CUNI Submission to MRL 2023

Jindřich Helcl, Jindřich Libovický

7 December 2023, Singapore

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Tasks

Named-Entity Recognition

PERS ORG

Founded by Václav Havel, the Civic Forum was a key player in the Velvet Revolution in LOC Czechoslovakia in November 1989, advocating for democracy and contributing to the downfall of communist rule. DATE

Question Answering

Context: *[see the above example]*
Question: What was the purpose of the Civic Forum?
Answer: *[see the underlined text above]*

Shared Task Experimental Setting

Models developed on (an extension to) the XTREME-UP dataset:

- Focus on general utility tasks in under-resourced languages
  - (besides QA and NER also e.g. MT, ASR, or OCR)

- Evaluation in (very) low resource languages, some without any task-specific training data.
  - Igbo, Indonesian (QA only), Alemannic, Turkish, Uzbek (QA only), and Yoruba.
  - Azerbaijani added as a surprise language

  no QA data

  neither NER nor QA data

- Very small (and out-of-domain) evaluation dataset!

Common Approach

Translate test

1. Translate everything into English.

2. Use task-specific model (in English).

3. Translate back into the original language.

Common Approach

Translate test

1. Translate everything into English. 🤗 NLLB-3.3B multilingual MT

2. Use task-specific model (in English).

3. Translate back into the original language.

Common Approach

Translate test

1. Translate everything into English. 🤗 NLLB-3.3B multilingual MT

2. Use task-specific model (in English). 🤗 RoBERTa-large

3. Translate back into the original language.

Common Approach

Translate test

1. Translate everything into English. 🤗 NLLB-3.3B multilingual MT

2. Use task-specific model (in English). 🤗 RoBERTa-large

3. ~~Translate back~~ Project into the original language. 🤗 NLLB-3.3B
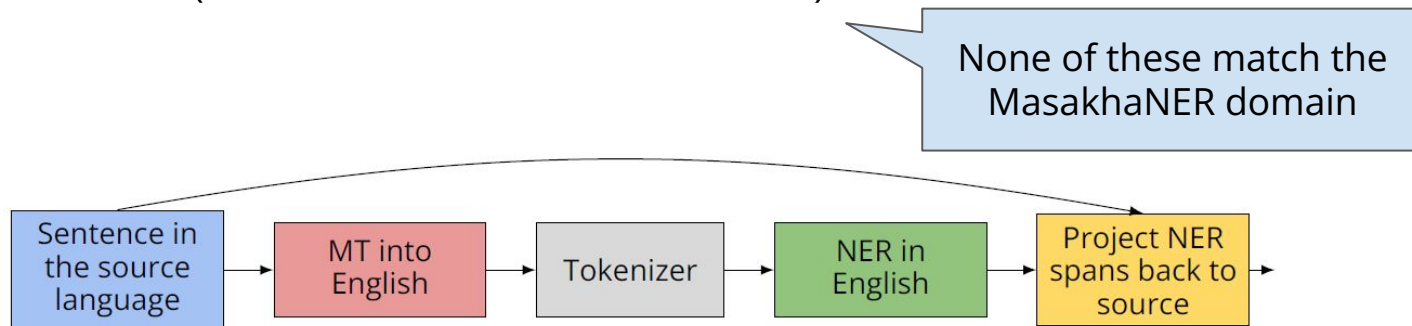
   ... finetuned for span projection 🤔

Translate-test with span projection

- Target sentence already known
- Use MT system finetuned for preserving <tags> in the sequence.
- Score every possible span (...subject to min/max length constraints).

```
Founded by Václav Havel , the …
Founded by Václav Havel , the …
Founded by Václav Havel , the …
Founded by Václav Havel , the …
Founded by Václav Havel , the …
```

Named-Entity Recognition

- Two RoBERTa-based models (tNER) 🤗, finetuned on:
  a. Ontonotes5 (collection of news, television and radio transcripts, web pages)
  b. CoNLL 2003 (international news from 1996-1997)

None of these match the MasakhaNER domain

| Sentence in the source language | → | MT into English | → | Tokenizer | → | NER in English | → | Project NER spans back to source | → |

- Finetuning experiments
  - Finetune tNER models on MasakhaNER data translated to English
  - Perhaps helps with double domain mismatch
    - MasakhaNER vs Ontonotes/CoNLL
   - Automatically translated data

Named-Entity Recognition - Results

- **MasakhaNER 1**
  - 63.3 average F1 score
  - Better than zero-shot transfer (32.9 avg F1)
  - Worse than supervised baseline (84.4 avg F1)
- **MasakhaNER 2**
  - Similar story, plus, for half of the added languages, our model performs better than zero-shot transfer from languages in MasakhaNER 1
  - In-domain finetuning helps for both versions of the benchmark
- **Shared task data**
  - Missing annotations in Yoruba dev/test sets
  - Domain mismatch - finetuning does not help

	als	aze	tur	yor	AVG
CoNLL 2003	38.4	54.6	48.0	4.4	47.0
Ontonotes5	40.3	62.8	54.7	3.1	52.4
+ finetuned	40.9	62.3	51.3	5.9	51.7

Question Answering

- RoBERTa finetuned for QA (deepset) 🤗
  - (always) finds an answer span in the context
- No-answer classification
  - some questions don't have answers in the context
  - RoBERTa-based classifier trained on translated XTREME-UP dataset
  - 93% accuracy on dev set
  - ... shared task validation/test data contained almost no "no-answer" examples
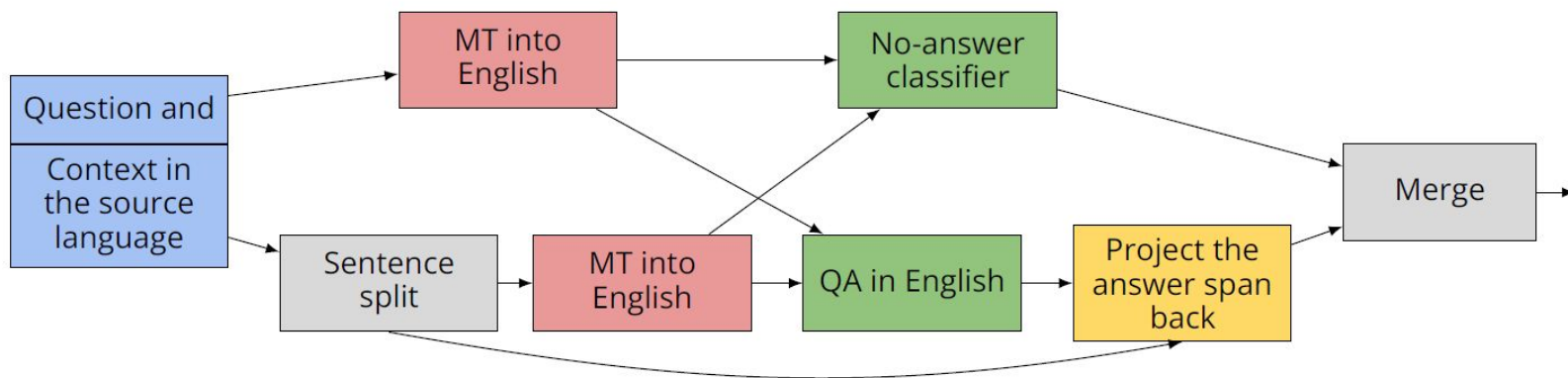
Question Answering

- ## In-domain fine-tuning
  - Similar to NER, we tune the RoBERTa model on translated XTREME-UP data
  - Use the span projection technique to project training answers into English

- ## Unlike XTREME-UP, shared task validation not extractive
  - We use Llama 2 to generate an answer
  - Nice results - submitted as contrastive experiment

Context: *Founded by Václav Havel, the Civic Forum was a key player in the Velvet Revolution in Czechoslovakia in November 1989, advocating for democracy and contributing to the downfall of communist rule.* *Question:* *What was the purpose of the Civic Forum?* *Short answer:*
The purpose of the Civic Forum, founded by Václav Havel, was to promote democracy and contribute to the downfall of communist rule in Czechoslovakia. [...]

Question Answering - Results

- XTREME-UP vs. shared task validation/test sets domain mismatch
  - In-domain finetuning not helpful
  - No-answer classification not needed

- Generating answers by Llama 2 scores higher

	als	aze	ind	tur	uzn	yor	AVG
Roberta-large	8.34	17.87	30.36	14.84	23.13	18.56	18.85
Finetuned	7.88	15.83	30.94	12.64	18.60	19.65	17.59
Llama 2	17.43	31.96	34.61	24.51	30.81	19.80	26.52

Conclusions

- Translate-test approach might prove itself as a strong baseline

- LLMs can be used for generative variants of the tasks

- Limited test data size, results not really reliable

`https://github.com/ufal/mrl2023-multilingual-ir-shared-task`