

Named Entity Alignment in Czech-English Parallel Data

RASLAN 2024

Zuzana Nevěřilová, Hana Žižková

December 6, 2024

Introduction to NER and NEL

Named Entity Recognition (NER)

- Detects entities like *persons*, *locations*, *organizations*, and *miscellaneous terms*.

Named Entity Linking (NEL)

- Maps identified entities to structured knowledge bases (e.g., Wikidata).

Goal Use cross-language NER and named entity alignment to create NEL-annotated Czech dataset.

The Parallel Global Voices (PGV) Corpus

- PGV is a corpus of citizen media stories translated into **756 language pairs**.
- Focus on Czech-English pairs: **450 documents**.

NER in PGV (previous work)

- **English** BERT-large-NER model, achieving precision **0.77** and recall **English (0.45)**
- **Czech** Czert-B model, achieving precision **0.79** and recall **Czech (0.2)**

Still in the previous work

- Manual annotation (quite fast 3-4 sec/sentence pair)
- NEL for English named entities (→ Wikidata)

PGV NER already published¹

¹<http://hdl.handle.net/11234/1-5533>

The Parallel Global Voices (PGV) Corpus

PER 1 | ORG 2 | LOC 3 | MISC 4

V rozhovoru, který Timčenkova poskytla v tomto roce, přirovnala sebe sama nikoli k Jelcinovi, nýbrž k Vsevolodu Rudněvovi, veliteli ruského křižníku Varyag, který Rudněv v roce 1904 raději potopil, než aby ho nechal padnout do japonských rukou.

PER 5 | ORG 6 | LOC 7 | MISC 8

In an interview earlier this year, Timchenko compared herself not to Yeltsin but Vsevolod Rudnev, commander of the doomed Russian cruiser Varyag, which Rudnev scuttled rather than let it fall into Japanese hands in 1904.

Named Entity Alignment

Objective Align named entities in Czech-English sentence pairs annotated for NER and NEL.

Process

- Input: Parallel sentences with pre-annotated NER.
- Named entities grouped by the same class (e.g., LOC → LOC, PER → PER).

Complexity Pre-existing annotations in both languages make the task straightforward.

Sentence Embedding Model and Results

Embedding Model

- Sentence transformer:
`distiluse-base-multilingual-cased-v2`.

Alignment Results

- **Precision:** 0.95
- **Recall:** 0.93
- **F1 Score:** 0.94

The alignment process achieves high-quality results, effectively matching entities across languages.

Interesting Observations

Entity Pair Similarities

- Example: *Nike* translated as *Adidas* (similar but incorrect).

Translation Challenges

- English named entity *Beijing police* translated as Czech *pekingská policie* (correct, but not an entity in Czech)
- not only English-Czech translations

PER 1 | ORG 2 | LOC 3 | MISC 4

Ukázkovým příkladem sdružení, které se soustředí na určité cílové skupiny, jako například na domácnosti s jedním rodičem nebo na děti, je asociace "ПОГЛЕД КОН ВИСТИНАТА" (Pohled na pravdu).

PER 5 | ORG 6 | LOC 7 | MISC 8

One case in point of focusing on particular target groups, such as single-parent homes and children, is the association "A View on the Truth."

NER Differences: English vs. Czech

Annotation Challenges (possibly different labels)

- English possessive constructions (e.g., *African stories*) are translated as Czech possessives (e.g., *africké příběhy*) or genitive phrase (e.g., *příběhy Afriky*)

Named Entity Discovery Using Sentence Embeddings

Objective Identify named entities in unannotated Czech sentences using sentence embeddings.

Process

- Generate candidates from all n -grams in Czech sentences (for reasonable ns).
- Match candidates to annotated English entities using cosine similarity.

Goal

- Extend NER discovery to languages without pre-existing NER models.

Named Entity Discovery Using Sentence Embeddings

Named Entities – n -gram Similarity ($n = [1, 2, 3]$)

Steve Oh described Irene as Malaysia's 'Joan of Arc' of maltreated migrants.

Steve Oh líčí Irene jako malajsijskou 'Johanku z Arku' utlačovaných migrantů

Steve Oh, Irene, Malaysia, Joan of Arc



Steve, Steve Oh, Steve Oh líčí, Oh, Oh líčí, Oh líčí Irene, ...

Named Entity Discovery Using Sentence Embeddings

Results on a subset (20 documents)

- Correct: 399
- Incorrect: 28
- Partial: 79
- Missed: 33
- Spurious: 94

[MUC-5]: precision=0.67, recall=0.74, F1=0.70.

Limitations

- Different annotations: *Amitié Hospital* translates as *nemocnice Amitié*.
- In Czech, only the word *Amitié* is annotated.

Conclusion and Future Work

Results for existing NE annotation

- High-quality bilingual NE alignment (**F1 = 0.94**).
- NEL dataset for Czech.

Future Directions

- Named entity discovery method on parallel data without pre-existing NER models for the target language.