Relation Extraction

Aeirya Mohammadi

Relation

A predicate between e1, ..., en

RDF: subject property object

Relational Graphs

Knowledge Graph

Semantic Graph

Methods

Traditional: using pos tags, regex

Graph-based: Created with the help of above features

Neural: CNN, GCN, RNN (BiLSTM), RSN

Deep: Transformers (T5, BERT)

And now, LLMs

Transformers

Most prominent

BERT can be used for:

- Better input encodings for other methods
- Classification of relations and entity types

T5 for seq2seq tasks, as in REBEL (sota)

LLM

LLMs can do relation extraction out of the box.

Fine Tuning takes a lot of computation resource

Two other interesting options: Instruction tuning, In-context learning

PiVE*

Iteratively improve the result of LLM

Uses a verifier, a T5 model fine-tuned on RE datasets.

. Online and offline mode

* Prompting with Iterative Verification Improving Graph-based Generative Capability of LLMs

Datasets

SMiLER (Multilingual, has farsi)

REBEL

DocRED, REDFM, ..

GenWiki, WebNLG, CONLL04, NYT

Re-TACRED (Relation classification)

T-REx: Uses an old entity linking tool

Persian Datasets available

PARLEX (available in farsbase website)

And that’s it!

Did not find links for RePersian, ...

PARLEX

The first Persian dataset for relation extraction.

Bilingual dataset (direct **translation** of SemEval-2010-Task-8 dataset)

But has only sentence-level examples.

Size: 4 MB

Available in: Farsbase

SMiLER

By Samsung

Cons:

- Very sparse data for Farsi
- Mediocre F1 score for predicting both entities and labels right

(Next slide)

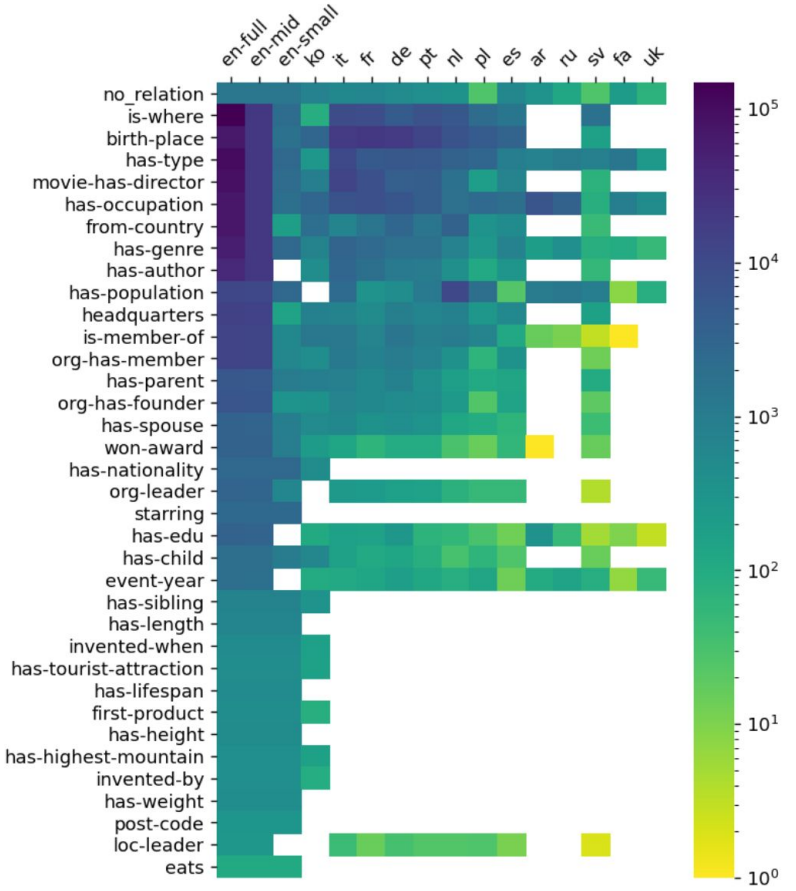

Figure 1: SMiLER: the number of sentences for each relation and for each language.

F1

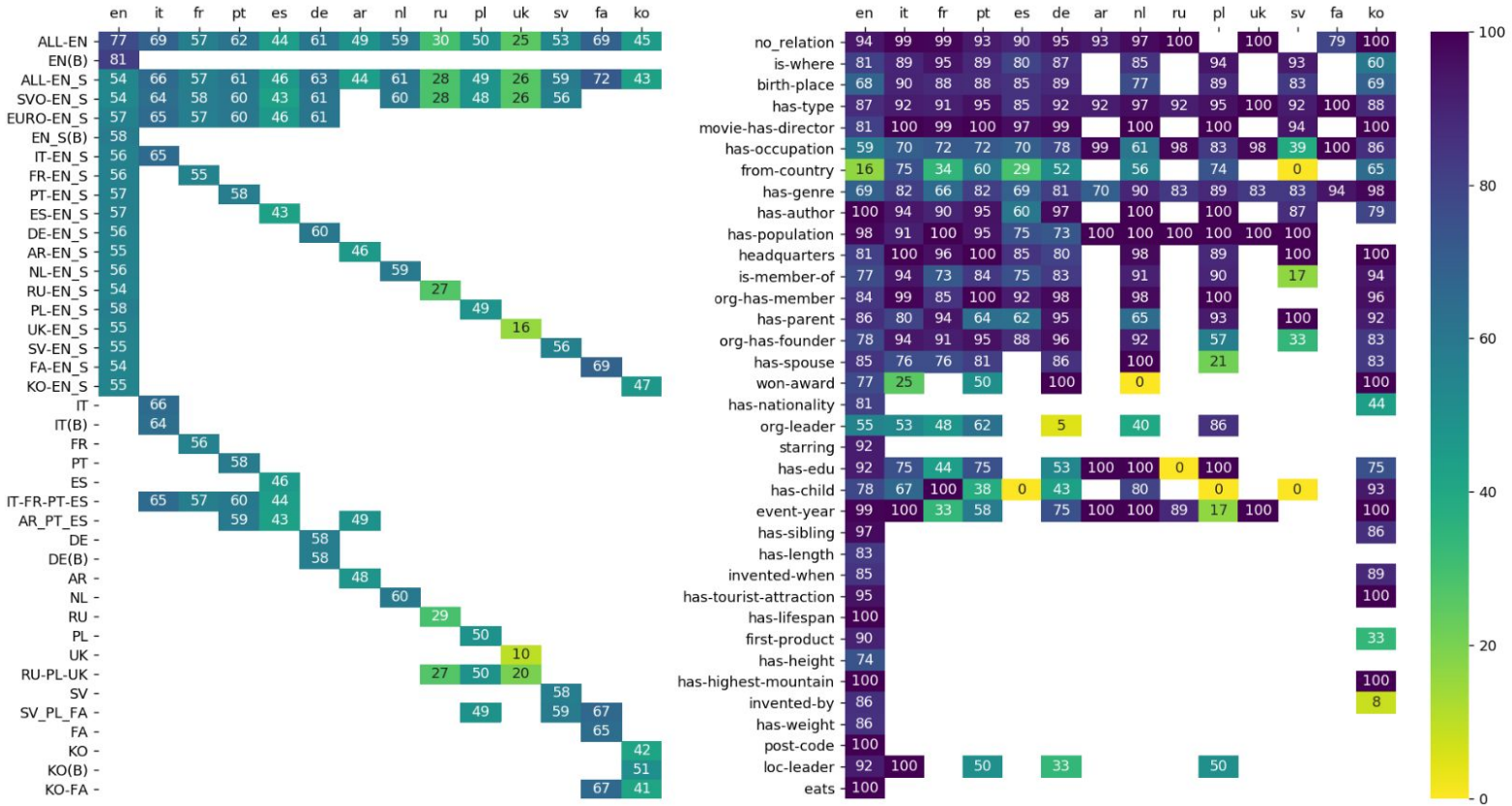

Figure 4: Left: F_1 (label and both entities correct) for all languages and models. Right: F_1 (label correct) for all languages and labels, averaged over the models available for each language.

Making a Persian dataset

1. Automatic ready extraction using tools like crocodile
2. Implement RePersian
3. Using GPT4 prompts to generate data
4. Translating existing datasets

Translation

We can leverage existing translation models.

PARS-BERT is better than pretrained mT5 for text summarization.

For translation task, we need to check if t5-fa competes with mBART.

Distant Supervision

Distant supervised paradigm is described as follows:

"If two entities participate in a relation, any sentence that contains those two entities might express that relation."

Datasets: Gold and Silver

Larger datasets are distantly supervised. And then:

(1.manually or 2.automatically) verify their validation and test dataset

A Distant Supervised Approach for Relation Extraction in Farsi Texts

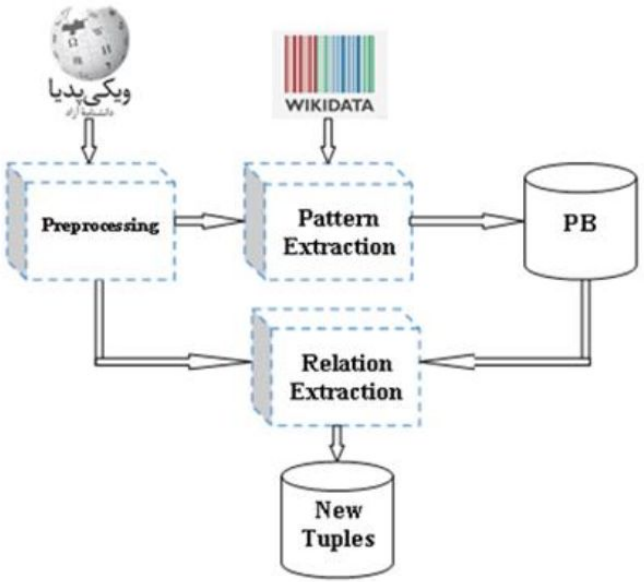

TABLE 4. PERFORMANCE FOR ALL APPROACHES

Research	Approach	Data	Language	Collection Method	Average Precision
Fadaei	Supervised	Wikipedia	Farsi	Manual	44%
Emami	Semi-supervised	Wikipedia Tabnak	Farsi	Manual	21%
Athena	Distant Supervised	Wikipedia	English	Automatic	84.1%
RePersian	Distant supervised	FarsBase	Farsi	Automatic	78.05%
Nasser	Distant supervised	Farsbase	Farsi	Automatic	76.81%
Proposed system	Distant supervised	Wikipedia	Farsi	Automatic	76.81%

NER

How to find entities? Wikipedia(url) -> DBpedia

Knowledge bases: Farsbase (Persian Freebase), Wikidata

Downfalls of RE methods:

Correferences, missing more than one or nested relations, computation, need for lots of data

My Project Proposal: PiVE for Persian

1. Verifier; google/flan-t5-small
2. LLM: mistralai/Mistral-7B-Instruct-v0.2
    a. Running LLM on colab and using an api to call it
3. Dataset:
    a. PARLEX for sentence-level FT
    b. GPT4 generated data
        i. Auto-verifier NLI module: xlm-roberta-large-xnli
        ii. Manually annotation of the wrong ones
    c. Gather more data with help of:
        i. Knowledge graph
        ii. Verifier (trained gradually)
        iii. Wikipedia, Wikidata, DBpedia

Version 0

- Finetune flanT5 on 20 GPT4 generated data
- Datasets:
  - PARLEX
  - 20 translated sentences from HIQ