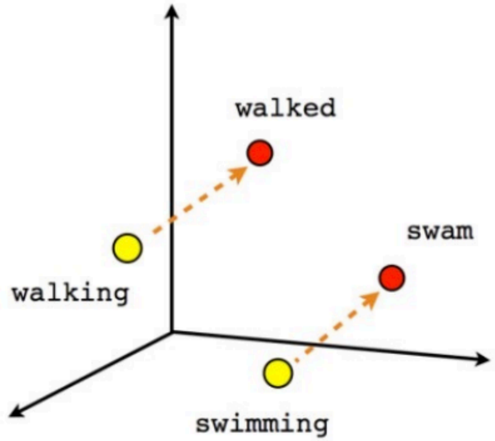
Word Embeddings



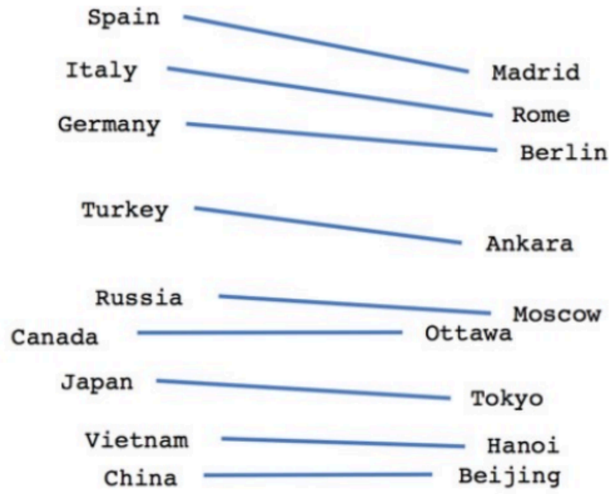
Word Embeddings



Male-Female



Verb tense



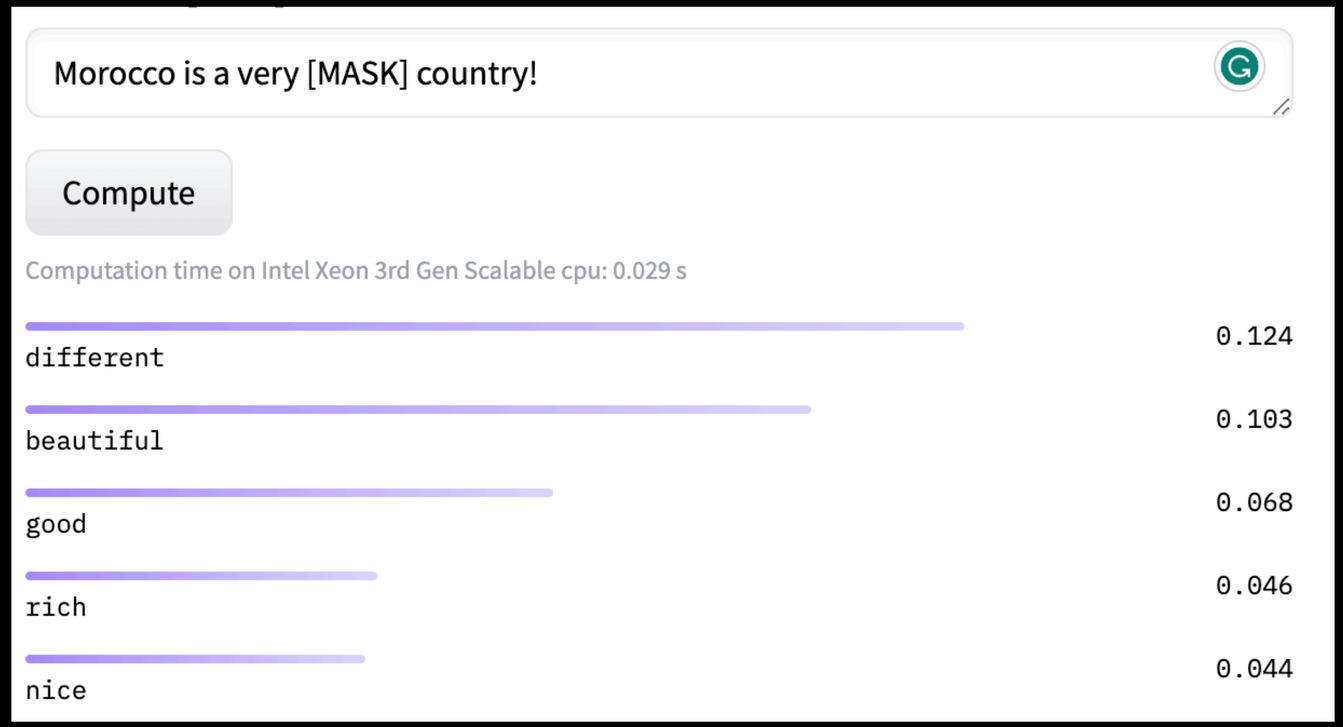
Country-Capital

Word Embedding - Fill the Blank

Morocco is a very … country!

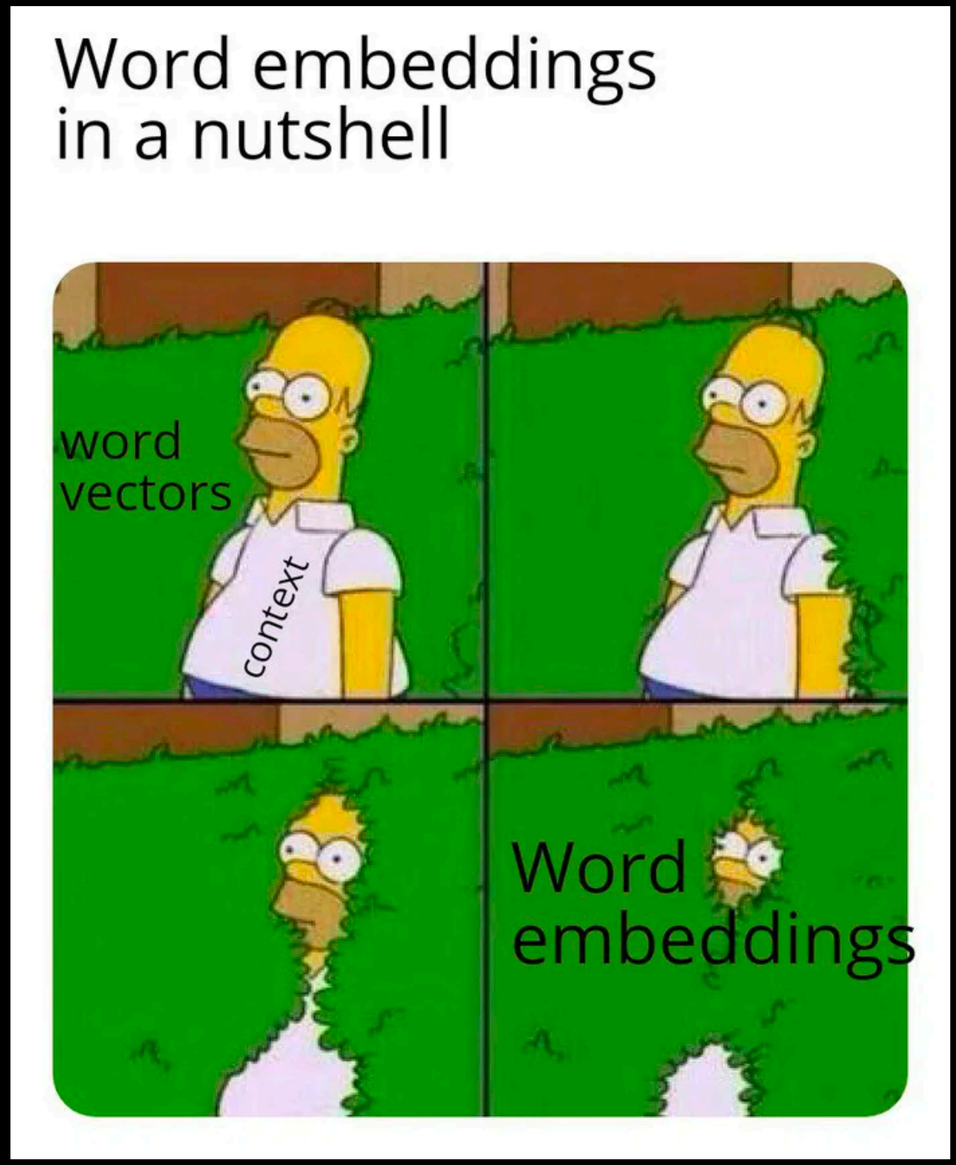
Word Embedding - Fill the Blank

Morocco is a very … country!

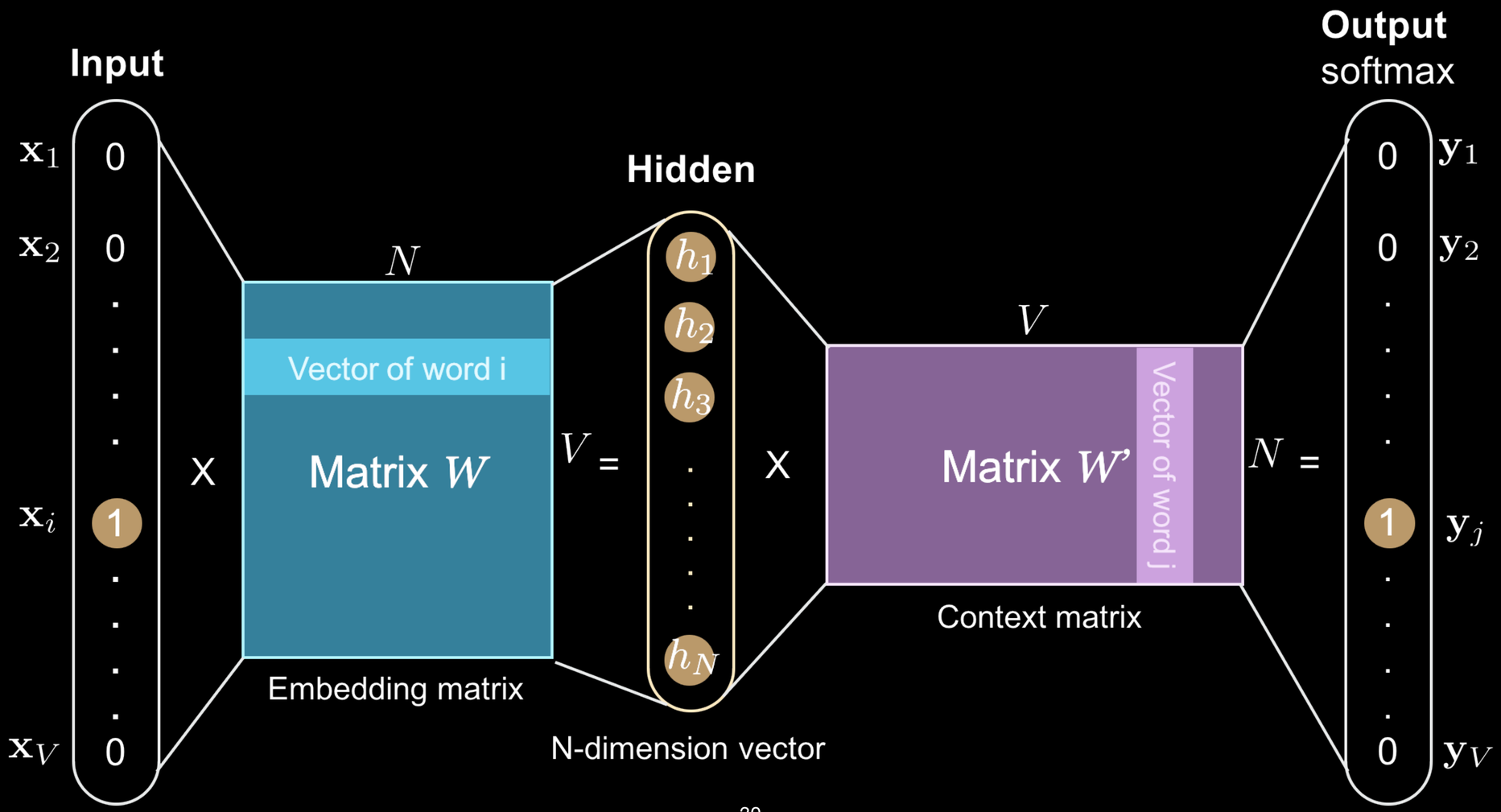


Word Embeddings

- Word2vec
  - CBOW
  - Skip-Gram
- Glove
- FastText



Word2vec (Skip-Gram)



How to use Word Embeddings?

This movie is good.

I did not liked this movie.

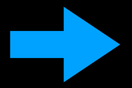
This movie is great.

How to use Word Embeddings?

This movie is good.

I did not liked this movie.

This movie is great.



[This, movie, is, good.]

[I, did, not, liked, this, movie.]

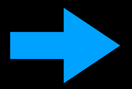
[This, movie, is, great.]

How to use Word Embeddings?

This movie is good.

I did not liked this movie.

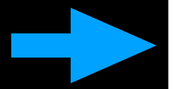
This movie is great.



[This, movie, is, good.]

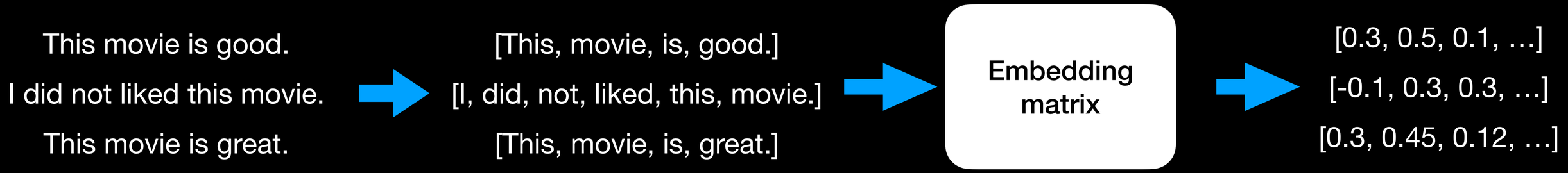
[I, did, not, liked, this, movie.]

[This, movie, is, great.]

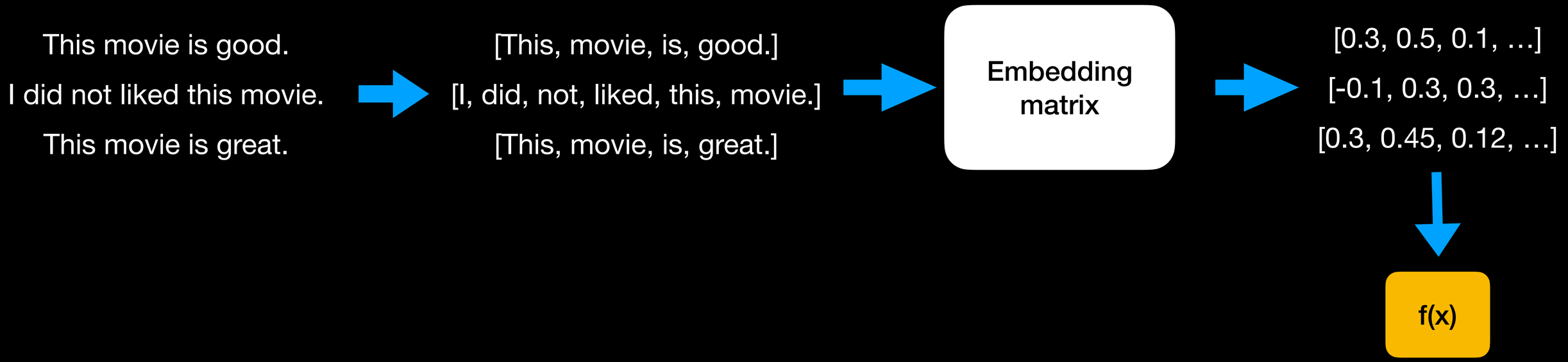


Embedding matrix

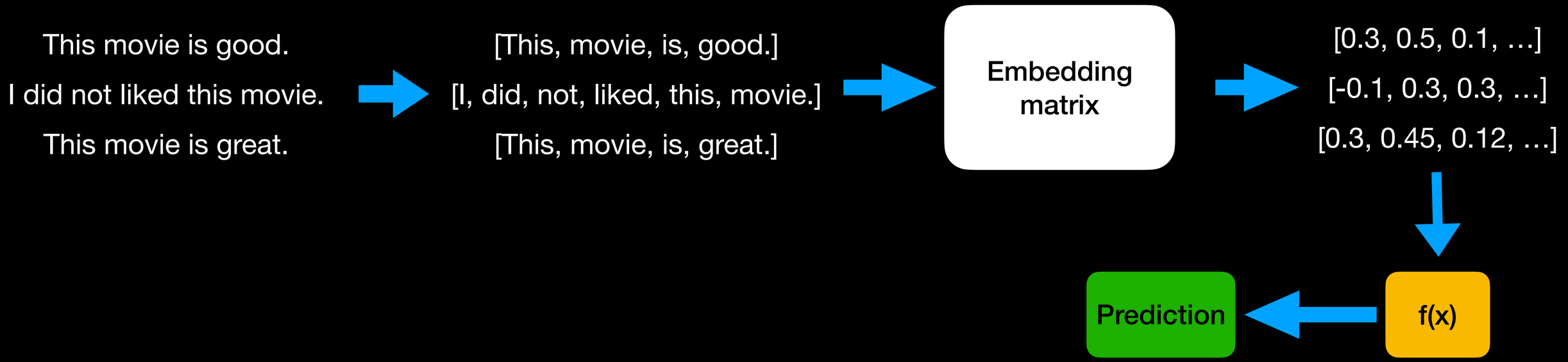
How to use Word Embeddings?



How to use Word Embeddings?



How to use Word Embeddings?



Transformers and Attention

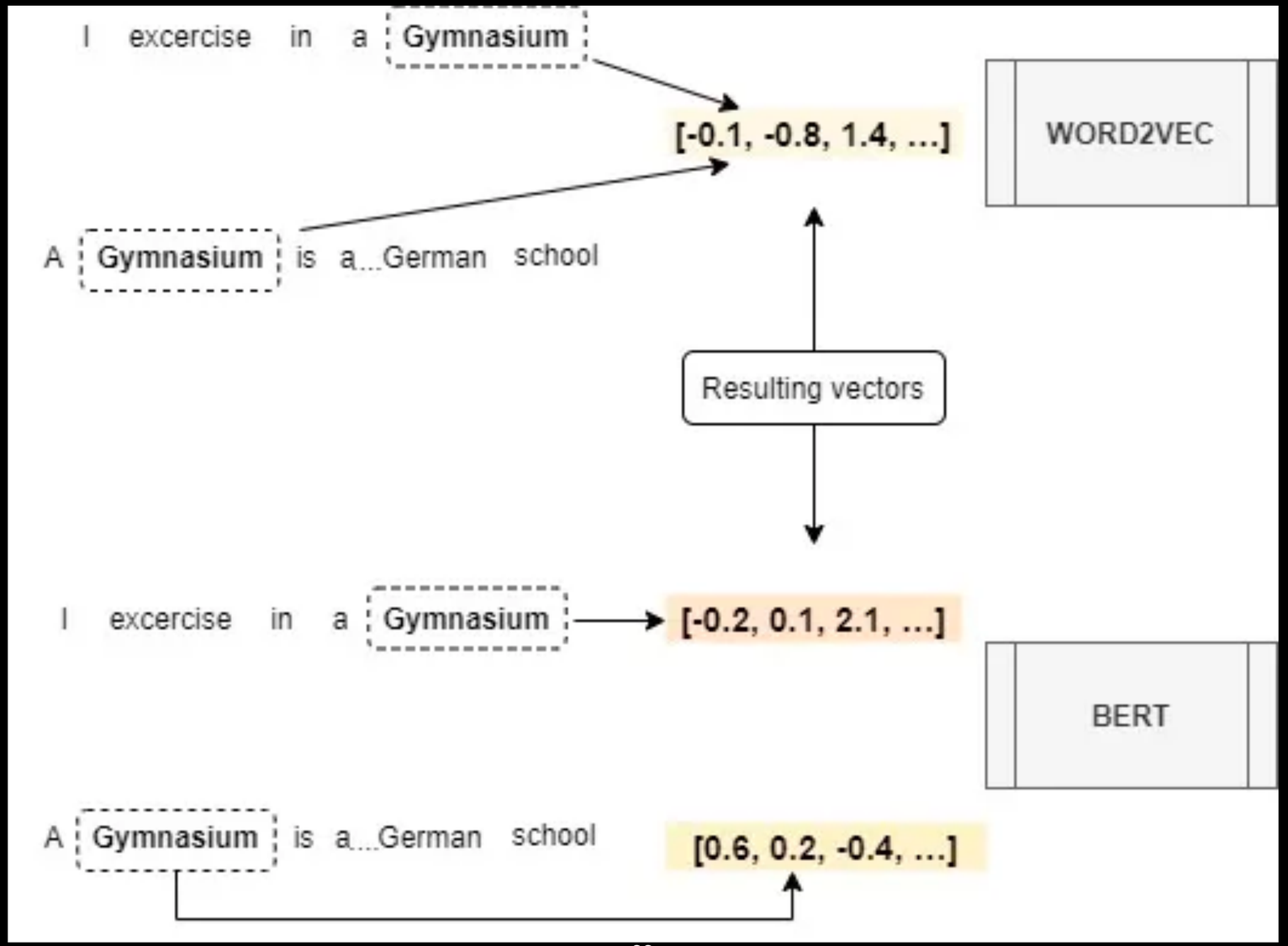
[-0.1, 0.3, 1.5, ...]

We went to the river **bank**.

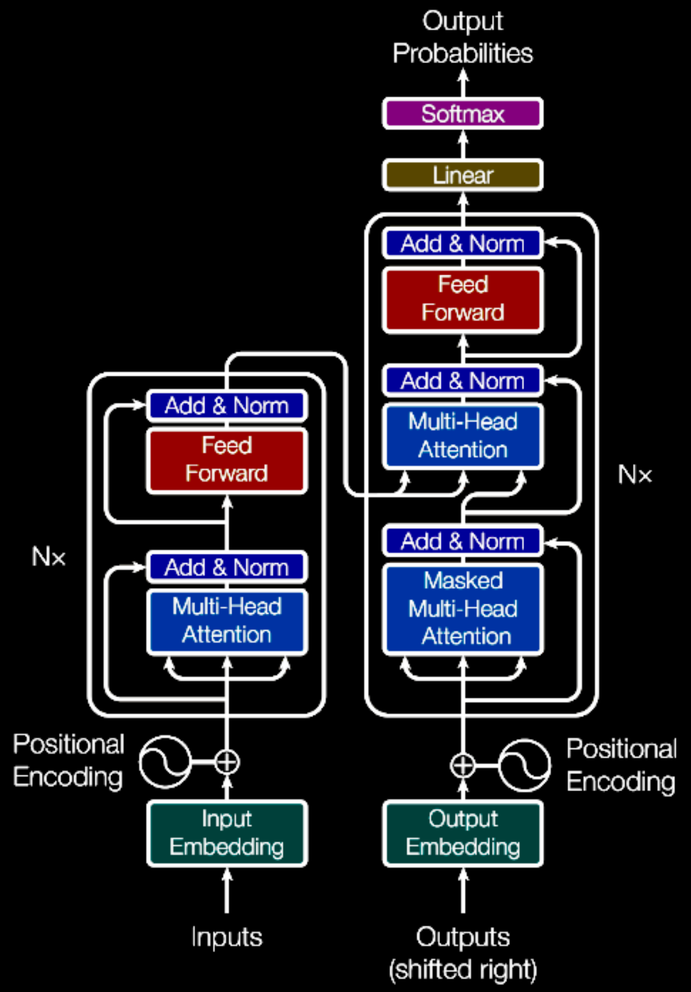
I need to go to the **bank** to make a deposit.

Word2vec

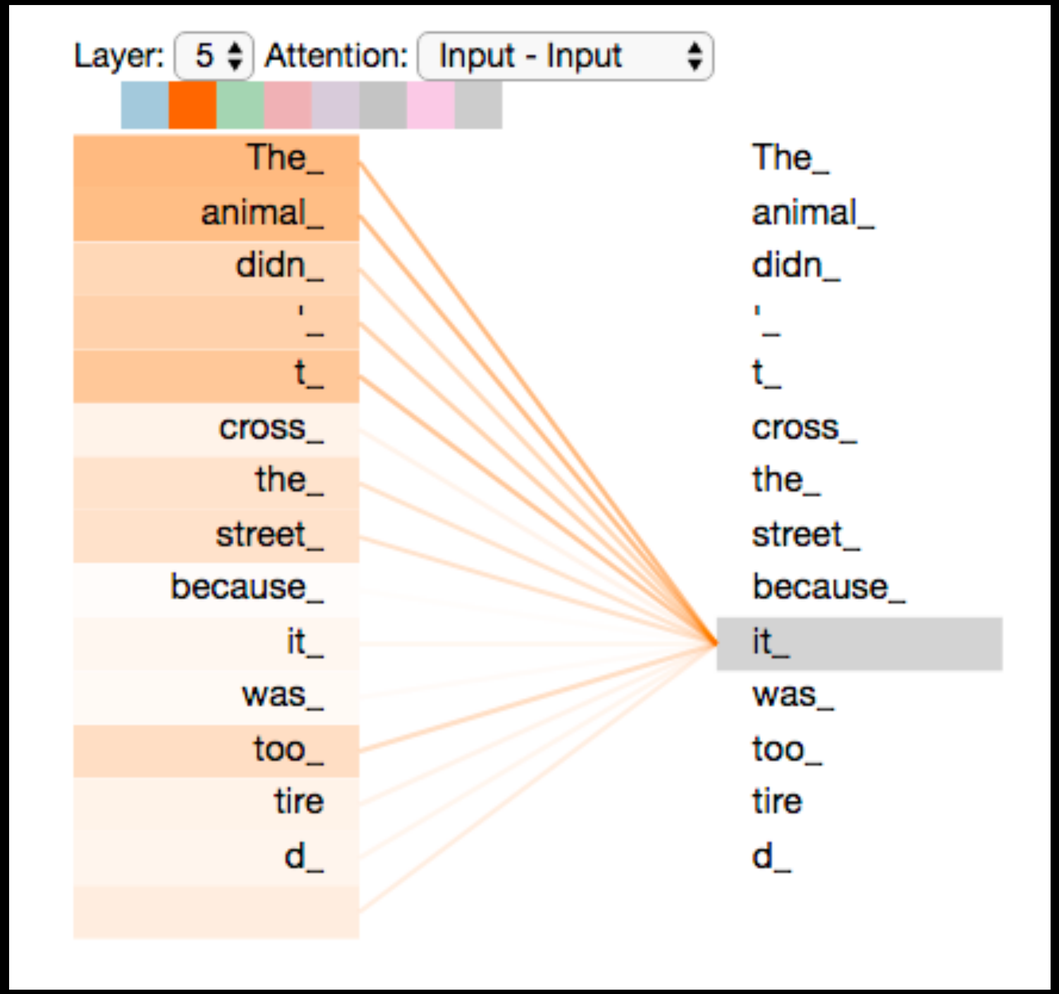
Transformers and Attention



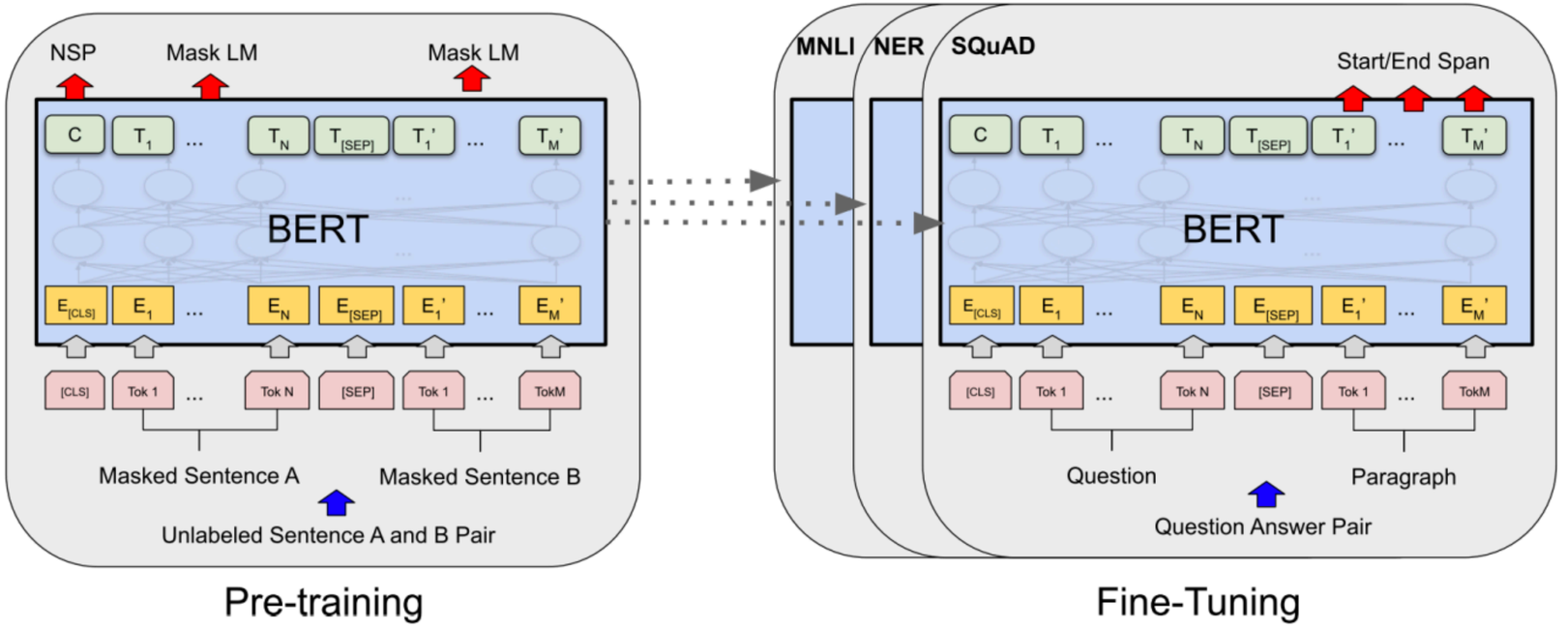
Transformers and Attention



Transformers and Attention



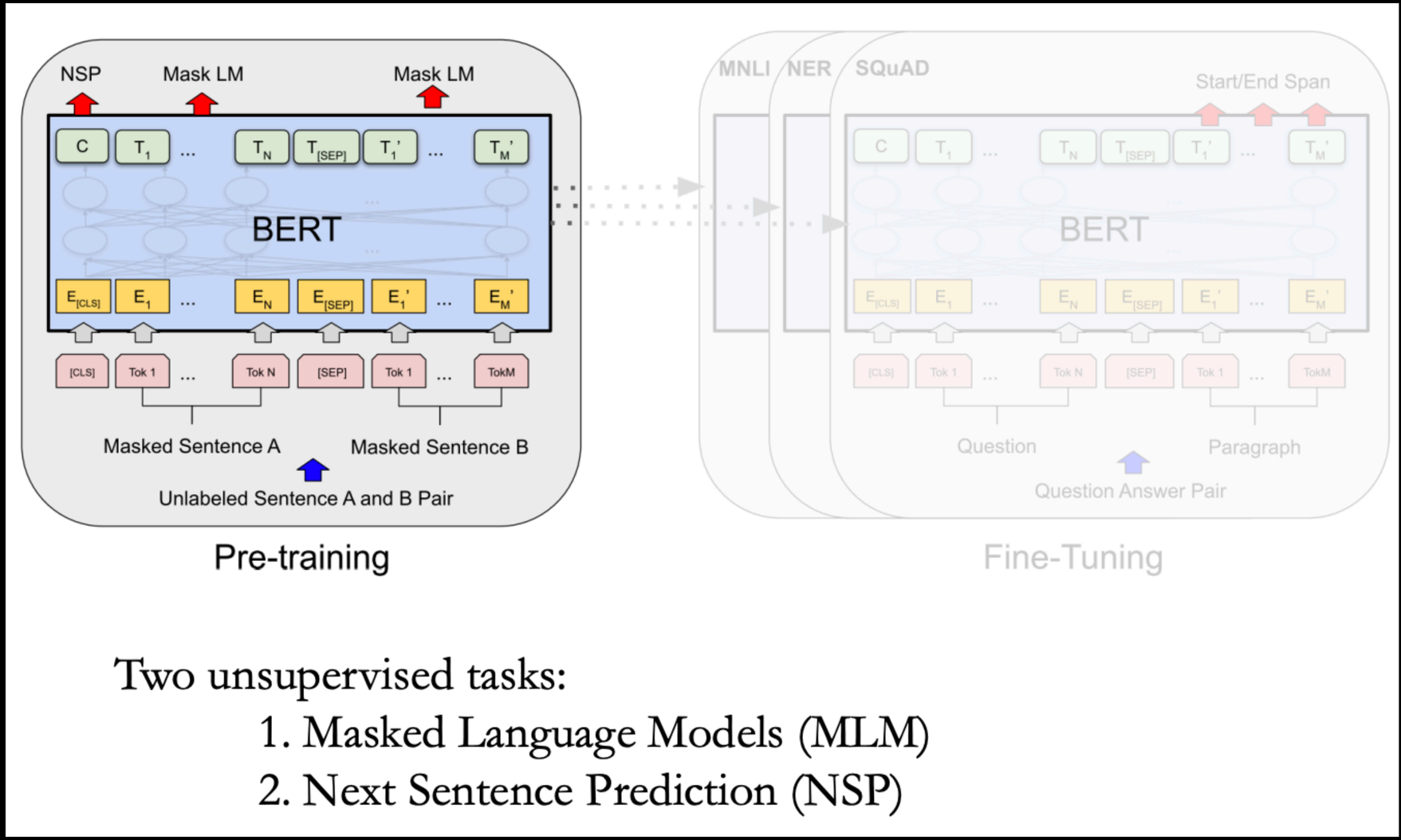
BERT: Framework



Pre-training : trained on unlabeled data over different pre-training tasks.

Fine-Tuning : fine-tuned parameters using labeled data from the downstream tasks.

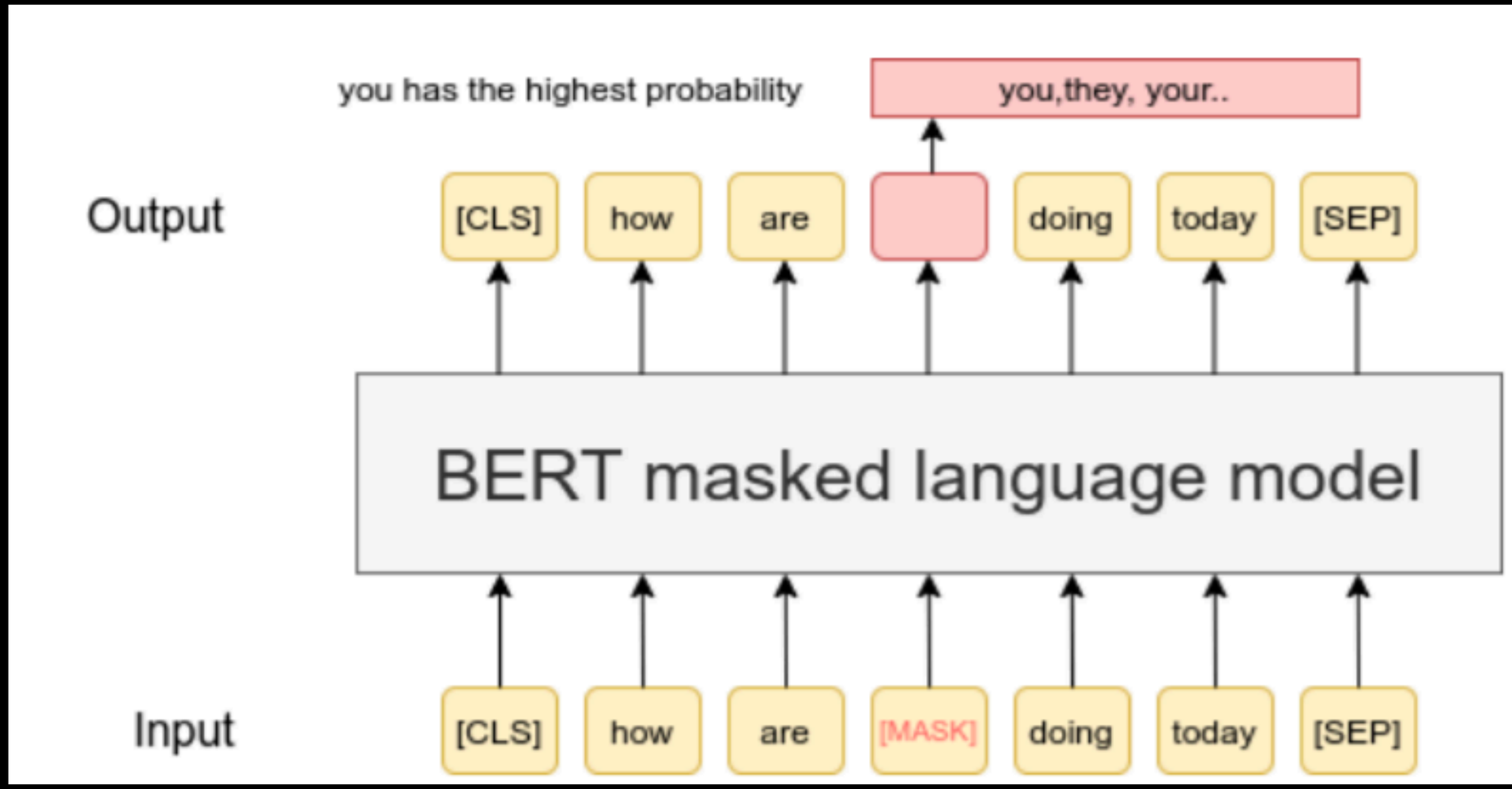
BERT: Pre-Training



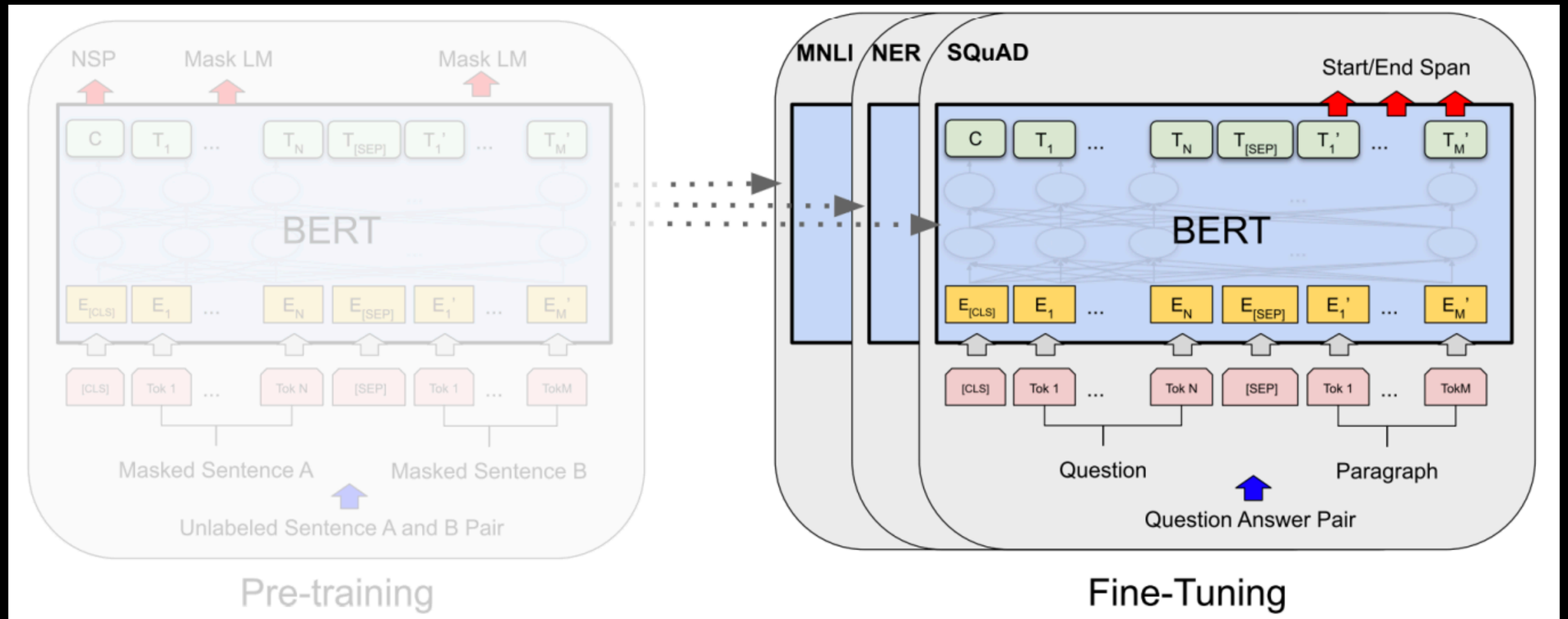
Two unsupervised tasks:

1. Masked Language Models (MLM)
2. Next Sentence Prediction (NSP)

BERT: Masked Language Modeling (MLM)

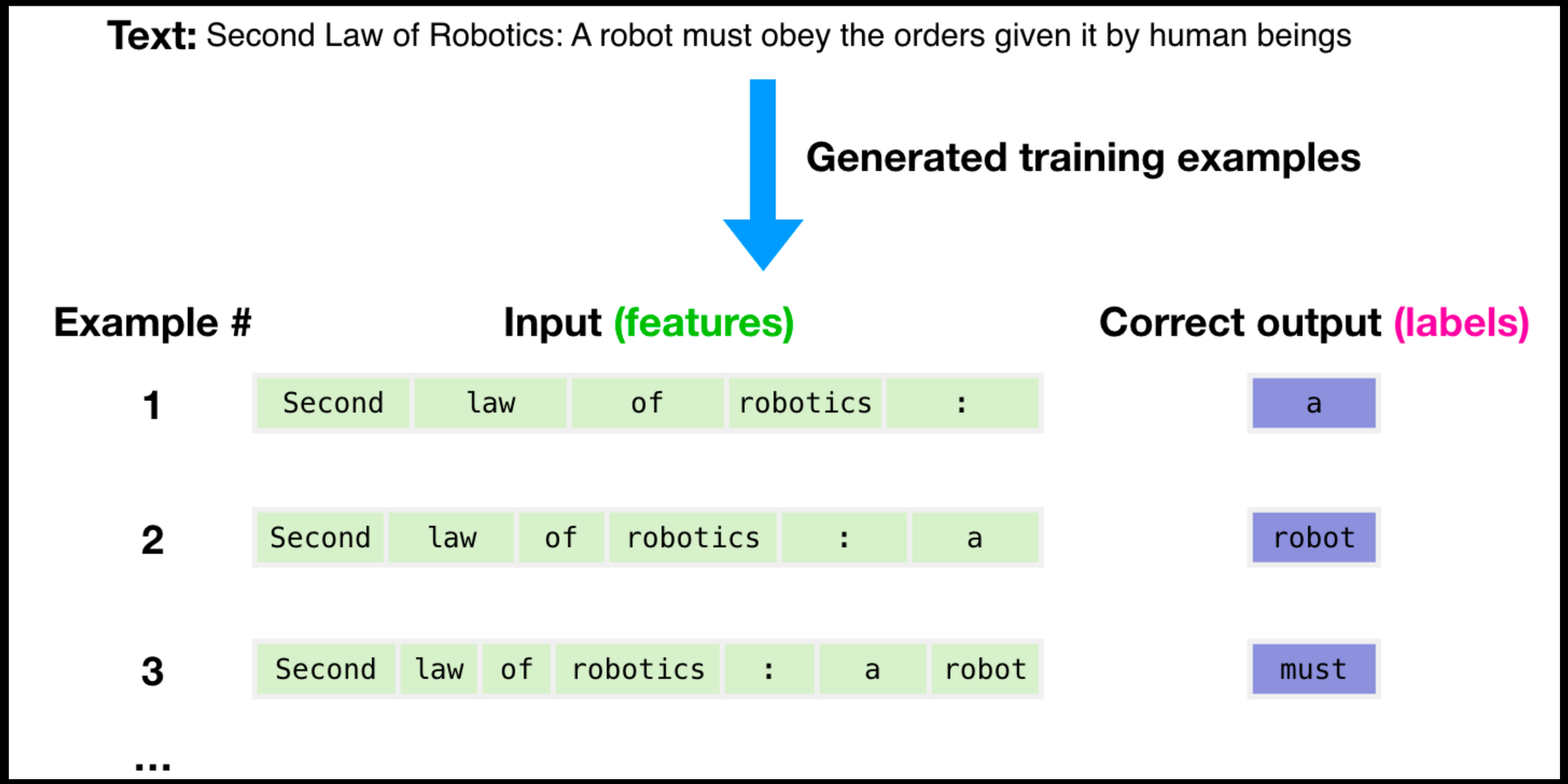


BERT: Fine-Tuning



Fine-Tuning : fine-tuned parameters using labeled data from the downstream tasks.

GTP-3: Training



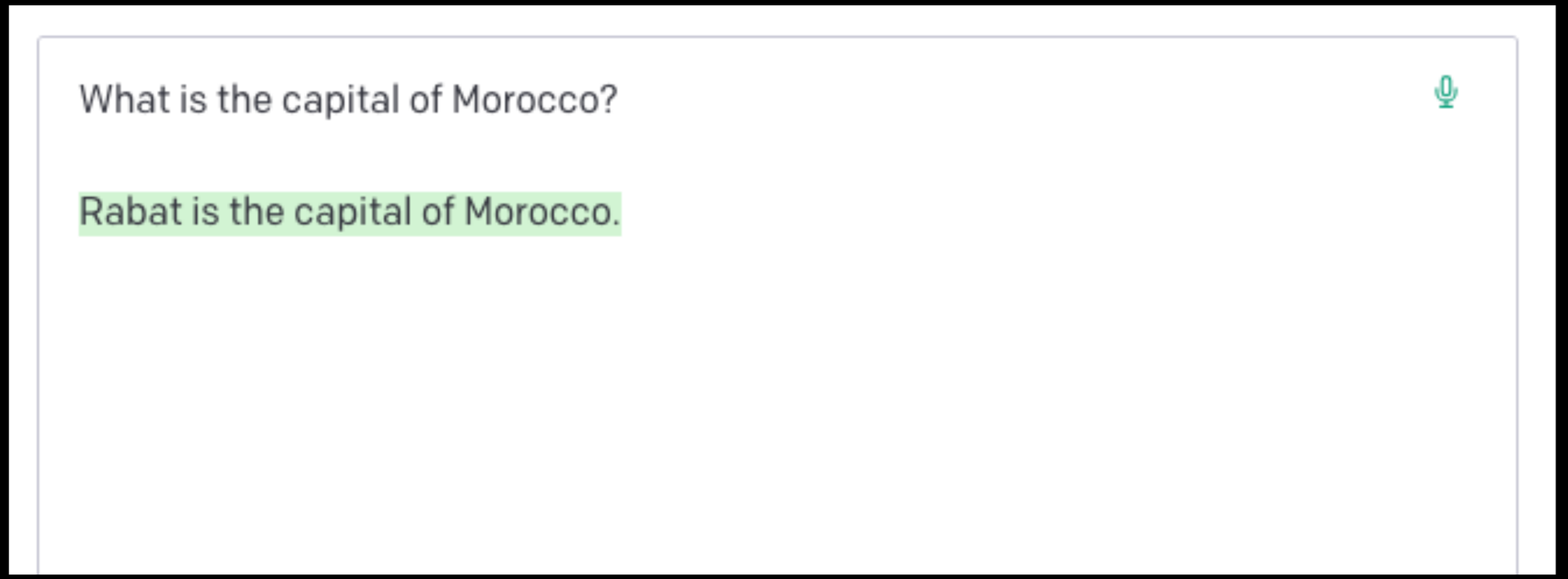
GPT-3: Base Model

What is the capital of Morocco? (Hint: The word Morocco is borrowed from Arabic.)

Stop by the Towers for evening ice cream service from 6:00 PM to 6:30 P.M. and enjoy the historic elegance of the old century room which features wood moldings and a stone fireplace. We will close at 8:45 PM due to the last train from Skyline Drive at 9:15.

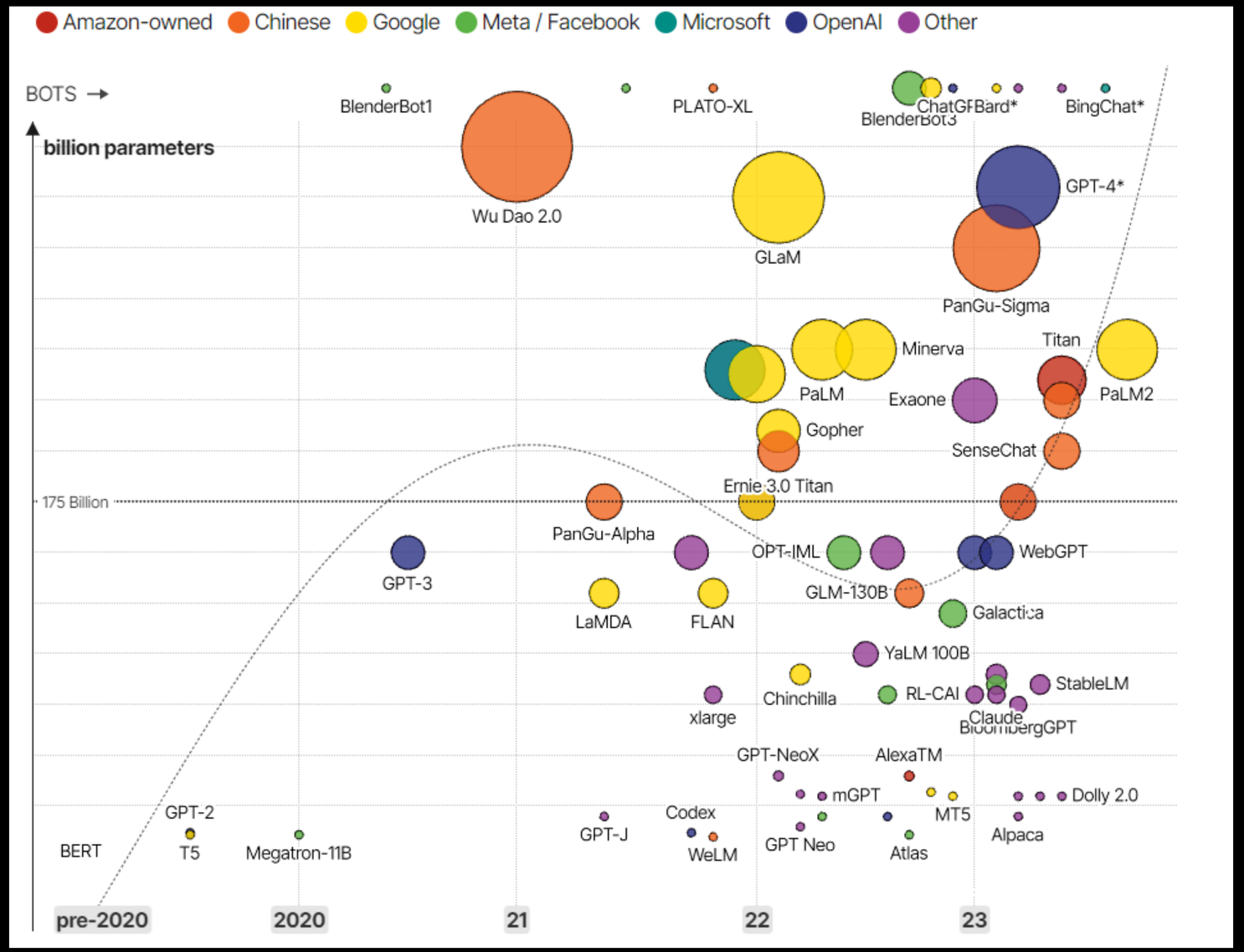
After dinner, we will head out to see the movie. Click here for a trailer...

GPT-3: ChatGPT



How can we leverage NLP achievements for underrepresented languages?

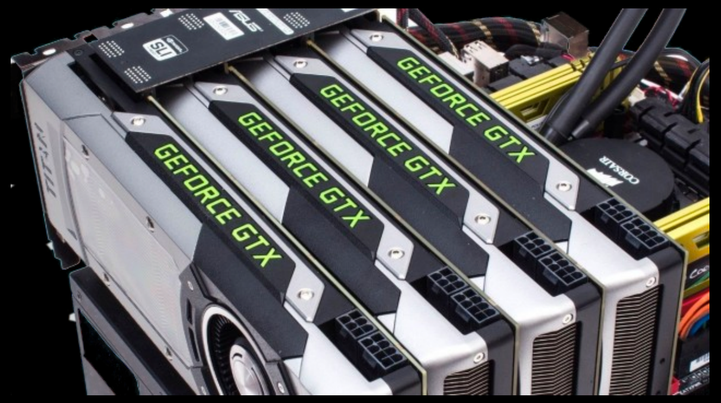
NLP ⇔ Large Language Models



How to build a similar language model?

LARGE TEXTUAL CORPORA

&

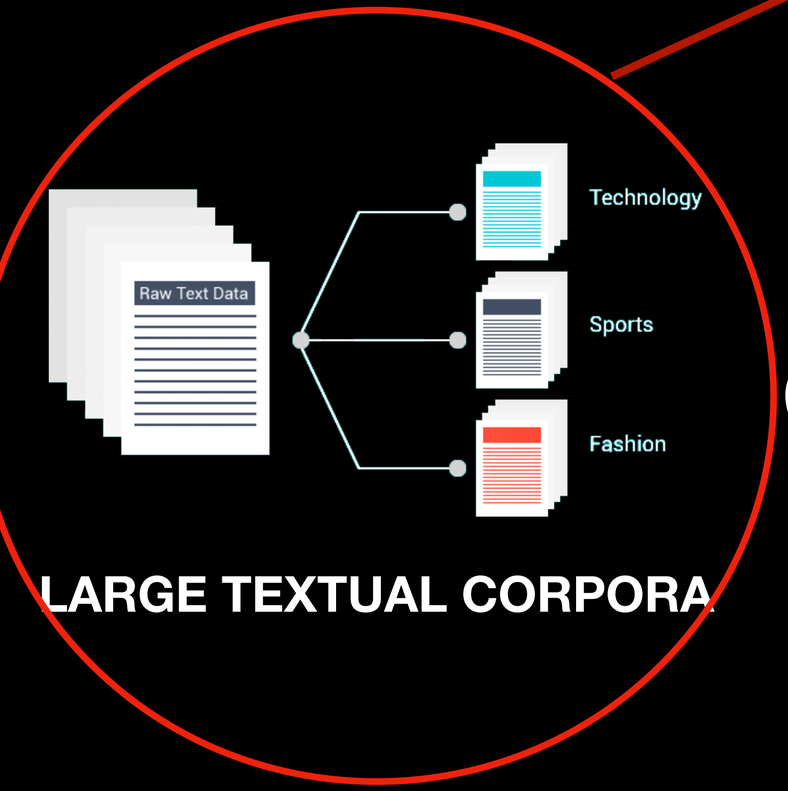


MANY GPUS …

=>

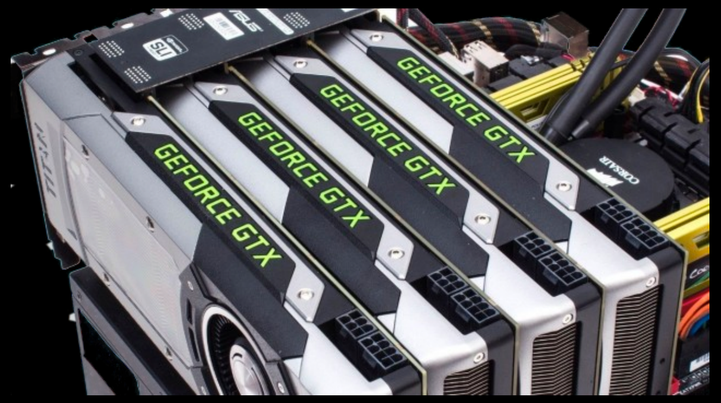
How to build a similar language model?

Hard to collect for underrepresented languages
(Arabic dialects, African languages…)



LARGE TEXTUAL CORPORA

&



MANY GPUS …

=>

Our team: UM6P-CC

College of Computing
UM6P

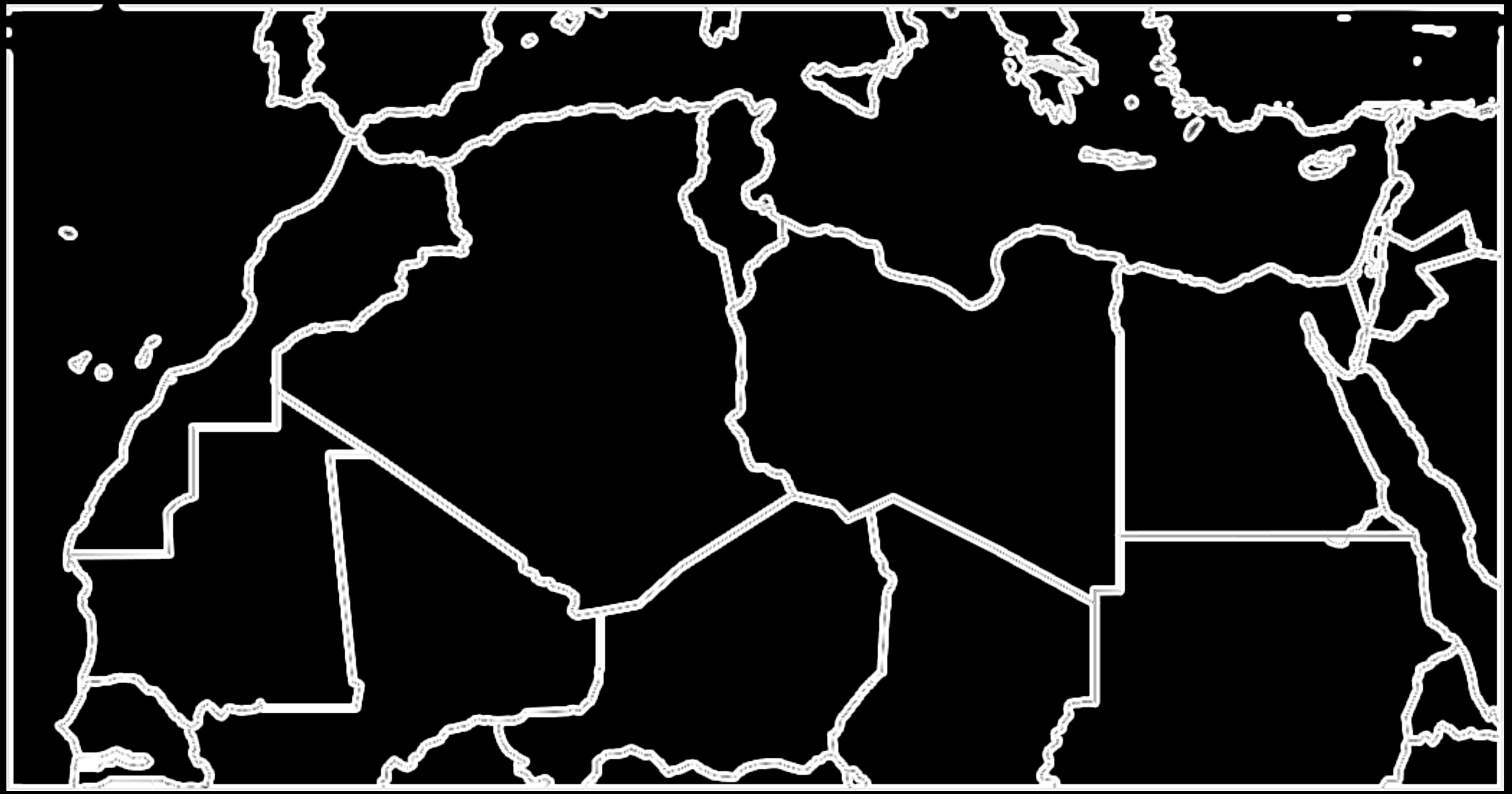
- Dr. Abdelkader EL MAHDAOUY
- Post-Doctoral Researcher In Natural Language Processing / Machine Learning
- University Mohammed VI Polytechnic, Morocco

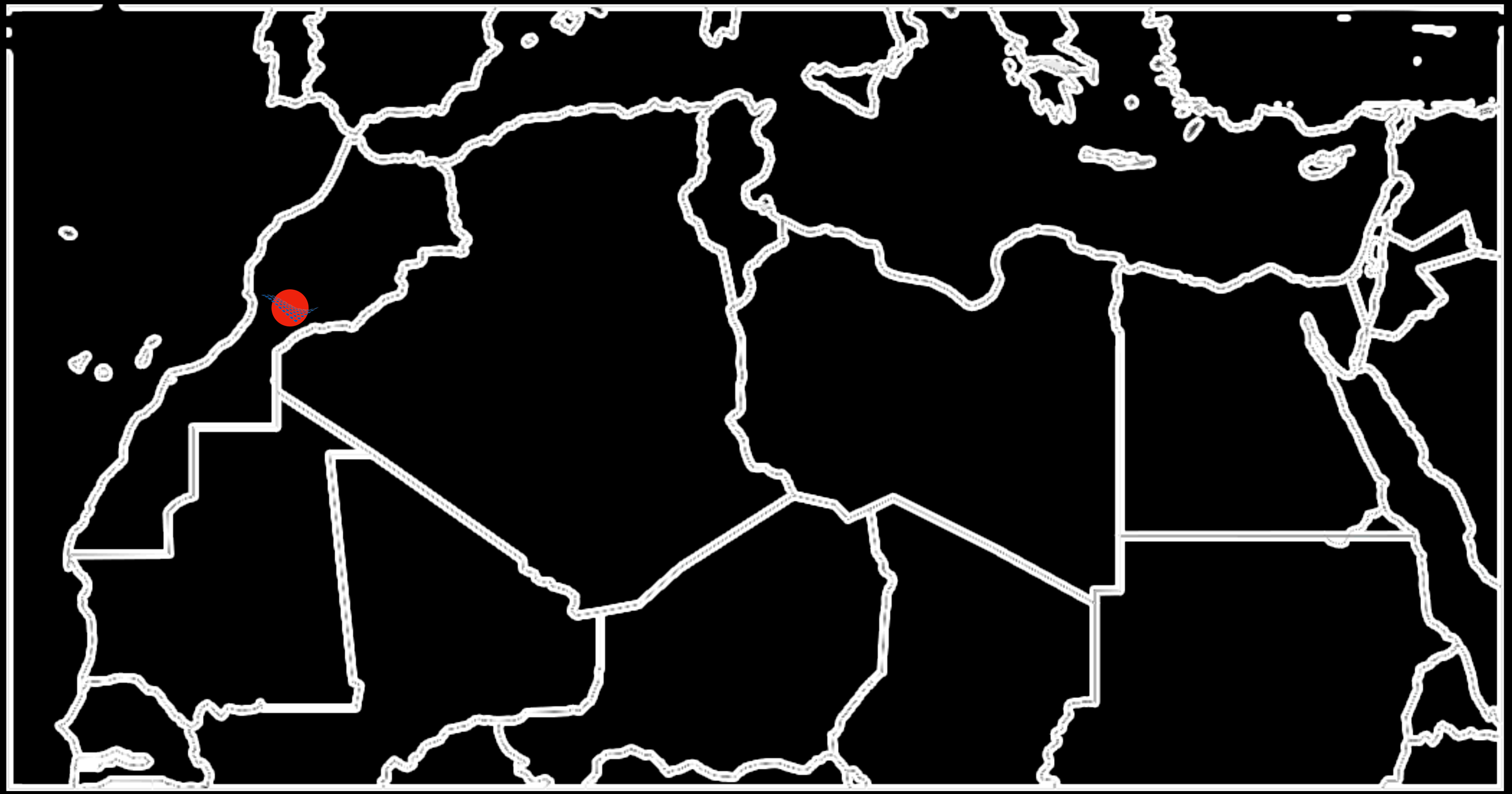
- Dr. Ismail BERRADA
- UM6-CC professor and researcher in applied ML
- University Mohammed VI Polytechnic, Morocco

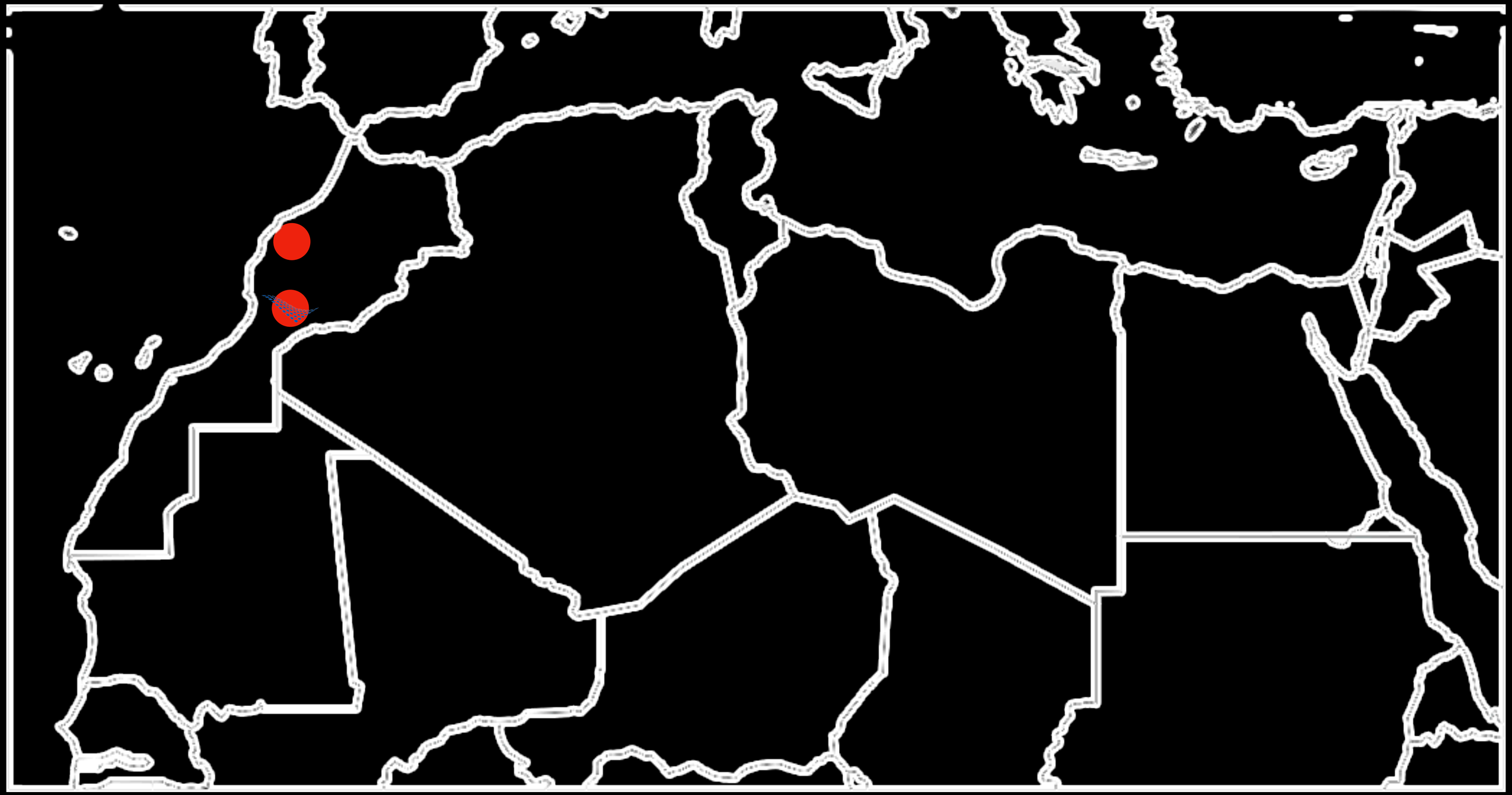
- Abdellah EL MEKKI
- PhD Student in Natural Language Processing / Machine Learning
- University Mohammed VI Polytechnic, Morocco

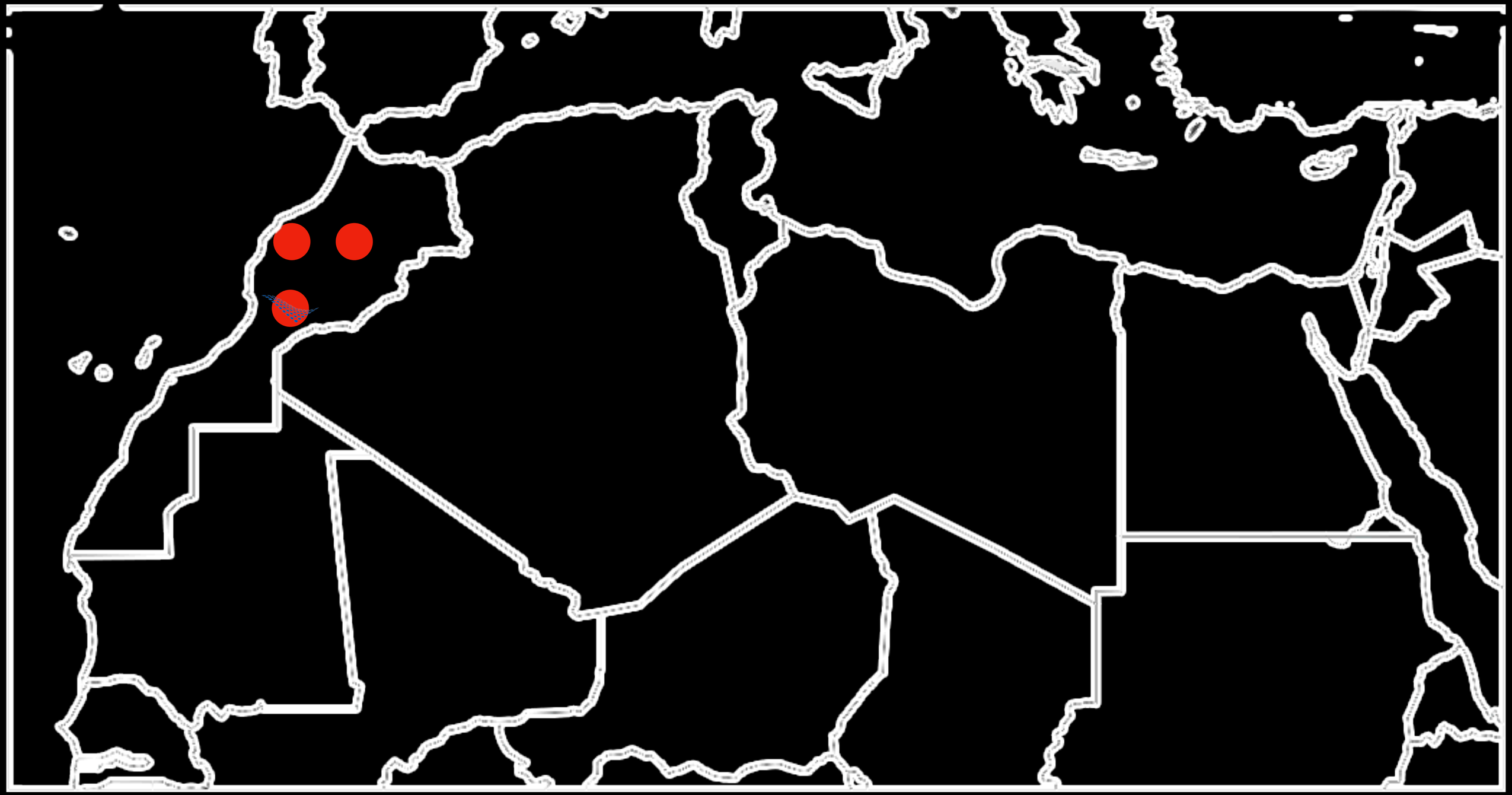
Modern Standard Arabic (MSA) NLP

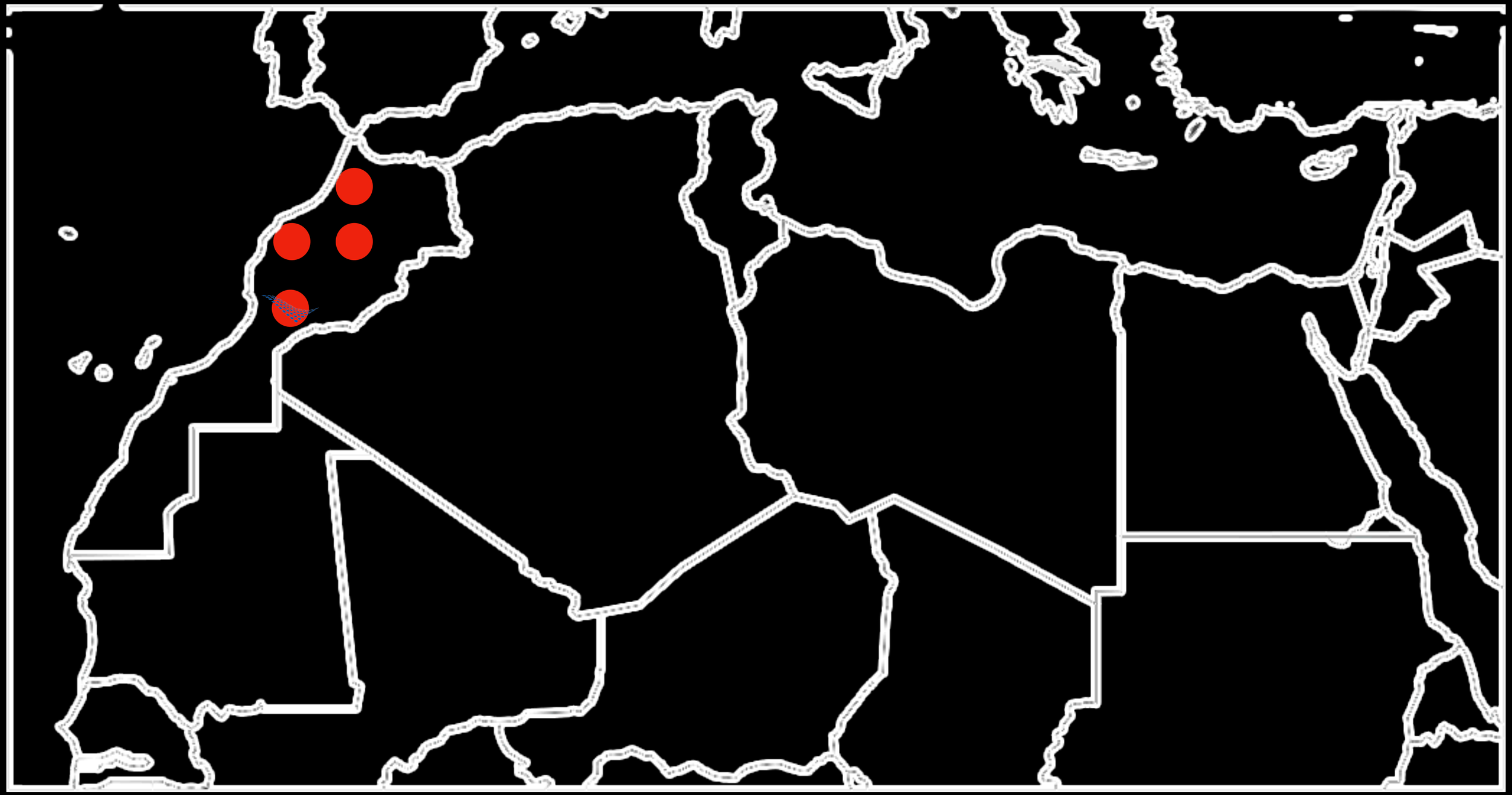
- Good Machine Translation results.
- Good conversation systems.
- Availability of corpora to train Arabic NLP systems on.

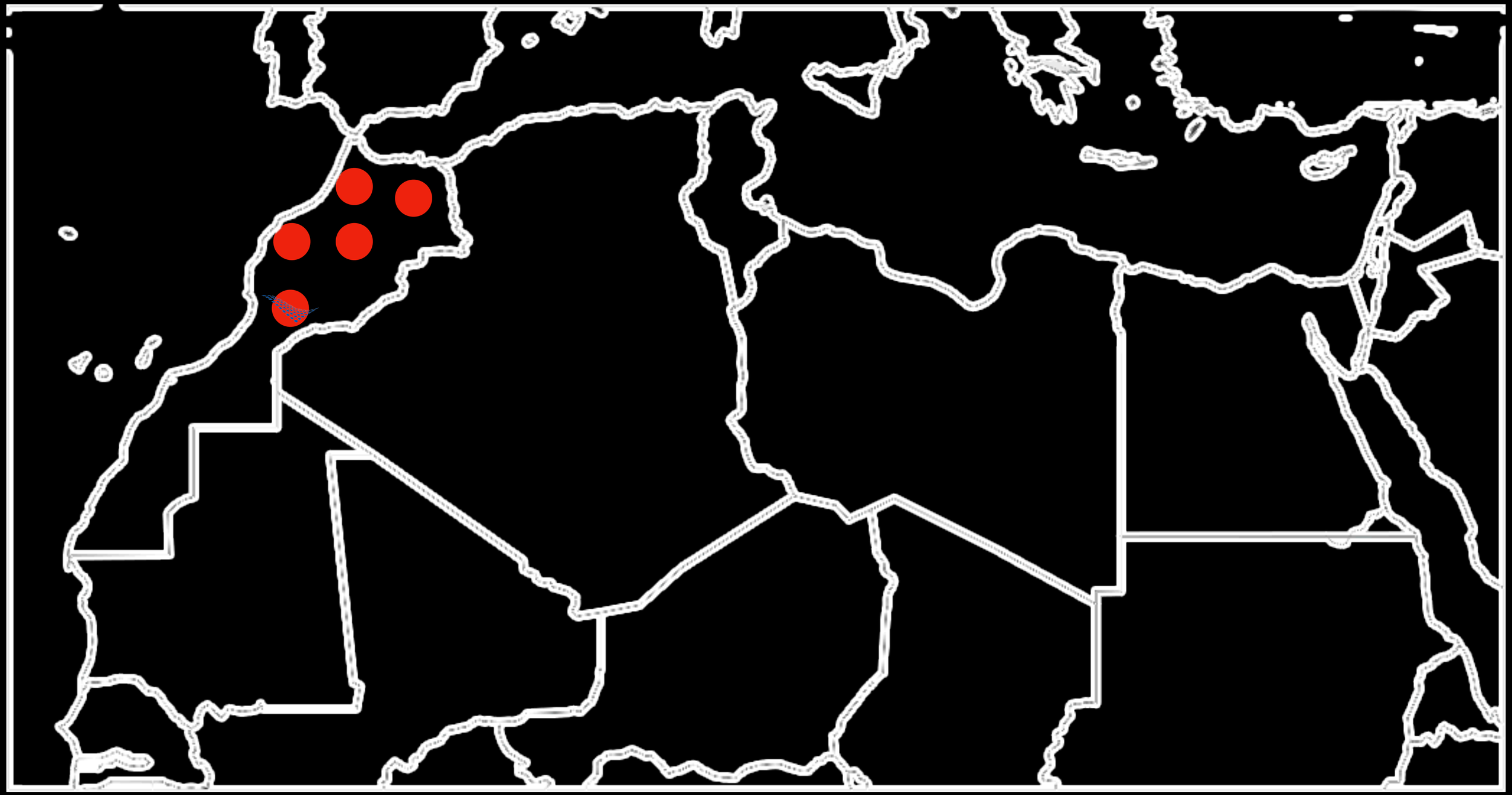


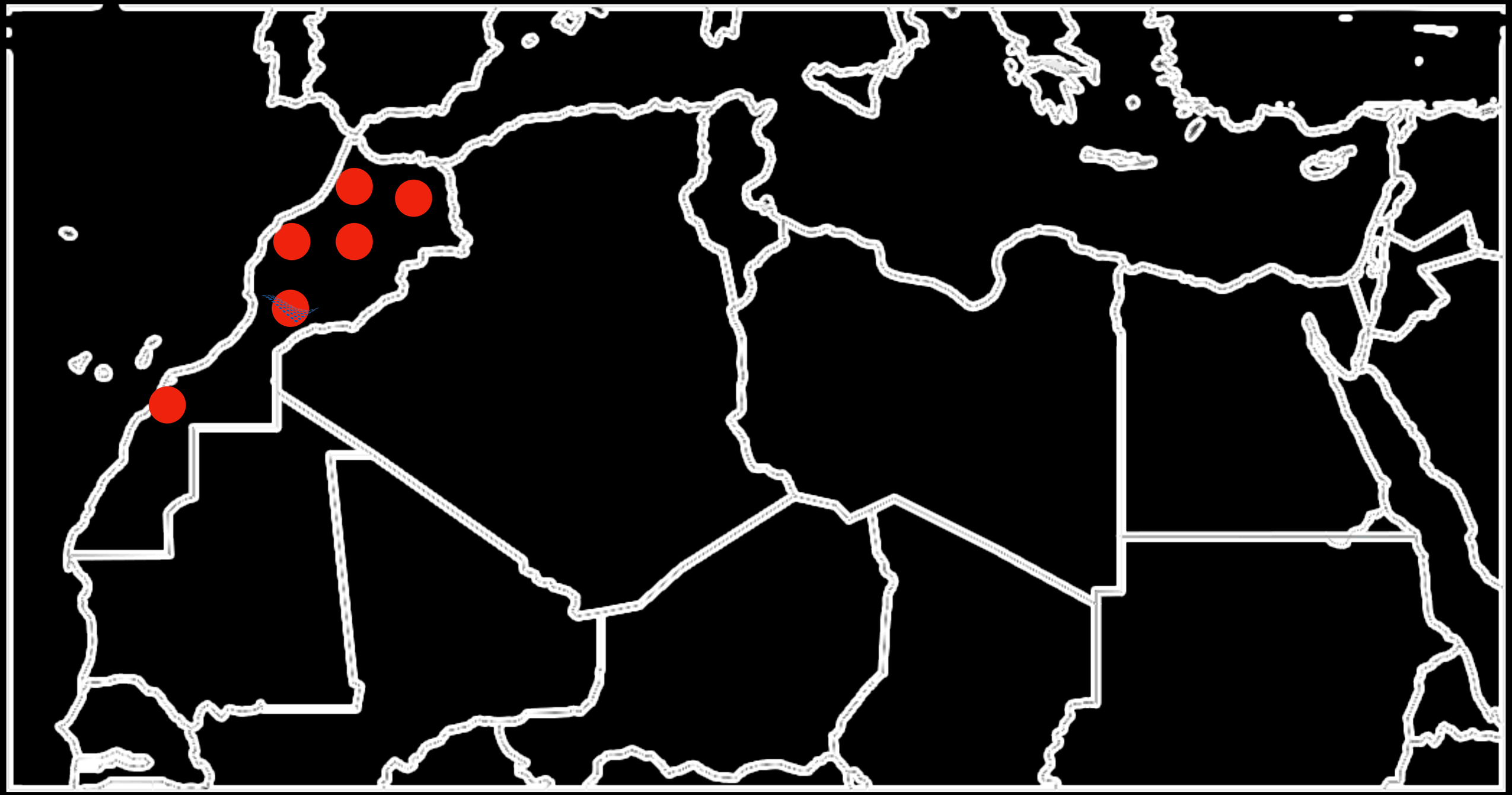


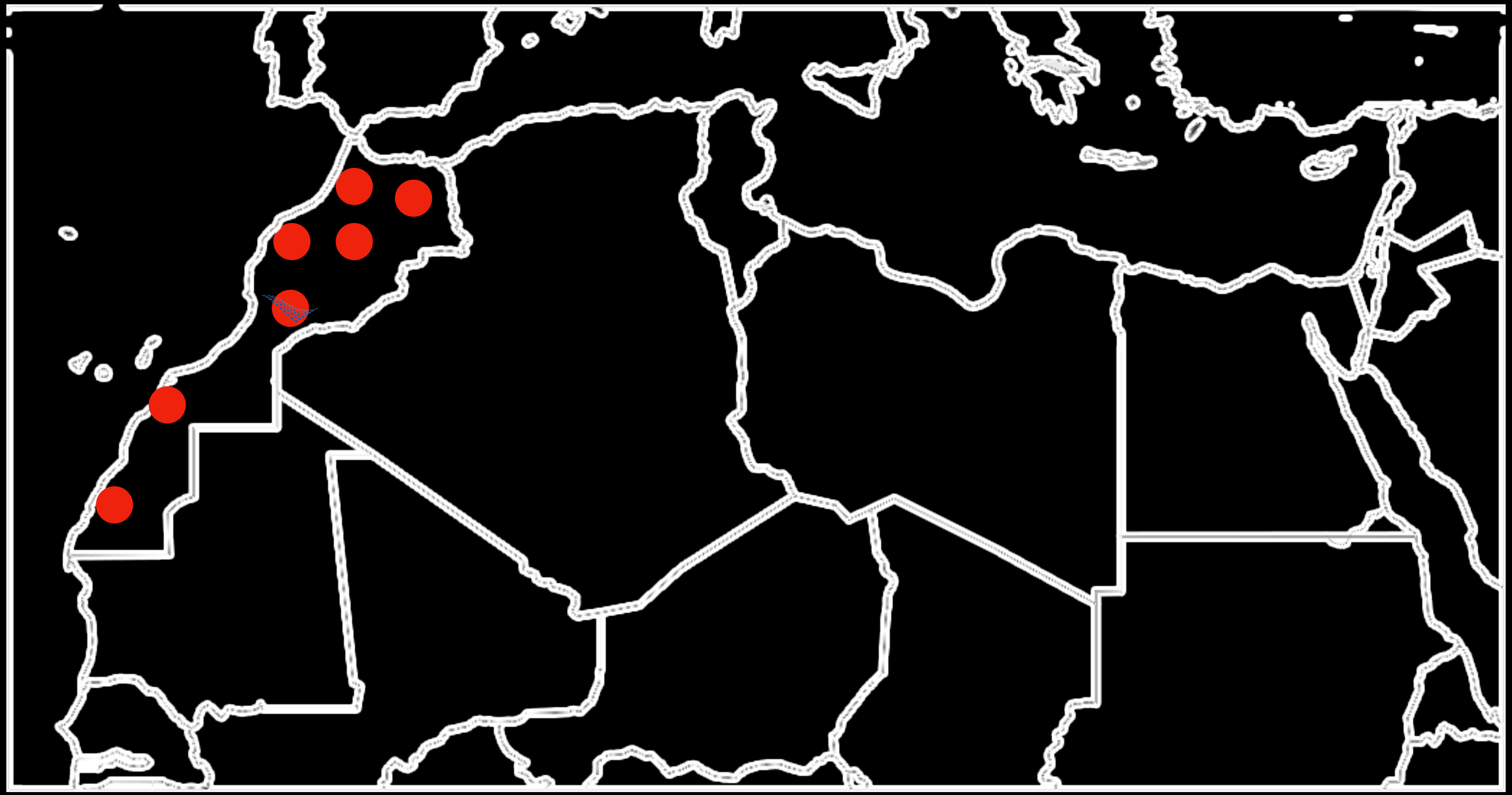


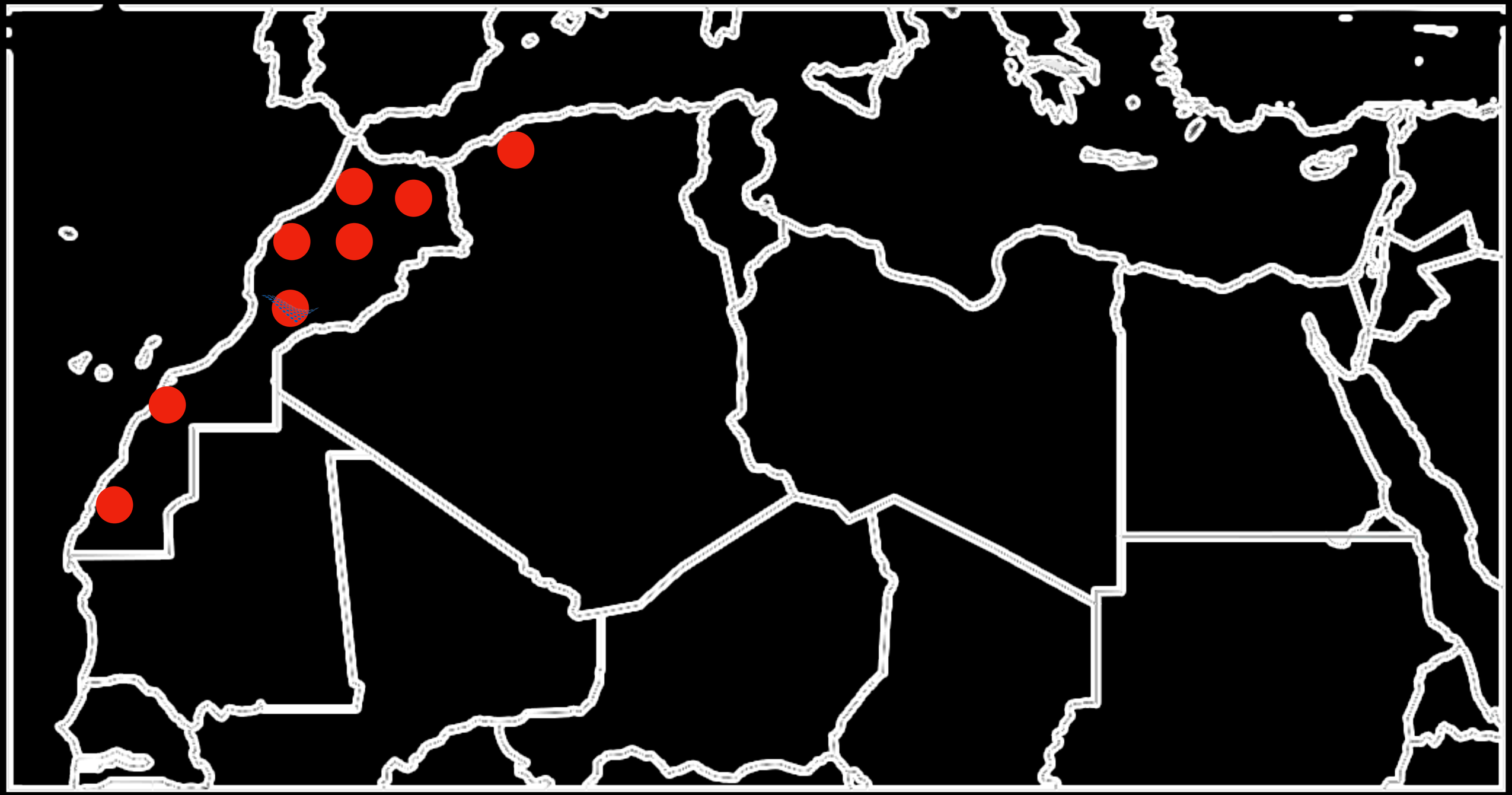


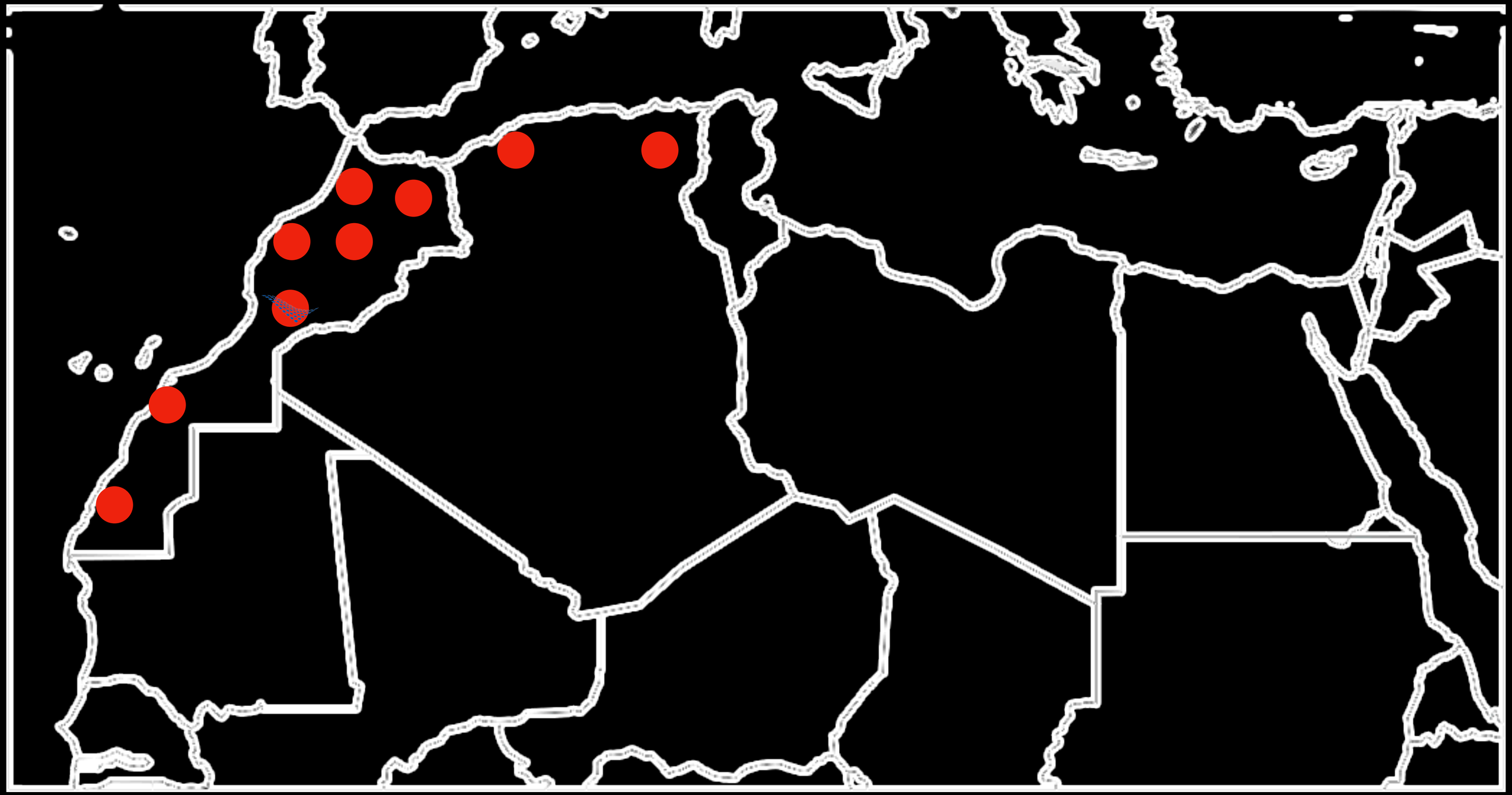


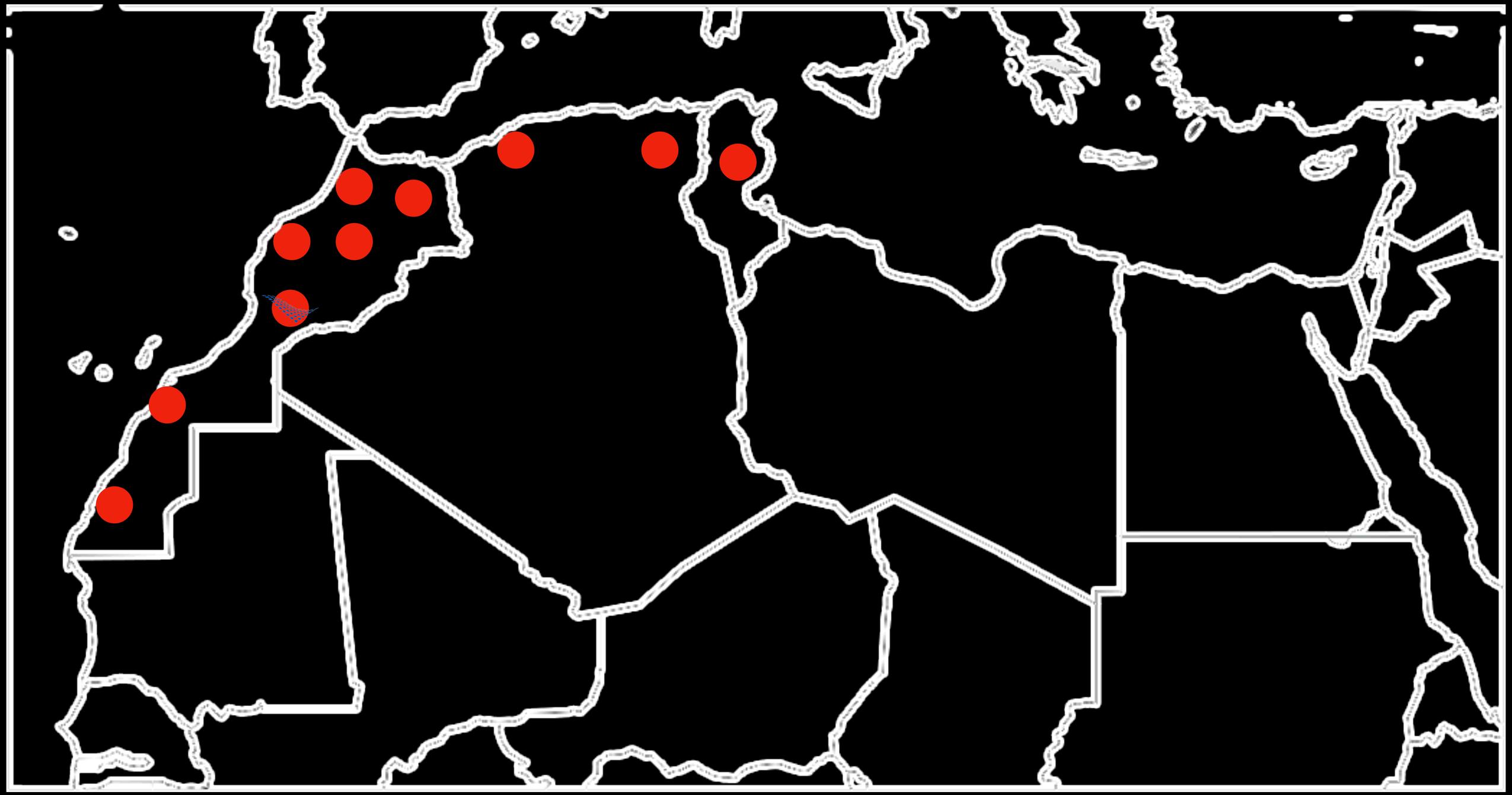


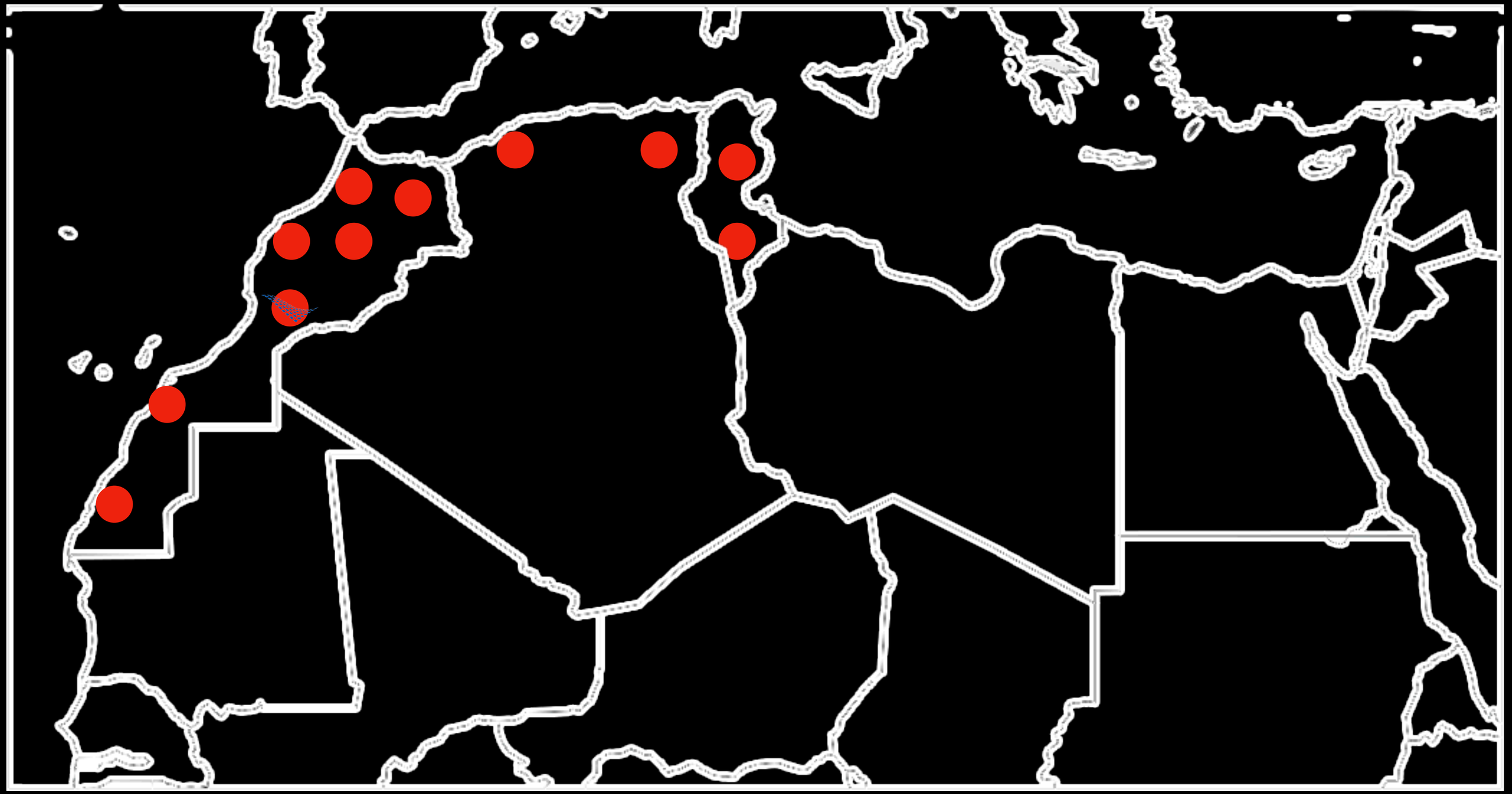


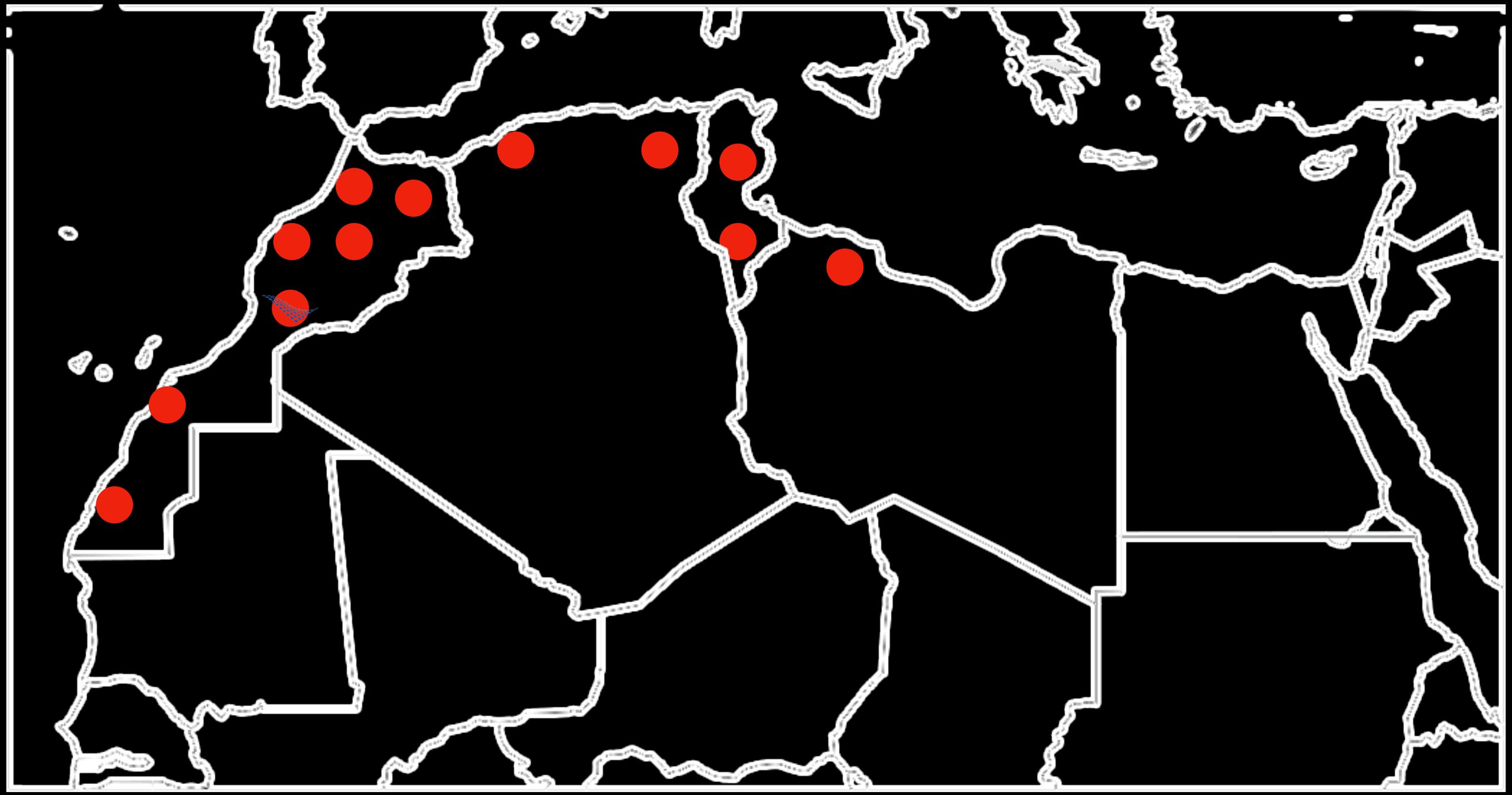


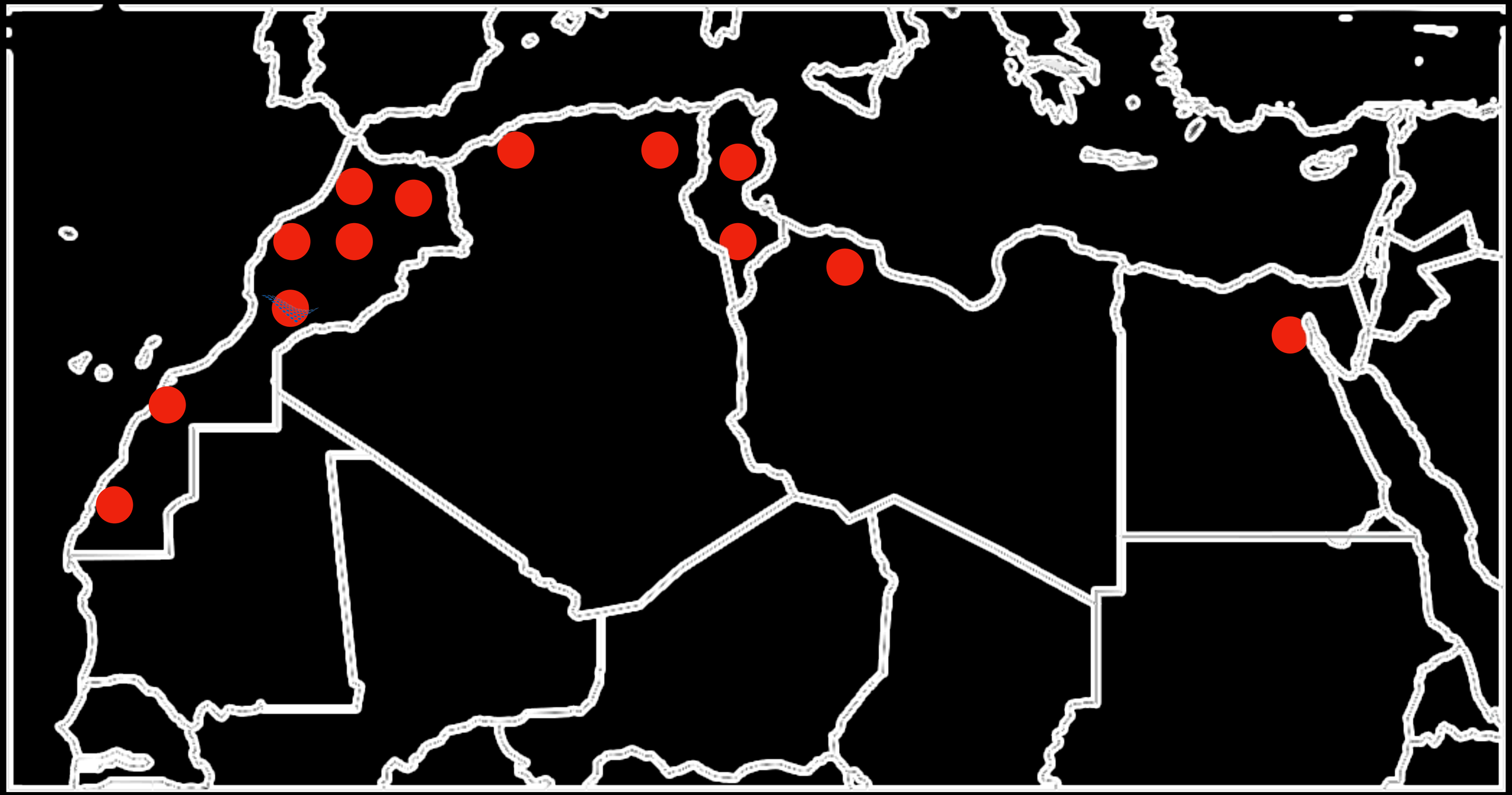


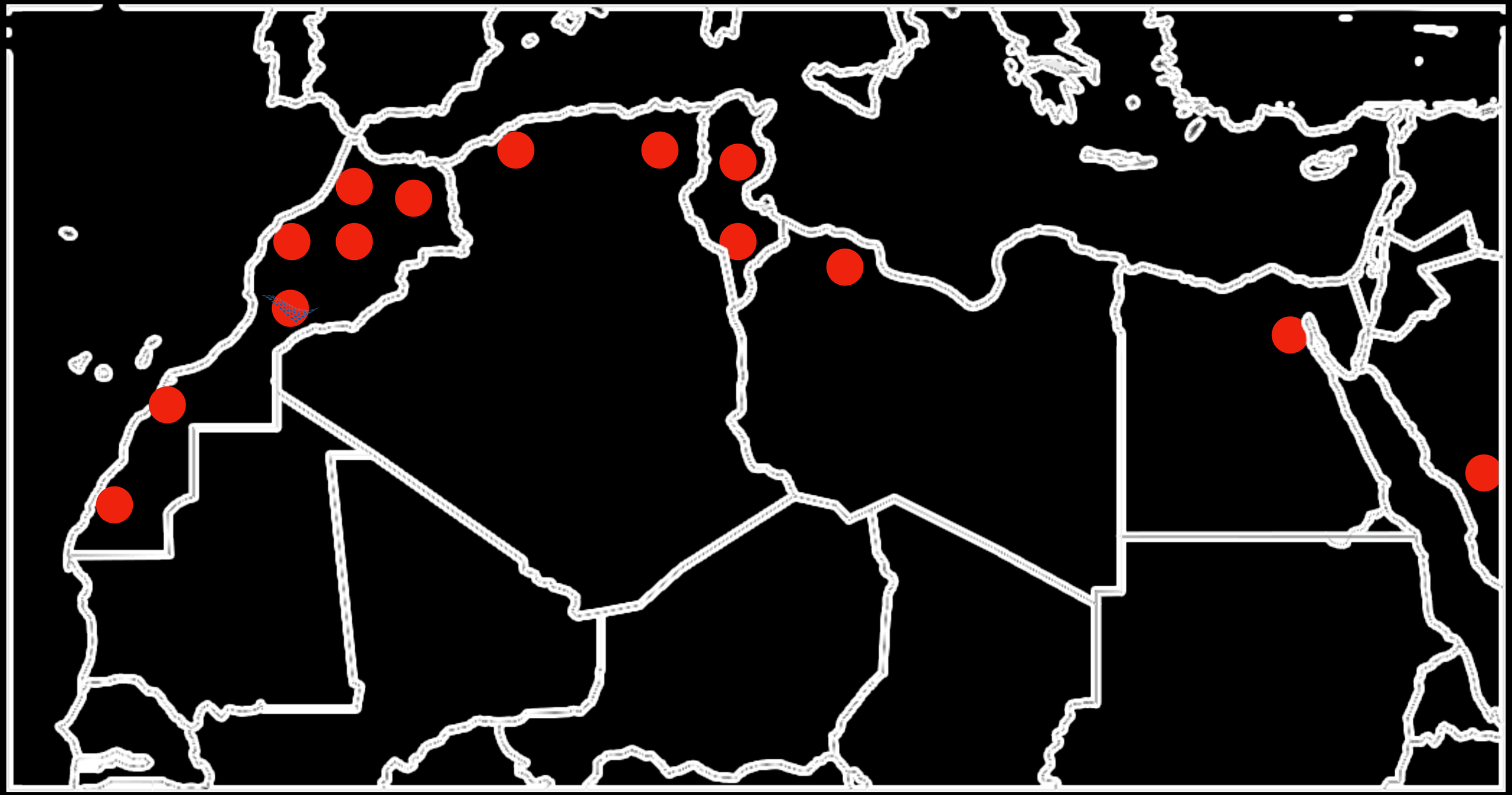


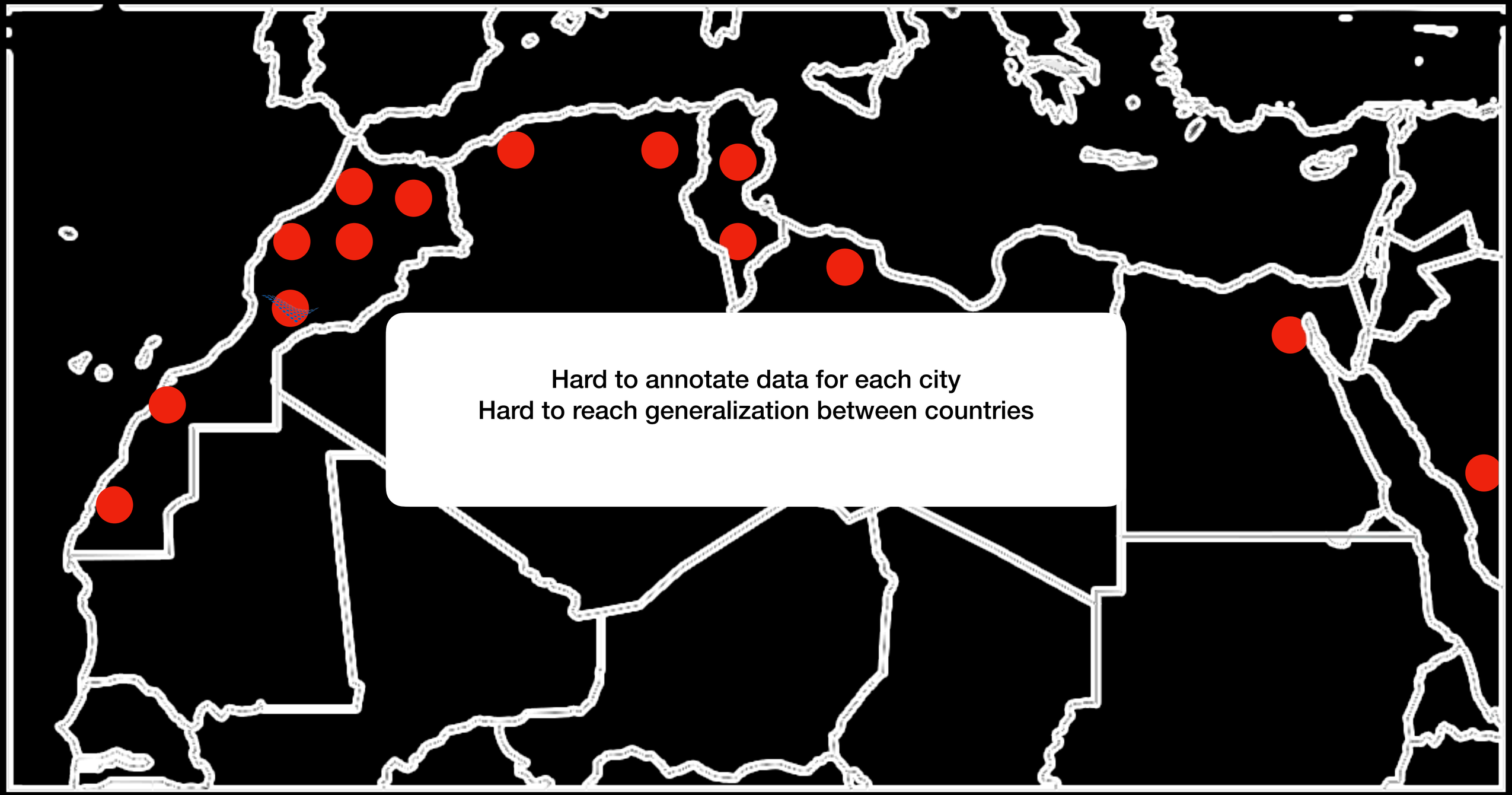


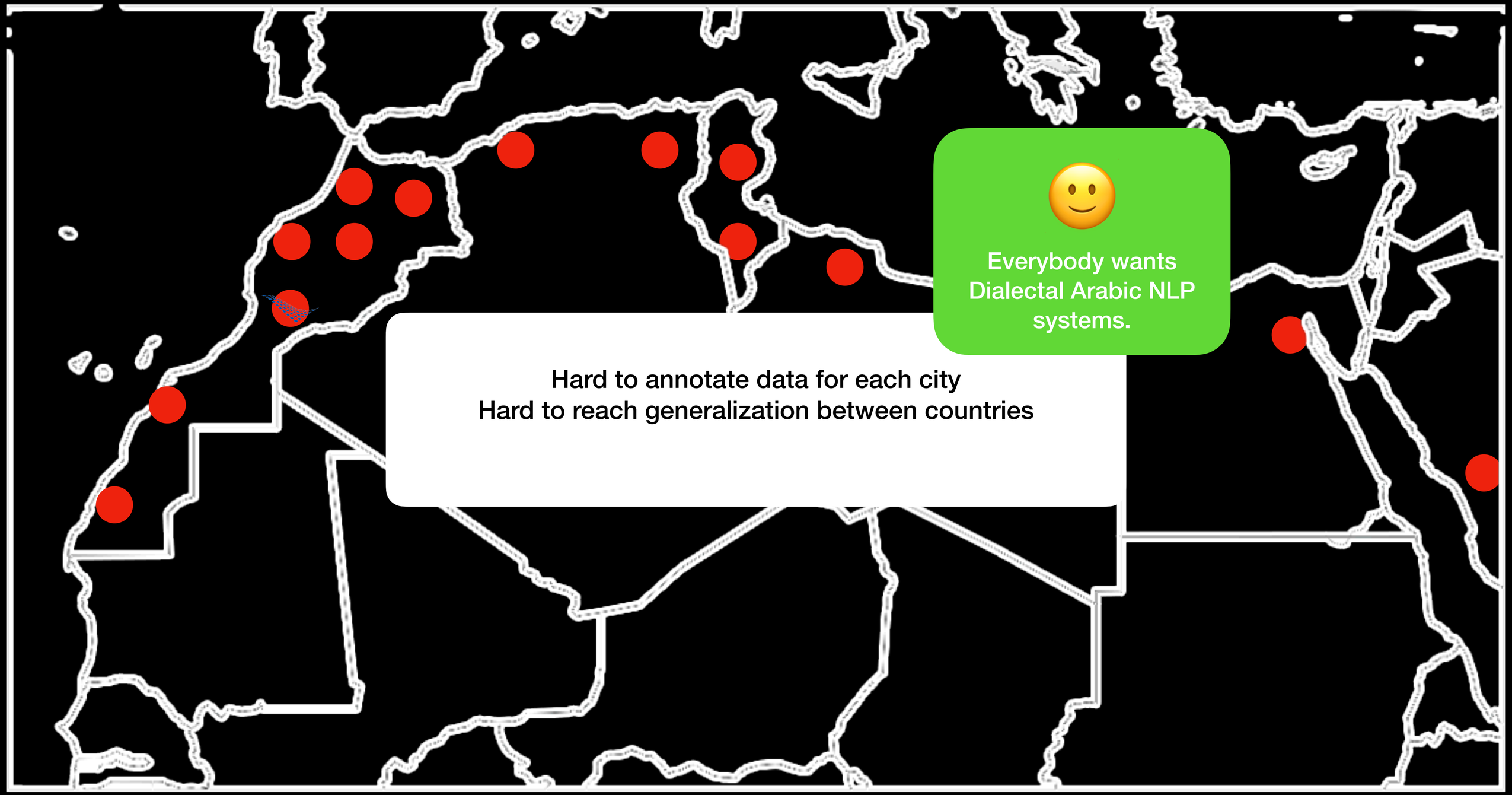


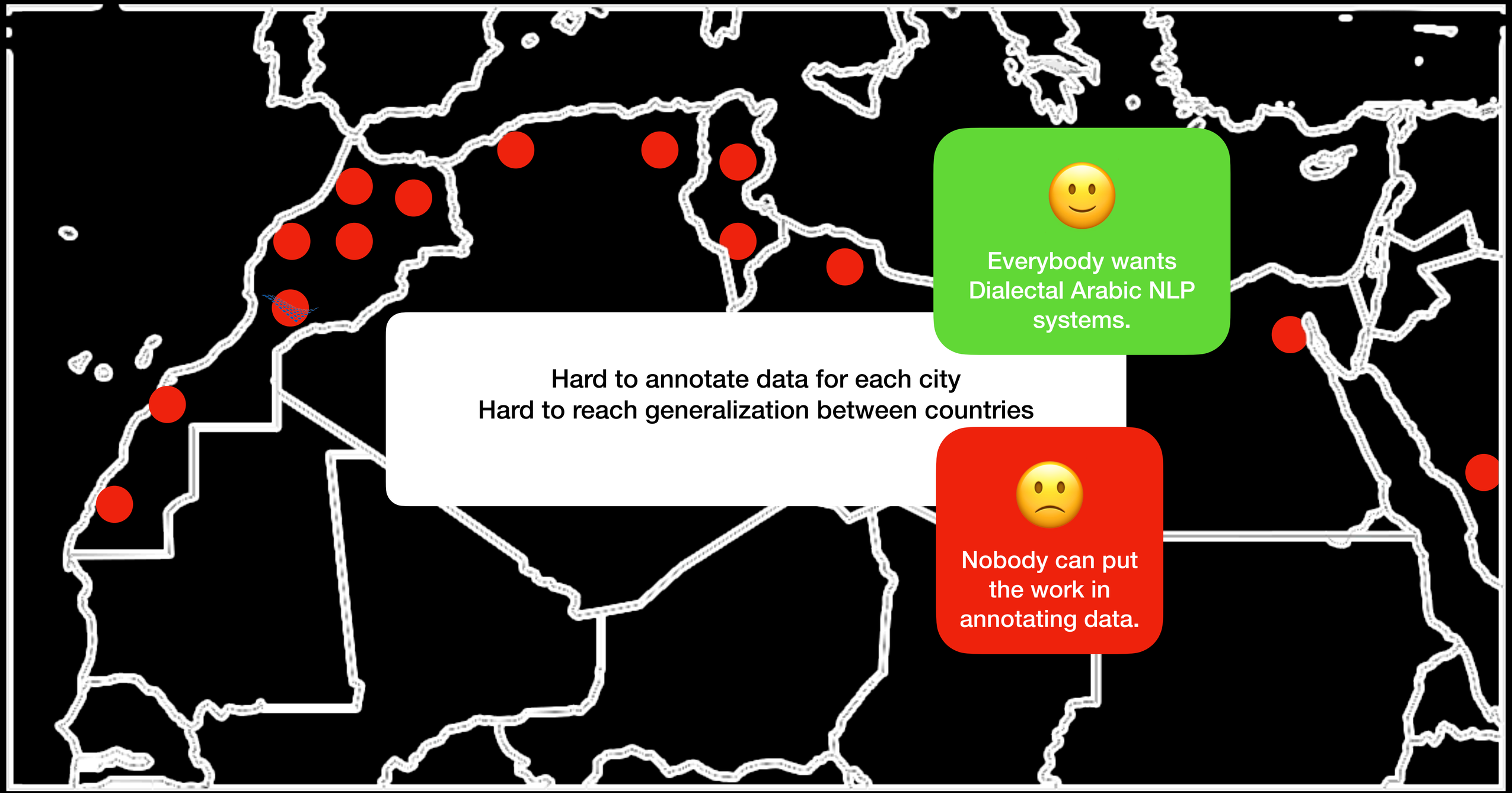


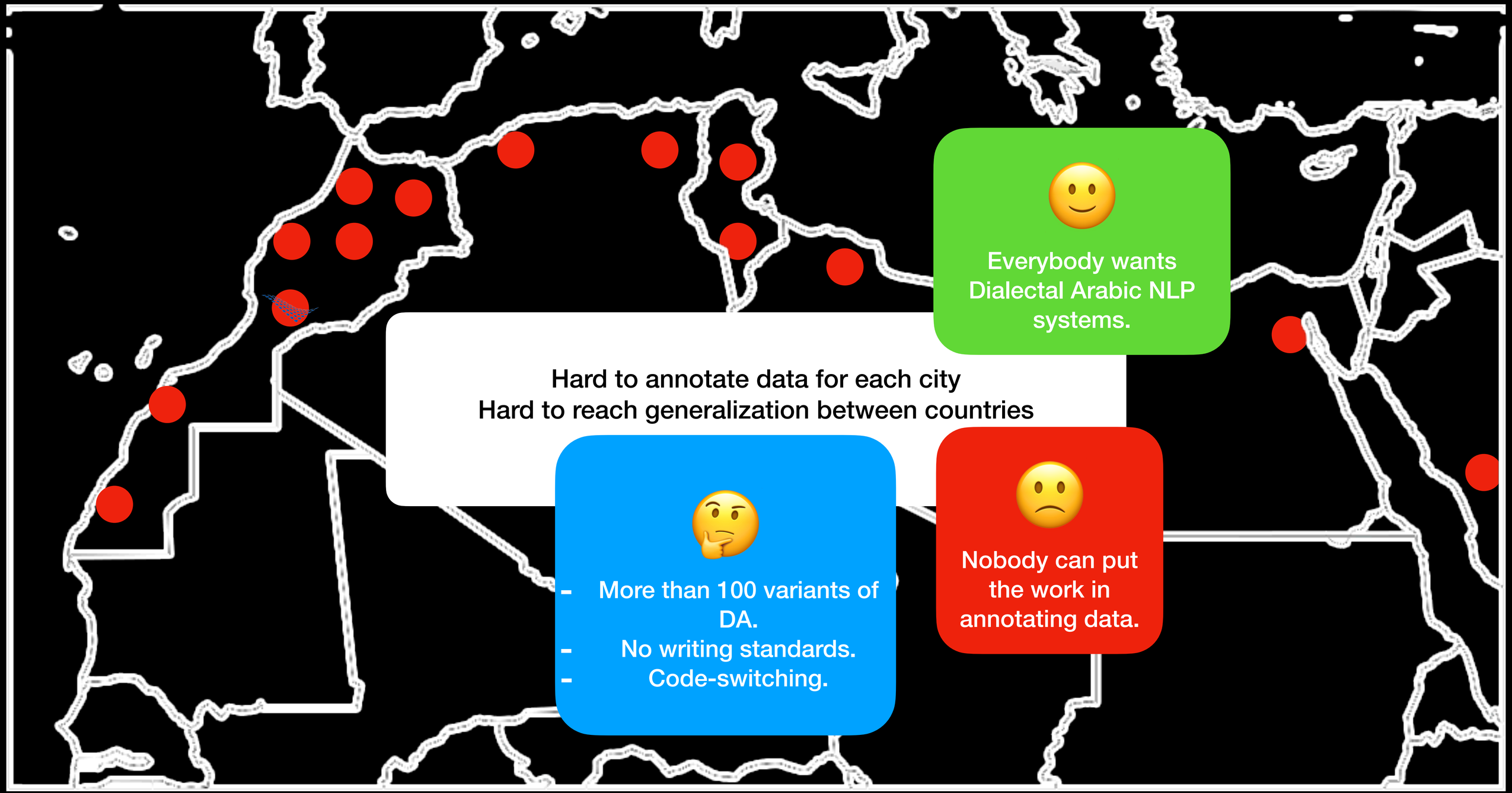


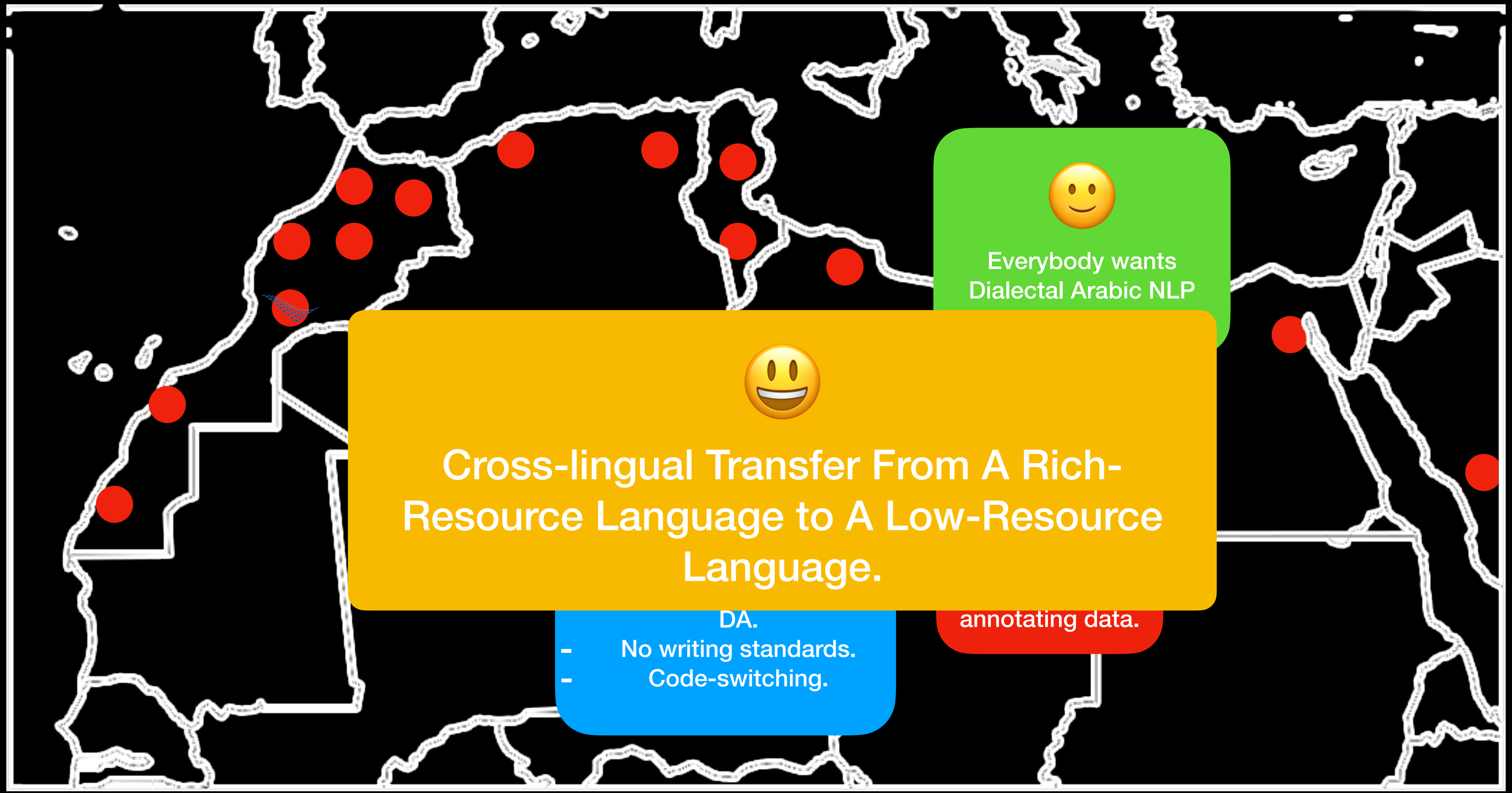



Hard to annotate data for each city
Hard to reach generalization between countries

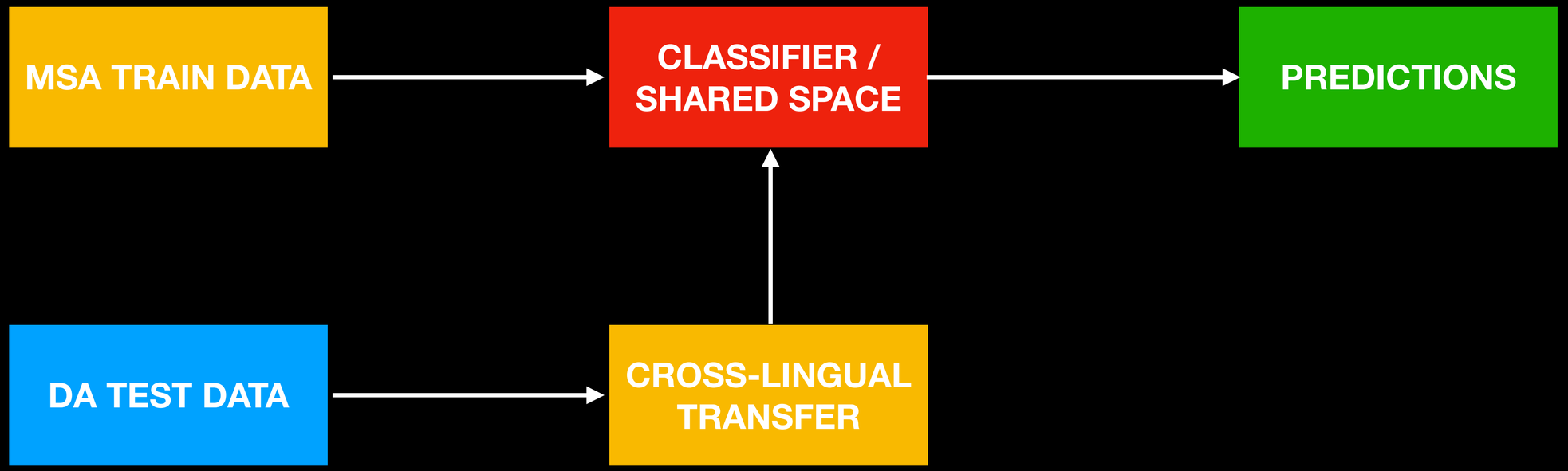

🙂
Everybody wants Dialectal Arabic NLP systems.
Hard to annotate data for each city
Hard to reach generalization between countries


Everybody wants Dialectal Arabic NLP systems.
Hard to annotate data for each city
Hard to reach generalization between countries
Nobody can put the work in annotating data.


🙂
Everybody wants Dialectal Arabic NLP systems.
Hard to annotate data for each city
Hard to reach generalization between countries
🤔
- More than 100 variants of DA.
- No writing standards.
- Code-switching.
😕
Nobody can put the work in annotating data.


Everybody wants Dialectal Arabic NLP
Cross-lingual Transfer From A Rich-Resource Language to A Low-Resource Language.
DA.
- No writing standards.
- Code-switching.
annotating data.

Cross-lingual from MSA to DA



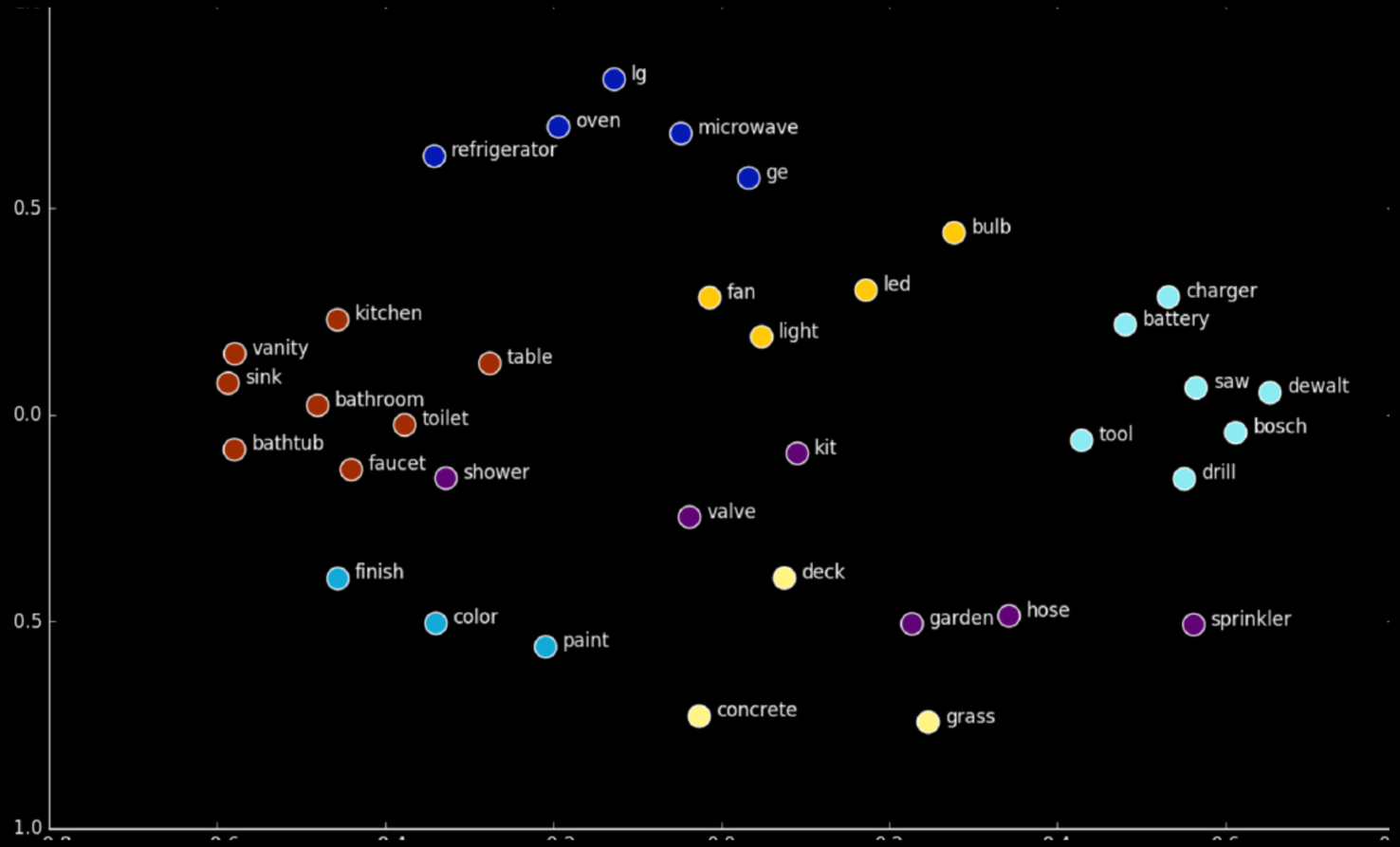
Motivations

- Modern Standard Arabic and Dialectal Arabic share similarities on multiple linguistic levels (orthography, morphology, phonology, etc).
- Availability of large Modern Standard Arabic pre-trained language models and annotated datasets.

Our research

- Word-to-word translation between Arabic dialects
- Unsupervised Sentence Classification for cross-domain & cross-dialects Arabic
- Unsupervised Sequence Labeling framework for Arabic dialects

Word-to-word translation between Arabic dialects



Word Embeddings

Word2vec (Mikolov, 2013)
FastText (Bojanowski, 2016)
BERT (Devlin, 2018)
XLM-RoBERTa (Conneau, 2019)
…

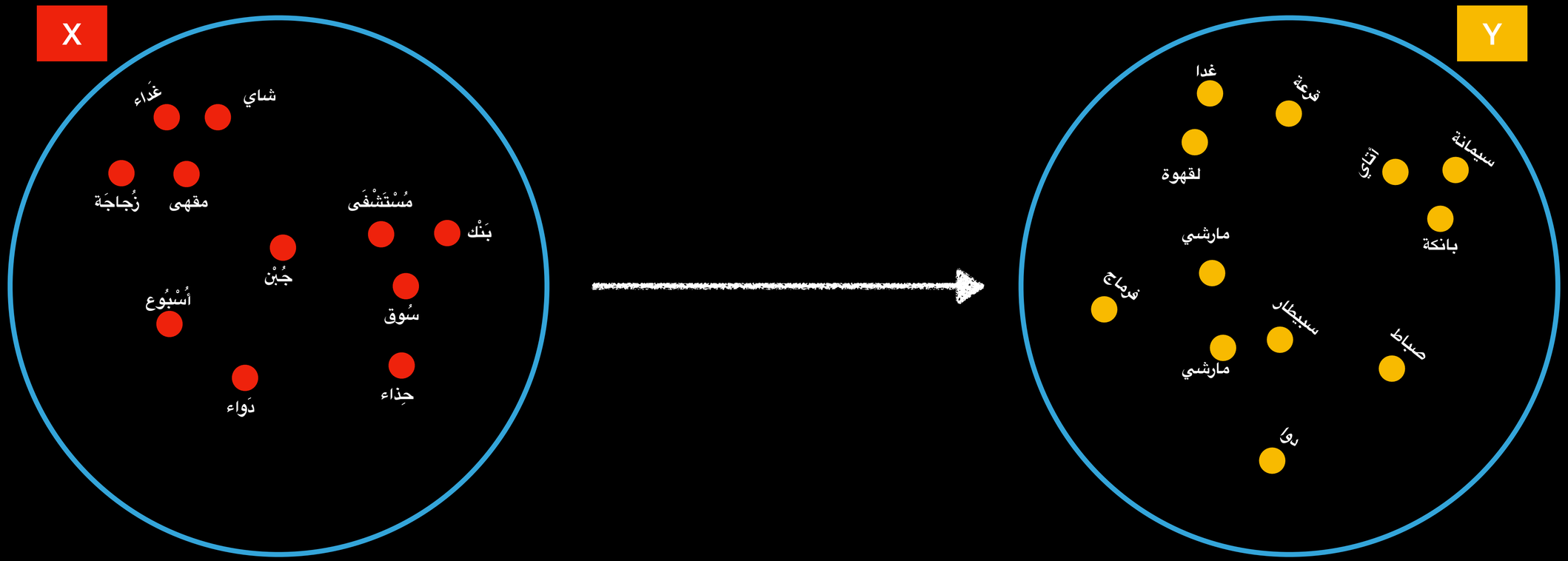
Multi-dialectal Arabic Word embedding

Approach

El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings. In Proceedings of the Canadian Conference on Artificial Intelligence.

Multi-dialectal Arabic Word embedding

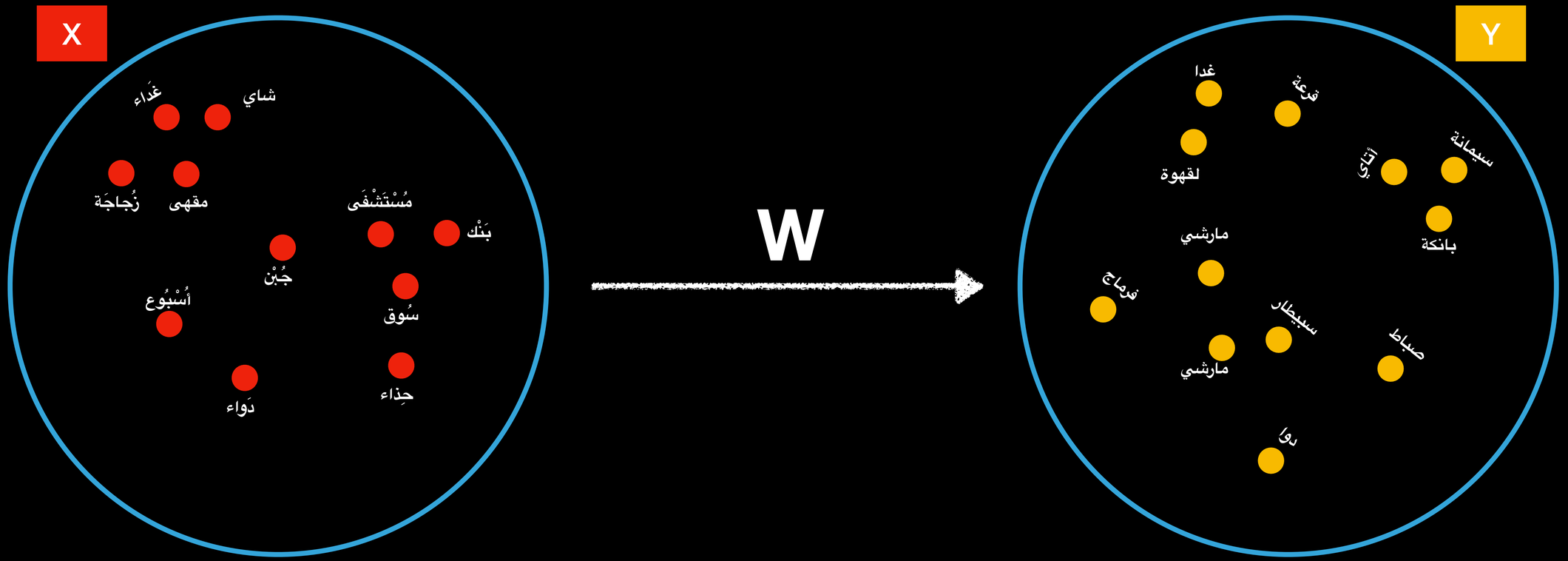
Approach



El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings. In Proceedings of the Canadian Conference on Artificial Intelligence.

Multi-dialectal Arabic Word embedding

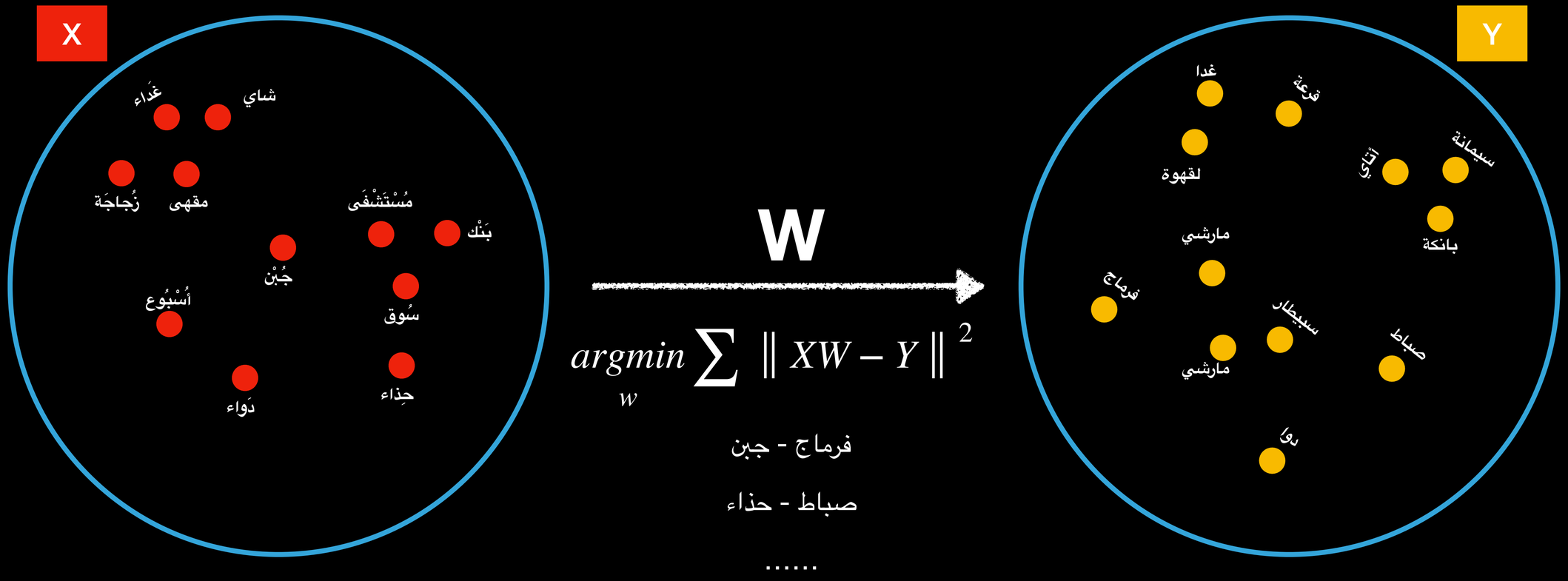
Approach



El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings. In Proceedings of the Canadian Conference on Artificial Intelligence.

Multi-dialectal Arabic Word embedding

Approach



El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings. In Proceedings of the Canadian Conference on Artificial Intelligence.

Multi-dialectal Arabic Word embedding

Approach

- Orthographic extension of the word embeddings.
- The encoding of the orthographic variations between Arabic dialects.

El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. On the Role of Orthographic Variations in Building Multidialectal Arabic Word Embeddings. In Proceedings of the Canadian Conference on Artificial Intelligence.

Extension of embeddings using orthographic features

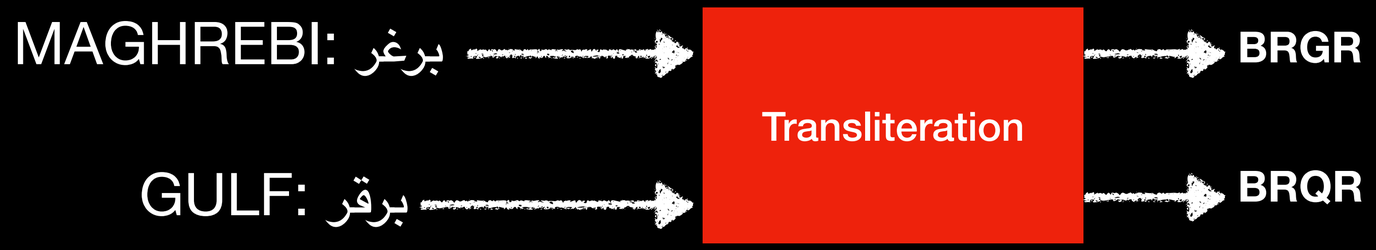
Approach

MAGHREBI: برغر

GULF: برقر

Extension of embeddings using orthographic features

Approach



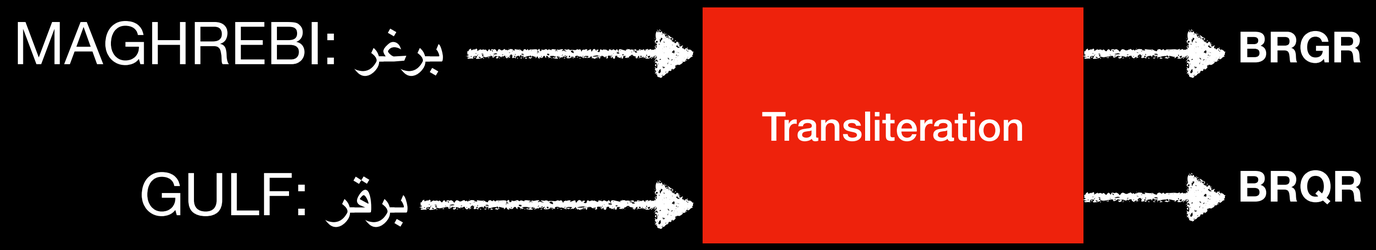
Vocabulary = {BRGR, BRQR}
Alphabet = {B, R, G, Q}
Embeddings dimensions: 100

Embedding vector

	d1	d2	d3	…	d100
BRGR	0.13	0.4	0.01	…	0.71
BRQR	0.4	0.8	0.2	…	0.19

Extension of embeddings using orthographic features

Approach



Vocabulary = {BRGR, BRQR}
Alphabet = {B, R, G, Q}
Embeddings dimensions: 100

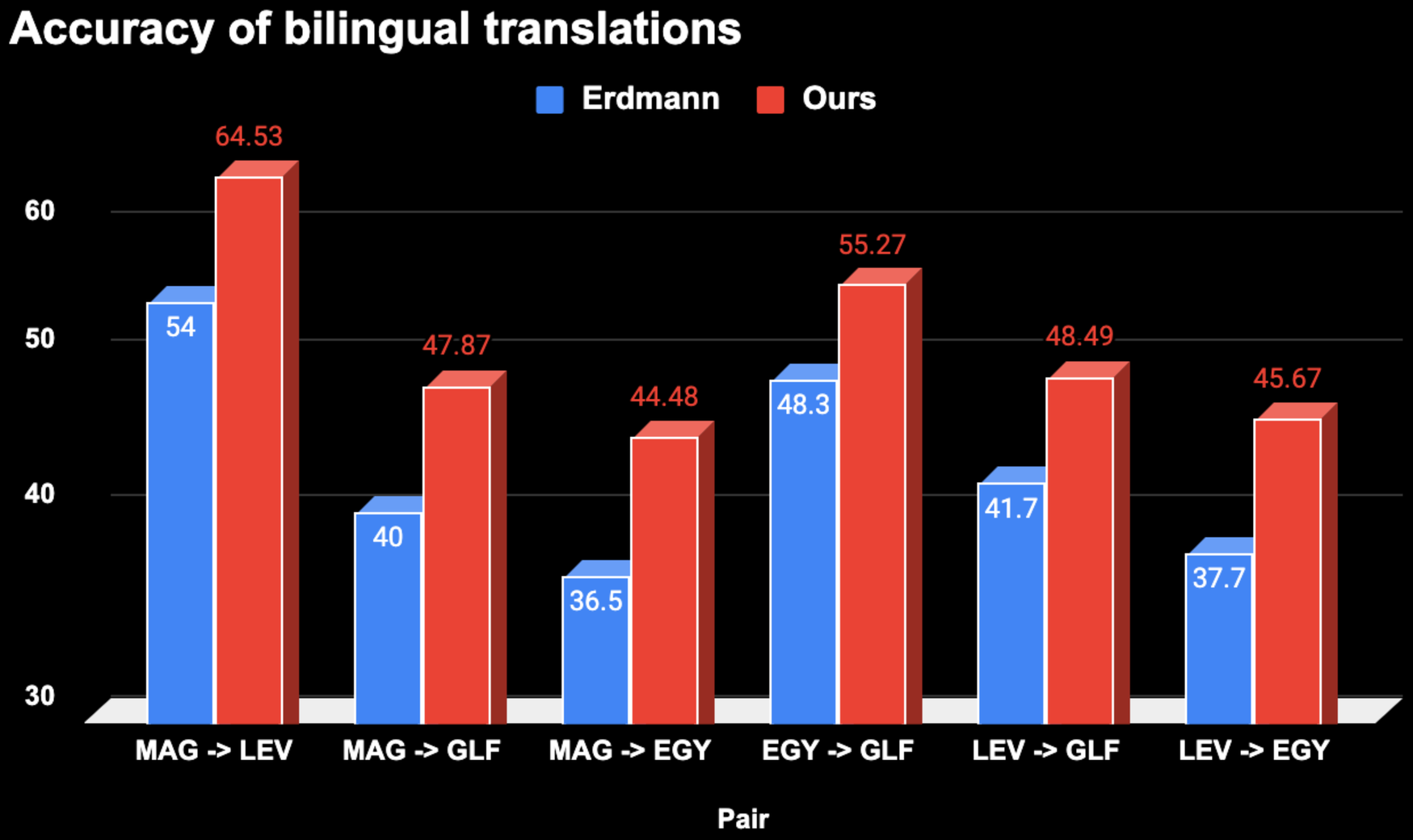
Embedding vector

Orthographic features
Count of letters

	d1	d2	d3	…	d100	B	R	G	Q
BRGR	0.13	0.4	0.01	…	0.71	1	2	1	0
BRQR	0.4	0.8	0.2	…	0.19	1	2	0	1

Multi-dialectal Arabic Word embedding

Results



Multi-dialectal Arabic Word embedding

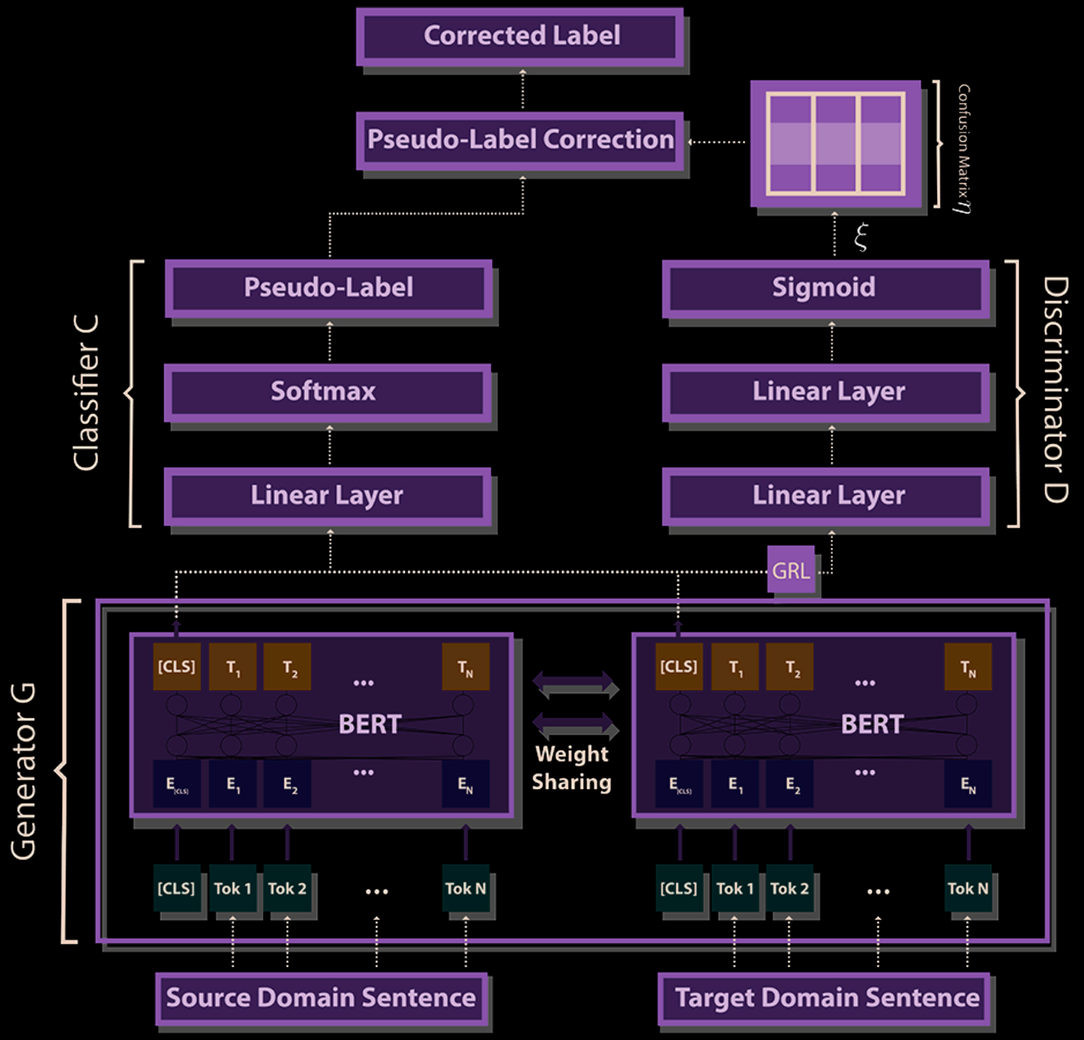
Limitations

- Performance is good in the word-to-word translation task. ✅
- Performance is low on downstream tasks such as sentence classification. ❌
- Hard and complex to extend to the sequence labeling tasks. ❌

Unsupervised Sentence classification for cross-domain cross-dialect Arabic dialects

Unsupervised sentence classification for DA

Approach



1. Self-training
2. Domain adaptation
    1. Alignment between the the sentences' embeddings of the source and target languages.
    2. Class-wise alignement.

El Mekki, A., El Mahdaouy, A., Berrada, I., and Khoumsi, A. 2021. Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 2824–2837).

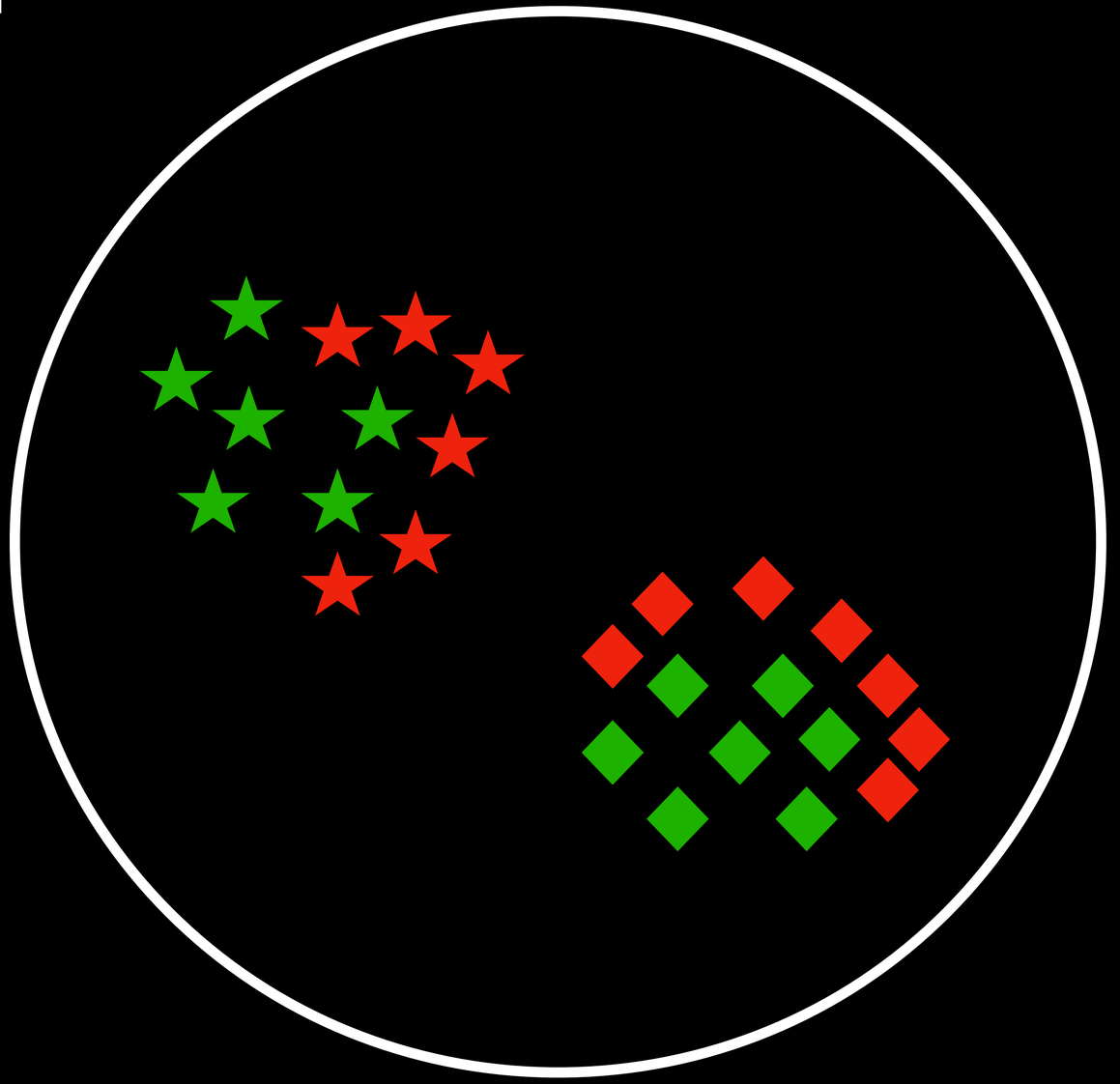
Domain Adaptation

Positive

Negative

Standard Arabic

Dialectal Arabic



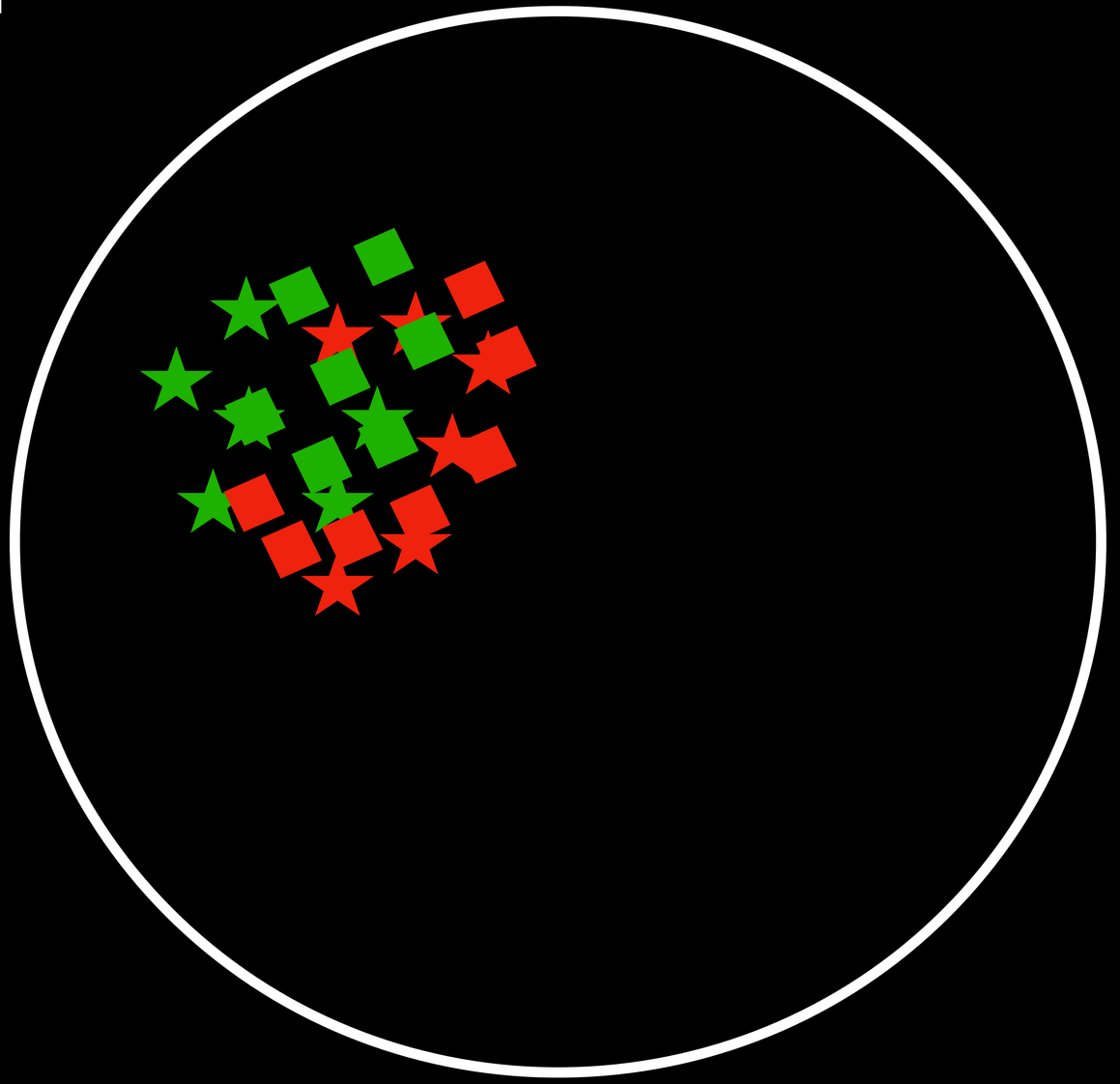
Domain Adaptation

Positive

Negative

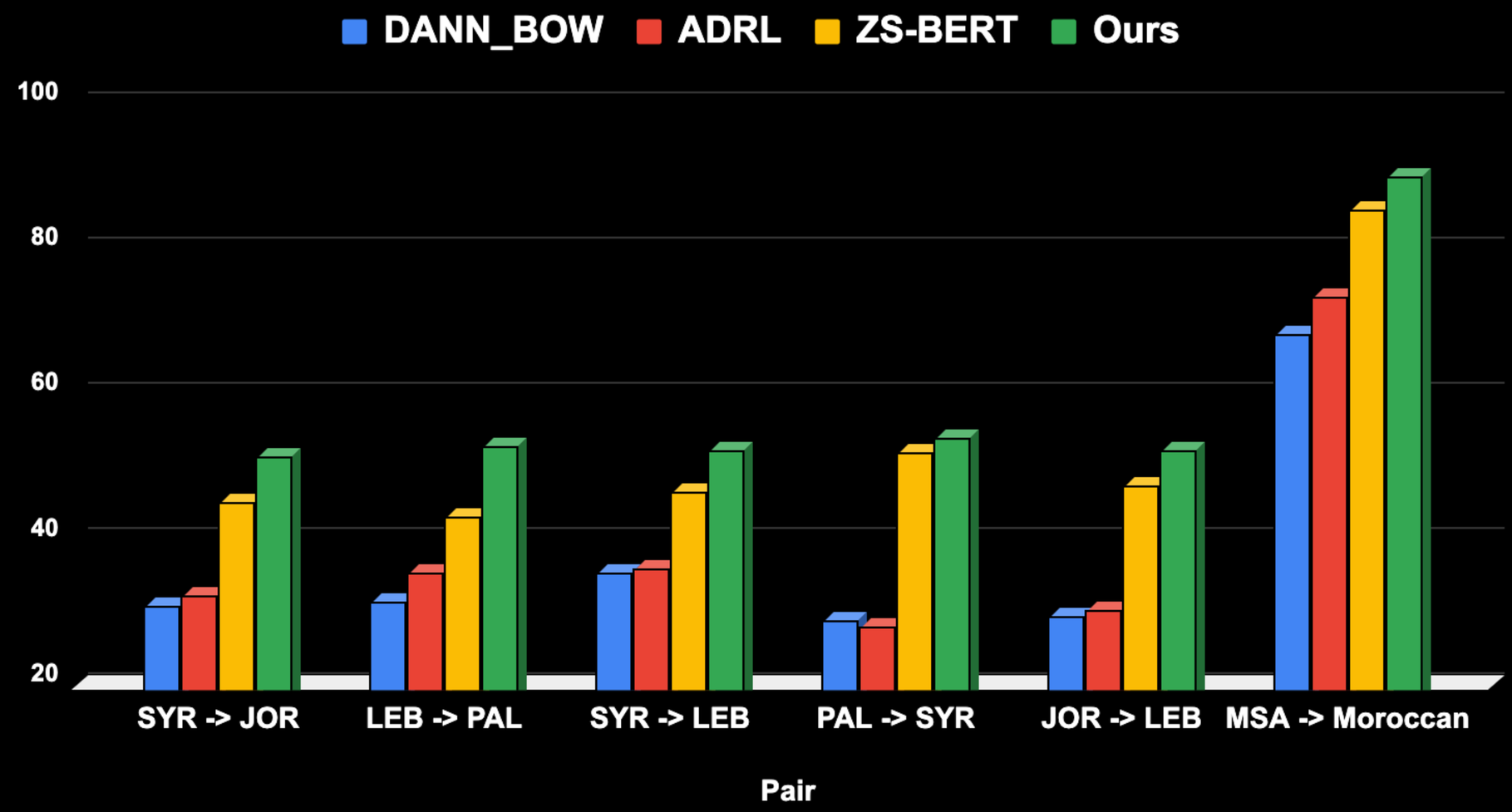
Standard Arabic

Dialectal Arabic



Unsupervised sentence classification for Dialectal Arabic

Results



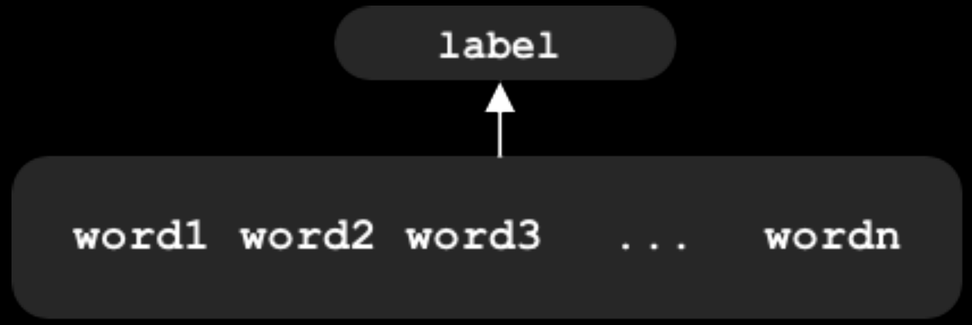
Unsupervised sentence classification for DA

Limitations

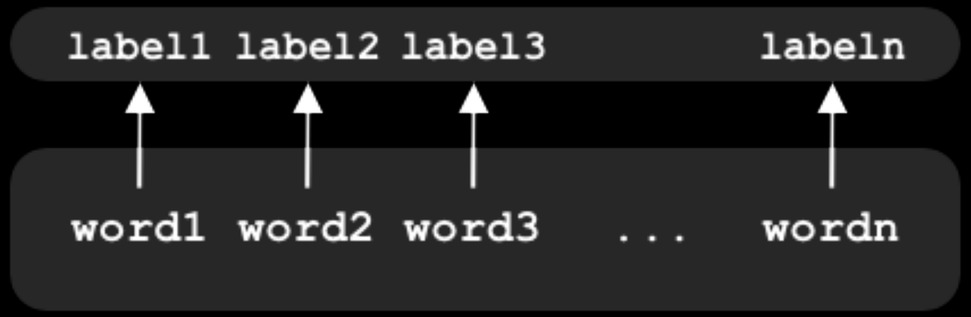
- Good performance on sentence classification tasks. ✅
- Hard and complex to extend to the sequence labeling tasks. ❌

AdaSL: An Unsupervised Domain Adaptation Framework For Arabic Multi-dialectal Sequence Labeling

Sentence classification

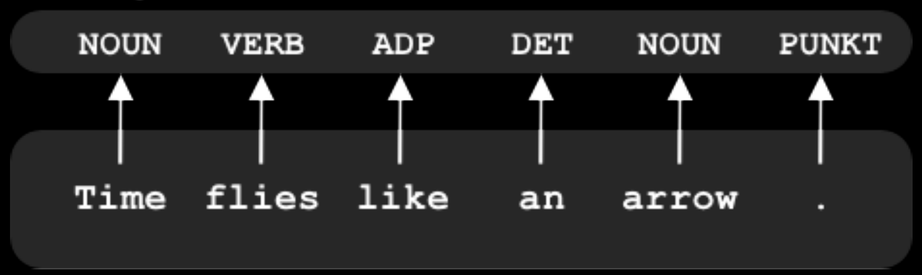


Sequence labeling

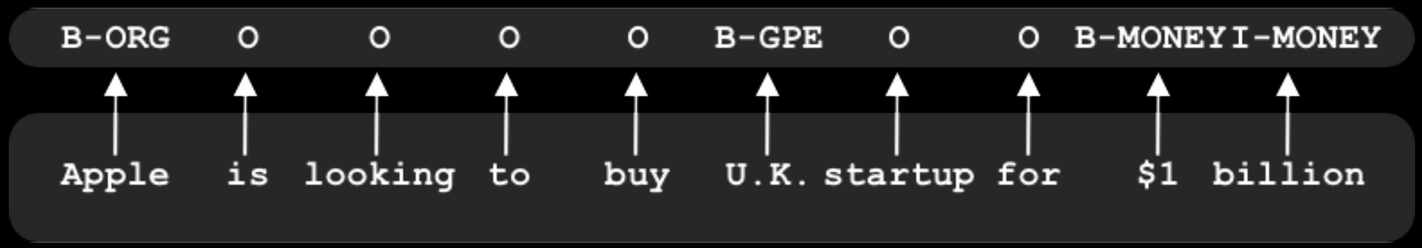


Examples of sequence labeling tasks:

Part-of-Speech (POS) tagging

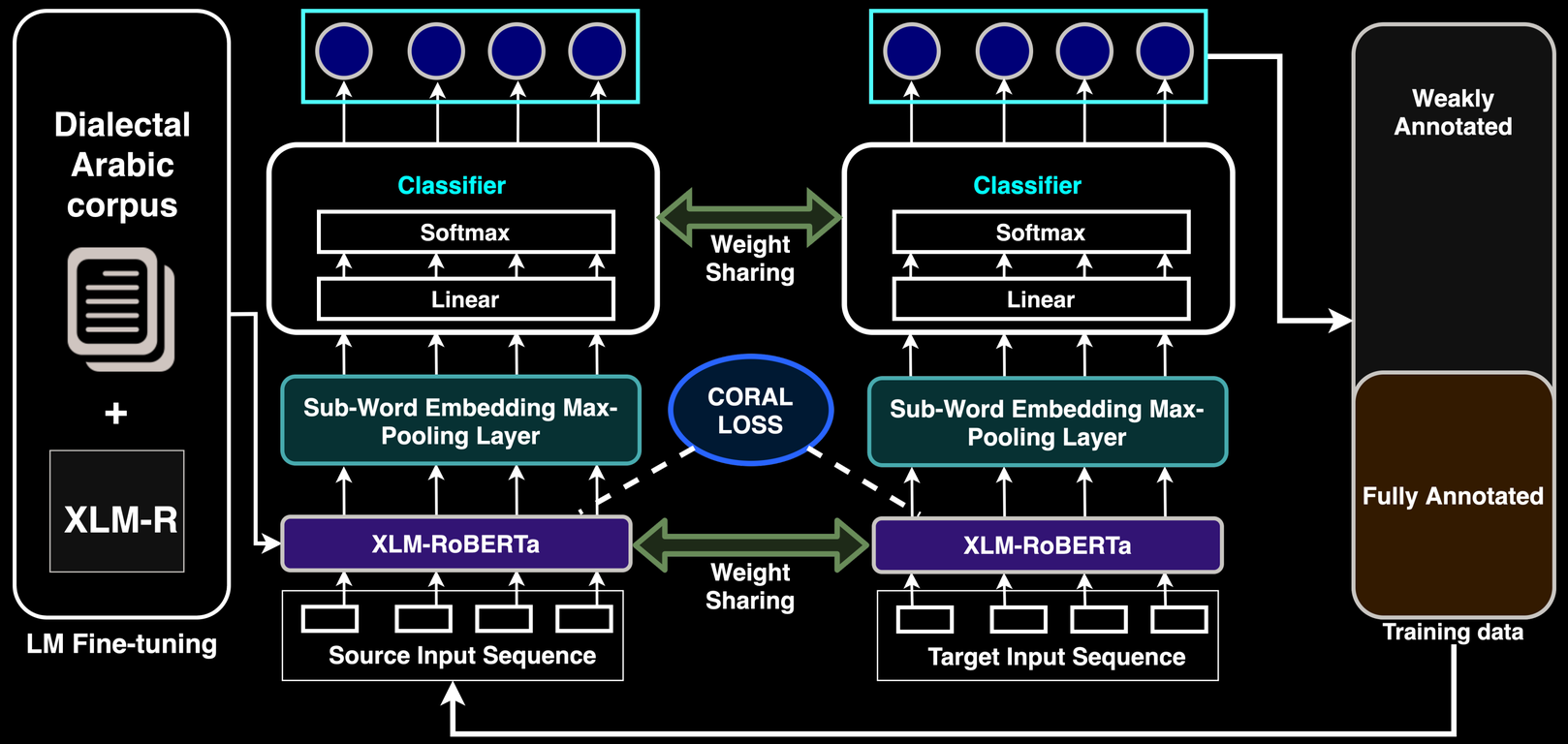


Named-entity recognition (NER)



Unsupervised Sequence Labeling for Dialectal Arabic

Framework



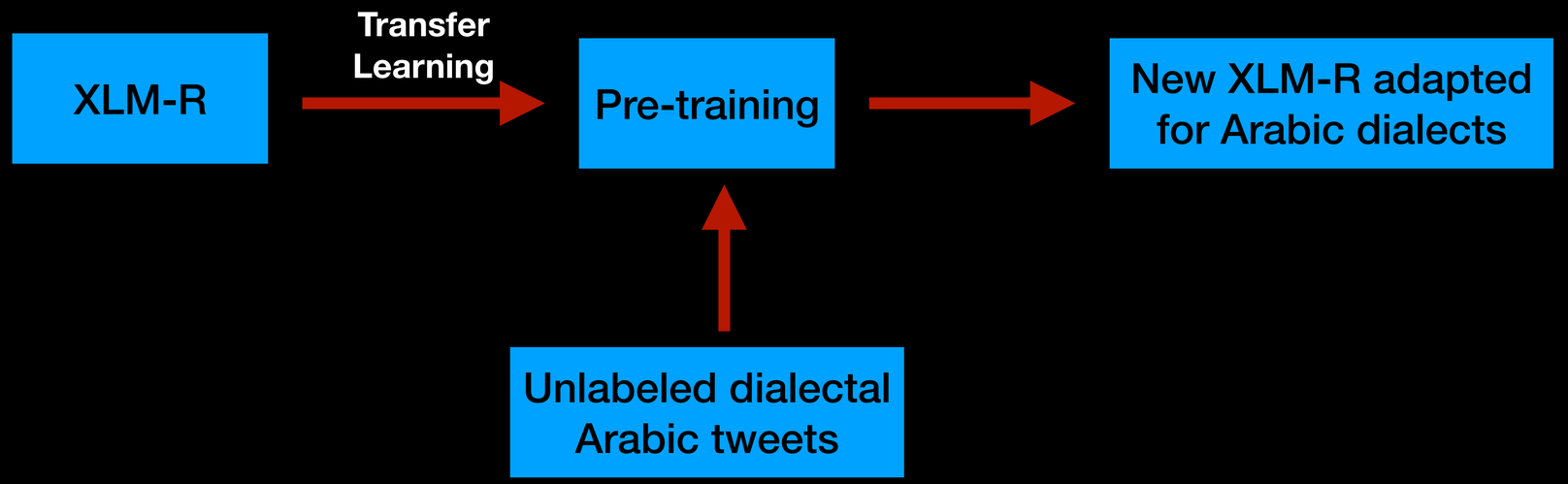
1) Language model domain adaptive fine-tuning
2) Sub-word embedding pooling
3) Unsupervised MSA-DA distributions alignment
4) Iterative self-training

El Mekki, A., El Mahdaouy, A., Berrada, I. & Khoumsi, A. AdaSL: An Unsupervised Domain Adaptation framework for Arabic multi-dialectal Sequence Labeling. Information Processing & Management. 59, 102964 (2022)

Unsupervised Sequence Labeling for Dialectal Arabic

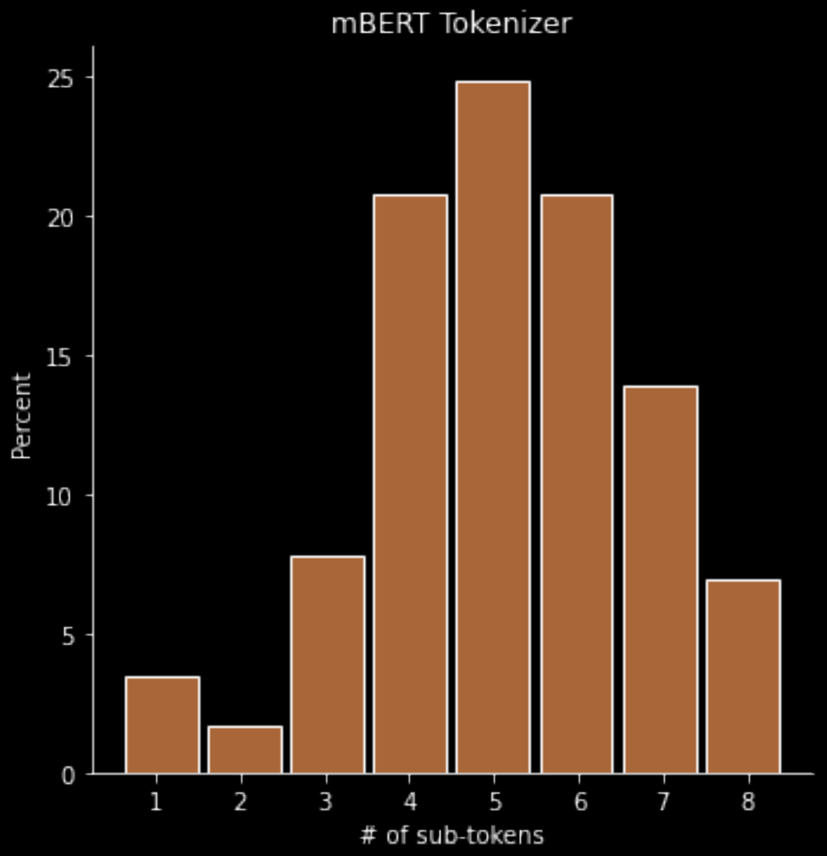
1) Domain adaptive fine-tuning (Masked-language modeling on the target dialect)

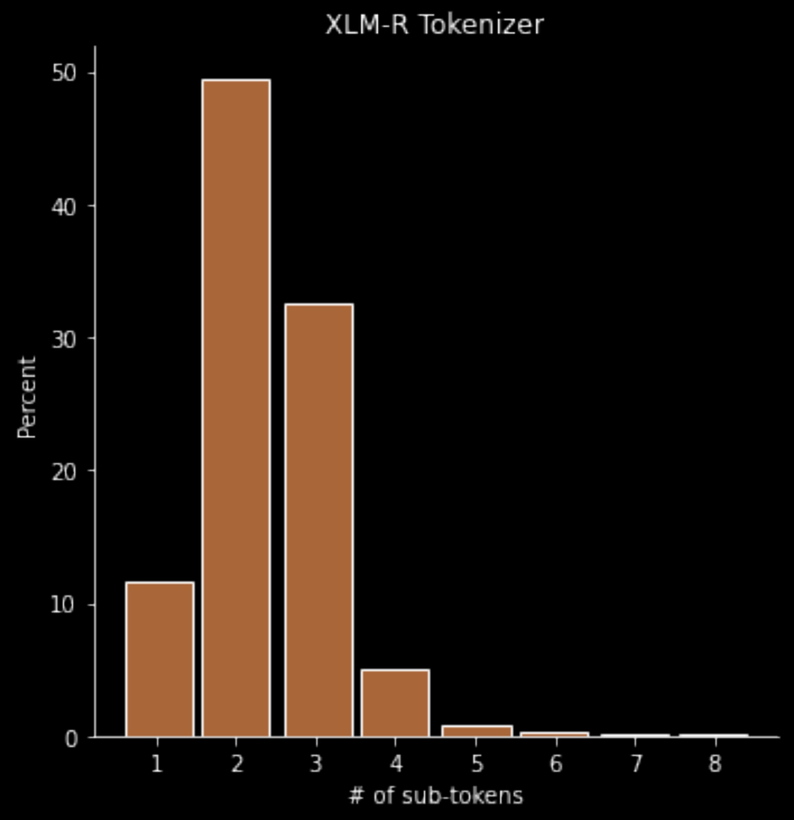
1) Collected 5 Million Dialectal Arabic Tweets.
2) Fine-tuning the multilingual language model on the collected tweets using masked-language modeling objective.



Unsupervised Sequence Labeling for Dialectal Arabic

2) Sub-word embedding pooling





1) Multilingual Language models tokenize Arabic words into multiple subwords.
2) Arabic is morphologically complex.
3) State-of-the-art sequence labeling methods use the 1st sub-token to label the word.

Tokenizers

- MSA tokenizer

```
tokenizer.tokenize("الجو جميل اليوم")
['الجو', '_جميل', '_اليوم']
```

- Dialectal Arabic tokenizer

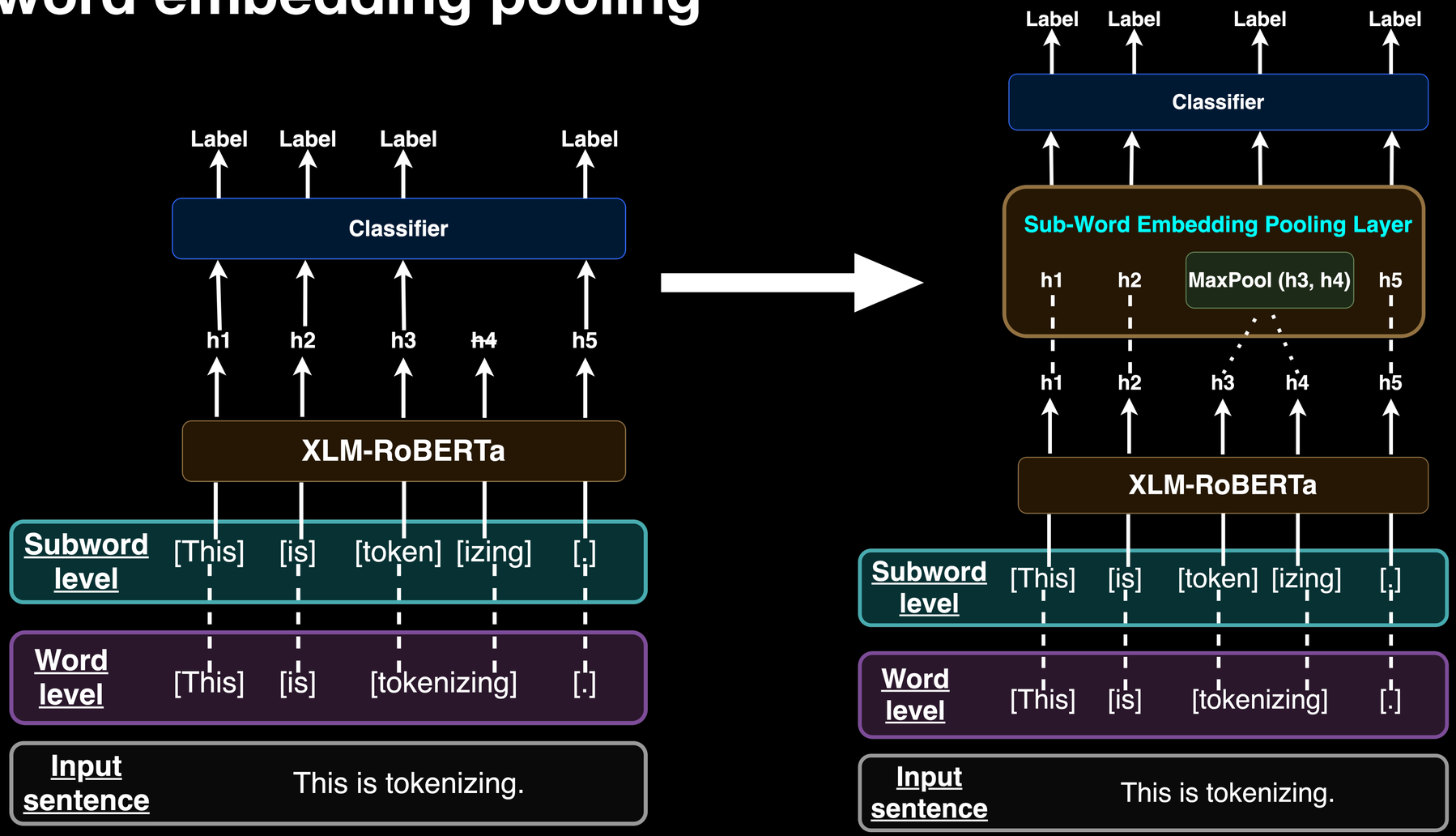
```
tokenizer.tokenize("الجو زوين ليوما")
['الجو', '_زو', 'ين', '_', 'ليو', '_ما']
```

- GPT-3 Tokenizer on English

- GPT-3 Tokenizer on Arabic

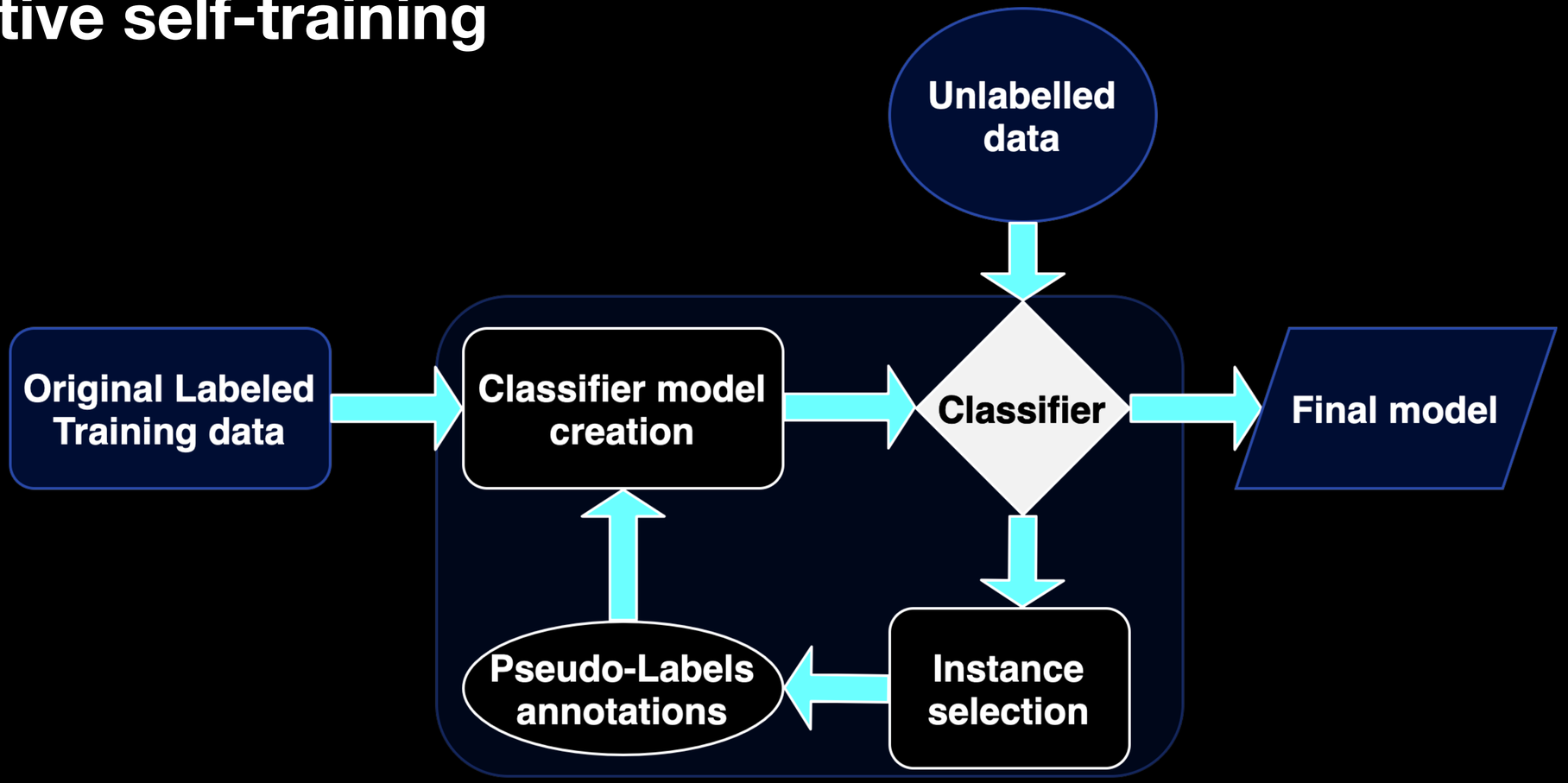
Unsupervised Sequence Labeling for Dialectal Arabic

2) Sub-word embedding pooling



Unsupervised Sequence Labeling for Dialectal Arabic

4) Iterative self-training



Unsupervised Sequence Labeling for Dialectal Arabic

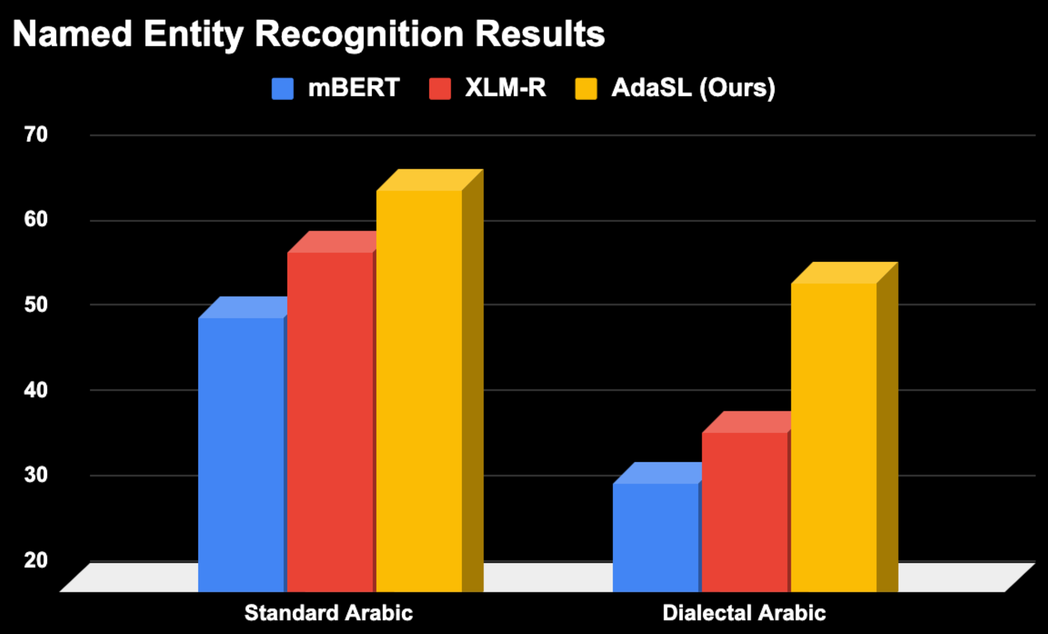
Evaluation scenarios

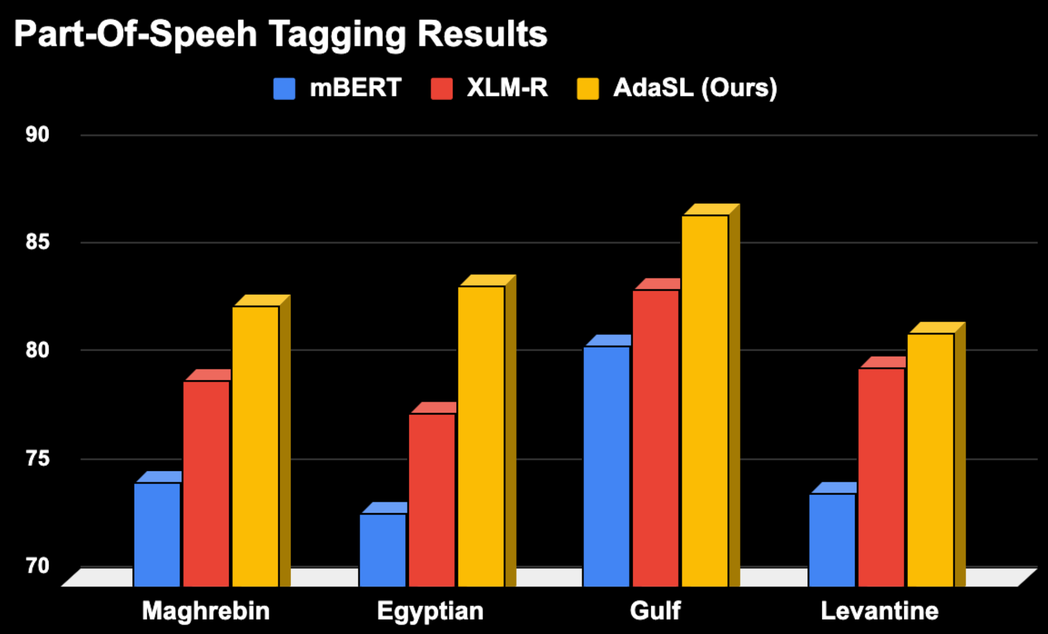
Train Sequence Labelling Model on MSA Data Only → Predict on Dialectal Data with Zero-labeled Data

- **Named Entity recognition task:**
  - Training dataset: MSA
  - Test dataset: MSA (Twitter), DA (Twitter)
  - Evaluation metric: Macro-F1 score
- **Part-of-speech tagging task:**
  - Training dataset: MSA
  - Test dataset: Maghreb (MAG), Egyptian (EGY), Leventine (LEV) and Gulf (GLF)
  - Evaluation metric: Accuracy

Unsupervised Sequence Labeling for Dialectal Arabic

Results





جامعة ORG محمد السادس PERS متعددة التخصصات التقنية مؤسسة دولية للتعليم العالي متخصصة في العلوم، والتكنولوجيا، والعلوم الاجتماعية، والأعمال. المقر الرئيسي ديالها كاين فمدينة ابن PERS جرير GPE لي جات حدا مدينة

مراكش GPE بالمغرب، GPE عندها فروع فالرباط GPE وفالعيون. صاوباتها مؤسسة المكتب الشريف للفوسفاط ORG OCP، ORG ودشنها الملك محمد السادس PERS نهار 12 يناير 2017. DATE

Conclusions

- A good performance on Arabic dialects NLP systems can be achieved with zero labeled data.
- Semi-supervised approaches can lead to better results (few-shot learning).
- More effort should be performed to make labeled data for underrepresented languages available (crowdsourcing, more use of local languages on the web rather than using 2nd languages).

Resources

- Stanford NLP course:
  - https://www.youtube.com/playlist?list=PLoROMvodv4rOSH4v6133s9LFPRHjEmbmJ
- Jay Alammar Blog:
  - https://jalammar.github.io/
- Coding:
  - Python
  - Pytorch
  - Huggingface

THANK YOU FOR YOUR ATTENTION!

Any questions?

abdellah.elmekki@um6p.ma