

Named Entity Discovery and Alignment in Parallel Data

Zuzana Nevěřilová^a

Natural Language Processing Centre, Faculty of Informatics, Masaryk University,
Botanická 68a, Brno 602 00, Czech Republic

Keywords: Named Entity Recognition, Named Entity Alignment, Named Entity Discovery, Named Entity Linking.

Abstract: The paper describes two experiments with named entity discovery and alignment for English-Czech parallel data. In the previous work, we enriched the Parallel Global Voices corpus with named entity recognition (NER) for both languages and named entity linking (NEL) annotations for English. The alignment experiment employs sentence transformers and cosine similarity to identify NE translations from English to Czech and possibly other languages. The discovery experiment uses the same method to find possible translations between named entities in English and Czech n-grams. The described method achieves an F1 score of 0.94 in finding alignments between recognized entities. However, the same method can also discover unknown named entities with an F1 score of 0.70. The result indicates the method can be used to recognize named entities in parallel data in cases where no NER model is available with sufficient quality.

1 INTRODUCTION

In the previous work (Nevěřilová and Žižková, 2024), we introduced an efficient method for creating parallel named entity (NE) datasets. We benefited from an existing resource, the Parallel Global Voices (Prokopidis et al., 2016), and existing named entity recognition (NER) annotation models. For English, we used the `dslim/bert-large-NER` model from HuggingFace (Devlin et al., 2018; Tjong Kim Sang and De Meulder, 2003). For Czech, we used the `Czert-B` multi-purpose model (Sido et al., 2021). In the project, we performed named entity linking (NEL) to Wikidata. Finally, we published a dataset where the parallel sentences have another two layers of annotation: NER annotation for both languages (classes PERSON, LOCATION, ORGANIZATION, and MISCELLANEOUS) and NEL into Wikidata QNames for the English part of the dataset.

For the disambiguation of English NEs, we used the OpenTapioca platform (Delpuch, 2020) with a re-ranking method that uses sentence transformers¹ (Reimers and Gurevych, 2020). We proved in (Nevěřilová and Žižková, 2024) that re-ranking via sentence transformers was more precise than the default OpenTapioca linking approach. We use sentence


transformers for NE alignment and translation in this follow-up work. The goal is to establish links to Wikidata from the Czech part of the dataset. To our knowledge, there is no NEL dataset for Czech NEs.

In general, the NEs can be linked to other ontologies as well. We selected Wikidata since it contains translations of many QNames to Czech, and later, we can compare the translations with those found in the corpus. The ultimate goal is to establish an efficient method for building datasets with NER and NEL annotations for mid- or low-resourced languages from parallel data.

We conducted two experiments: one was to find alignments between already discovered NEs, and the second was to find the translated NEs. In the second experiment, we removed the previous annotation and let the model discover translations using similarity measures between English NEs and all possible n -grams with n up to 3.

1.1 Paper Outline

Section 2 lists similar projects focusing on NE translation from English to less-resourced languages. Section 3 describes the dataset we used, and Section 4 describes the algorithm that can tackle cases different from 1:1 translations. In Section 5, we present two experiments. The first is relatively straightforward, producing alignment between already recognized enti-

^a  <https://orcid.org/0000-0002-7133-9269>

¹Particularly, we used the model from <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

ties. The second shows that sentence transformers can be used to discover NE translations. We discuss differences that can occur in translations of the NEs, especially the change of part-of-speech and entity type. Section 7 discusses the need for new datasets with NE annotation.

2 RELATED WORK

In the early days, named entity translation has been described, e.g., in (Al-Onaizan and Knight, 2002), which proposes an algorithm for translating NEs between English and Arabic. In (Awadallah et al., 2007), the authors propose a system for NE translation between English and Arabic.

Authors of (Fu et al., 2014) propose a general framework to generate large-scale NER training data from parallel corpora using an English-Chinese parallel corpus. The aim is to improve Chinese NER using English data. A chunk symmetry strategy and English-Chinese transliteration model are used in (Li et al., 2021b).

Named entity translation is closely related to terminology extraction and translation as explored, e.g., in (Deléger et al., 2006). The paper describes using word alignment in parallel corpora to extract new term translations automatically. The work focuses on translating medical terminology from English to French. Terminology extraction in multilingual data has also been proposed as a CLEF-ER challenge (Rebholz-Schuhmann et al., 2013).

In our first experiment, the translation candidates were found in previous work, so the task is only to establish alignment between appropriate candidates. A similar (but harder) task is aligning English NE annotations with other languages. Transformer models are used, e.g., in (Li et al., 2021a), for alignment of NEs in German, Spanish, Dutch, and Chinese, with F1 ranging from 0.71 to 0.81.

Named entity recognition and linking are also related to ontologies and knowledge graphs. In the paper (Stanković et al., 2024), the authors prepared a NER-annotated Italian-Serbian corpus comprising literary works translations. The paper focuses on semantic interoperability as one of the key aspects of linked data and digital humanities.

3 THE PARALLEL DATASET

Parallel Global Voices (PGV (Prokopidis et al., 2016)) is a massively parallel (756 language pairs), automatically aligned corpus of citizen media stories

translated by volunteers. The Global Voices community blog contains several guides, including the Translators' guide². It contains recommendations to "localize" whenever possible. Also, it mentions English as the most significant source language. However, according to authors of the PGV (Prokopidis et al., 2016), the source language for the translation cannot always be reliably identified.

PGV contains texts crawled in 2015, reporting "on trending issues and stories published on social media and independent blogs in 167 countries" (Prokopidis et al., 2016).

The corpus contains the Global Voices (GV) topics about politics and elections; civil, sexual, and socio-economic rights; disasters and the environment; demonstrations and police reaction; labor; and specific geographic regions. In addition, the corpus contains articles about the organization of the GV network, culture, and online media.

The sentence-level alignment has been done automatically. Sometimes, sentence boundaries are incorrectly detected, e.g., on initials inside people's names. The Czech-English pairs (450 documents) are in aligned 1:1 in 86% of cases, the rest are 1:2, 2:1, 1:0, and 0:1 alignments.

We used existing NER models for pre-annotation. They performed well in precision: a BERT-based (Devlin et al., 2018) model achieved 0.70 precision in the MUC-5 strict evaluation scheme (Chinchor and Sundheim, 1993), and Czert-B achieved 0.73 precision in the same evaluation scheme. On the other hand, the recall of English and especially Czech models was low: 0.41 and 0.18, respectively, in the MUC-5 strict evaluation scheme.

In (Nevěřilová and Žižková, 2024), we set up the annotation task with detailed instructions following the UniversalNER (Mayhew et al., 2024) annotation scheme. We have shown that manual annotation could be performed relatively efficiently using high-precision/low-recall pre-annotations.

When set up wisely, the annotation environment allows the production of high-quality annotations quickly: annotation median time was around 4.5 seconds, and the inter-annotator agreement was Cohen- $\kappa=0.91$. We used the LabelStudio³ for annotation. The screenshots in Figures 2 and 1 are from this tool.

²<https://community.globalvoices.org/guide/lingua-guides/lingua-translators-guide/>

³<https://labelstud.io/>

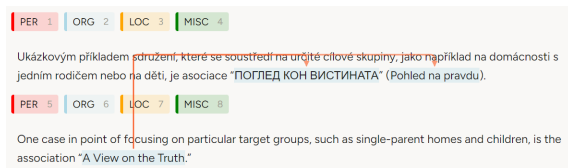


Figure 1: Example with multiple translations: The organization name is mentioned and translated. The model establishes two alignments.

4 ALGORITHM FOR NAMED ENTITY ALIGNMENT

For the parallel sentence pairs with marked NEs, we established the following algorithm that finds the translation:

Data: pairs of English–Czech sentences with annotated named entities
Result: pairs of English–Czech named entities
 encode all NEs into embeddings using sentence transformers⁴
 calculate cosine similarities matrix for NEs in source and target languages
while *similarity matrix* \neq *zero matrix* **do**
 find the most similar pair (i, j) **if** *NEs at positions (i, j) share the same class* **then**
 establish an alignment $R(i, j)$; set all similarities in row i and column j with a similarity lower than a threshold to 0;
end
end

Algorithm 1: NE alignment for annotated pairs of sentences.

The reason for such an apparently complicated algorithm is that the translations are not necessarily 1:1. As shown in Figure 1, in some cases, a foreign NE is mentioned and translated to Czech in the same sentence. The algorithm does not discard an already used NE; instead, it tries to find other similarities with the NE.

The downside of this approach is that the algorithm cannot distinguish the order of the NEs. Figure 2 shows the annotation corrected manually. The model proposed relations between all mentions of *Nigeria*.



Figure 2: Example with multiple NE occurrences: *Nigeria* is mentioned two times. The model can find the translation but finds relations between all occurrences without preserving their order.

5 RESULTS

We evaluated the proposed alignments against manual annotation. We selected the MUC-5 evaluation scheme (Chinchor and Sundheim, 1993).

The MUC-5 distinguishes five cases:

- CORrect – the predicted value equals the ground truth
- INCorrect – the predicted value does not equal the ground truth (in NER evaluation, this is used for an incorrect label)
- PARTIally correct – in NER, a partial overlap between predicted and ground truth data exist
- MISsed – prediction did not find a value
- SPURious – prediction found a false positive value

5.1 Experiment 1: Find Alignments Among Annotated NEs

We only considered COR, MIS, and SPU cases where the alignment was correctly established, missed, or added extra, respectively.

We selected a subset of 20 documents for manual annotation. The subset contains 590 sentence pairs, 373 containing entities, and 320 containing relations. The total numbers of entities are 763 and 684 for sources and target languages, respectively. One NE pair per sentence pair is the most common situation. The ground truth contains 539 entity pairs.

The results for different similarity thresholds are presented in Table 1. The threshold does not affect the results significantly since, in most cases, there is only one possibility of aligning the NEs.

5.1.1 Discussion

The similarity threshold can be the subject of further experiments. When set too high, similar NEs with

Table 1: Number of CORrect, MISsed, and SPURious alignments, together with precision (P), recall (R), and F1 for different similarity thresholds T .

T	COR	MIS	SPU	P	R	F1
0.2	569	43	32	0.95	0.93	0.94
0.3	568	44	33	0.95	0.93	0.94
0.4	568	44	34	0.94	0.93	0.94
0.5	568	44	35	0.94	0.93	0.93
0.6	568	44	37	0.94	0.93	0.93
0.7	568	44	38	0.94	0.93	0.93
0.8	567	45	39	0.94	0.93	0.93
0.9	564	48	40	0.93	0.92	0.93

different parts of speech can be missed. In addition, the embedding similarity can differ from language to language, i.e., the similarity between an English NE and its Czech translation can be higher than between an English NE and its Macedonian translation.

The similarity threshold was set to 0.2. However, we could see that the model incorrectly translated similar NEs, such as *Adidas* to *Nike*.

On the other hand, the model can find NE translations, even if they are incomplete or contain typos. This observation encouraged us in the second experiment.

5.2 Experiment 2: Discovering New Entities

In this case, we simplified the algorithm only to take the most similar pair of English NE and Czech n -gram. We set up the maximum n to 3 since the majority of the NEs are one-word, two-word, or three-word expressions. We were aware that longer NEs can only be found partially with this method.

We selected the MUC-5 schema for evaluation since it is more informative than an F1. In addition, we were aware that we should not include longer NEs, so incorporating a partial match seemed to be a good option.

For the comparison, we discarded trailing punctuation, including quotes and brackets. We also removed the possessive 's since there is no strong agreement on whether it should be part of the NE.

The results with the test subset are as follows:

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

According to the MUC-5 evaluation schema, the precision of the method is 0.67, recall is 0.74, and F1 score is 0.70.

The error analysis showed that some partial matches contained additional common words, such as articles. In other cases, the partial matches were correct translations but incorrect NEs. For example, *Amitié Hospital* translates as *nemocnice Amitié*; however, in the Czech data, only the word *Amitié* is annotated as an NE. A similar situation holds for many spurious NEs, where, e.g., *Chinese* is annotated as an NE, while its Czech translation *čínský* is not.

5.2.1 Discussion

In this experiment, we did not consider other possible n -gram features, such as similarity to English NE or uppercase. Even though the method was set up straightforwardly, we achieved reasonable results. In experiments with other languages, we need to consider different coverages of the other languages in the sentence embedding model. We expect low-resourced languages to have lower similarities with English than Czech, a mid-resourced language.

5.3 Cross-Lingual Named Entity Annotation Issues

The UniversalNER community does not agree yet on how to annotate possessives. Since English possessives are (proper) nouns with the 's, they are easily considered (proper) nouns within the NER task. On the other hand, Slavonic languages use two competing strategies on how to express possession: genitive noun phrases (e.g., *kniha pana profesora*, *book of the professor* meaning *professor's book*) and possessive adjectives (e.g., *Alicina kniha*, *Alice's book*). As described in (Janda and Townsend, 2000), adjectives formed from names of human males (with suffix *-ův*) and females (with suffix *-in*) have a paradigm distinct from other adjectives in Czech. The translations found in the dataset are not exact since sometimes the part-of-speech is not preserved.

A related issue is in translation pairs that contain a noun in English but an adjective in Czech. For example, *pekingská policie* is (correctly) translated as *Beijing police*. From the NER annotation point-of-view, the Czech expression is not an NE, while *Beijing* is a LOCATION. Our method cannot find an NE translation, although the expressions can be translated.

In some cases, we observed the inverse case. For example, *African stories* are translated as *příběhy Afriky* (using a genitive noun phrase). In such cases, the NEs are not translated at all since *African* is an entity of type MISC (similar to national origin). In contrast, *Afrika* (*Africa*) is a LOCATION.

6 DISCUSSION

Although NER is a common NLP task that is solved very well in high-resourced languages, we argue that using well-established benchmarks such as ConLL-2003 (Sang and Meulder, 2003) for training new models is not always the best choice.

The English part of the ConLL-2003 is from the news stories of the Reuters corpus from 1996 to 1997. Although the PGV dataset is the same domain – news texts – we observed a considerable shift in topics and style.

In the English ConLL-2003 training data, the most prevalent LOCations are the U.S., Germany, Australia, Britain, and France. In PGV, the most frequent LOCations are China, Hong Kong, Mexico, Russia, and Iran. A similar shift in focus is visible from the most prevalent person names – Clinton in ConLL-2003 and Erdogan in PGV. In PGV, many person names do not start with uppercase since they are social media nicknames (e.g., @realDonaldTrump). Sometimes, social media nicknames refer to organizations (e.g., @greenpeaceindia).

In the previous work (Nevěřilová and Žižková, 2024), we discovered that even the state-of-the-art NER model⁵ that performed well on the ConLL-2003 benchmark had issues with recall on PGV. The precision of the model was lower on PGV (0.77 for exact match, 0.82 for partial match) than reported on the ConLL-2003 data (0.91). The recall dropped more significantly from 0.92, reported by the author of the model, to 0.45 and 0.49 for exact and partial matches, respectively, on PGV data.

The most frequently missed entities were person names that do not start with uppercase and organization names. The incorrect annotation appeared between classes ORG, PER, and MISC since MISC contains artwork, product, and media names and nationalities, languages, and social group names. We think that providing more recent data that reflect the shift of journalism towards social media can be beneficial even for high-resourced languages, not speaking about the mid- and low-resourced ones.

7 CONCLUSION AND FUTURE WORK

The result of our task is two-fold: we have proposed a straightforward method that aligns already annotated data. We also discovered the same method can be

used to annotate parallel data using only English NE annotation. Possible improvements comprise variable maximum n -gram length and additional information, such as capitalization.

We used the same method (sentence embedding similarity) for two cases: finding alignments between English and Czech NEs and discovering NE translations in Czech. We plan to experiment with other language pairs. The goal is to provide an efficient method for NE annotation in mid- and low-resourced languages. The efficiency can be achieved by pre-annotations composed of high-precision (possibly low-recall) NER model annotations and NE translation.

Apart from the method, this work’s contribution is a new NEL-annotated dataset for Czech. The dataset is published in the CLARIN repository⁶ under the Creative Commons – Attribution 4.0 International (CC BY 4.0), and all Python codes are published in the GitLab repository⁷.

ACKNOWLEDGEMENTS

This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT/CLARIAH-CZ project LM2023062.

REFERENCES

- Al-Onaizan, Y. and Knight, K. (2002). Translating named entities using monolingual and bilingual resources. In *ACL*, pages 400–408.
- Awadallah, A., Fahmy, H., and Hassan Awadalla, H. (2007). Improving named entity translation by exploiting comparable and parallel corpora.
- Chinchor, N. and Sundheim, B. (1993). MUC-5 Evaluation Metrics. In *Fifth Message Understanding Conference*.
- Deléger, L., Merkel, M., and Zweigenbaum, P. (2006). Contribution to terminology internationalization by word alignment in parallel corpora. *AMIA Annu Symp Proc*, 2006:185–189.
- Delpeuch, A. (2020). OpenTapioca: Lightweight Entity Linking for Wikidata.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Fu, R., Qin, B., and Liu, T. (2014). Generating chinese named entity data from parallel corpora. *Front. Comput. Sci.*, 8(4):629–641.

⁵Particularly, BERT Large NER from <https://huggingface.co/dslim/bert-large-NER>

⁶<http://hdl.handle.net/11234/1-5533>

⁷https://gitlab.fi.muni.cz/nlp/named_entity_linking

- Janda, L. and Townsend, C. (2000). *Czech. Languages of the world: Materials*. Lincom Europa.
- Li, B., He, Y., and Xu, W. (2021a). Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment.
- Li, P., Wang, M., and Wang, J. (2021b). Named entity translation method based on machine translation lexicon. *Neural Computing and Applications*, 33(9):3977–3985.
- Mayhew, S., Blevins, T., Liu, S., Šuppa, M., Gonen, H., Imperial, J. M., Karlsson, B. F., Lin, P., Ljubešić, N., Miranda, L., Plank, B., Riabi, A., and Pinter, Y. (2024). Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark.
- Nevěřilová, Z. and Žižková, H. (2024). Named entity linking in english-czech parallel corpus. In E. Nöth, A. Horák, P. S., editor, *Text, Speech, and Dialogue, 27th International Conference, TSD 2024, Part I*, pages 147–158, Switzerland. Springer International Publishing. Mezinárodní význam, Recenzováno.
- Prokopidis, P., Papavassiliou, V., and Piperidis, S. (2016). Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, SI. European Language Resources Association (ELRA).
- Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., and Kors, J. A. (2013). Entity recognition in parallel multi-lingual biomedical corpora: The clef-er laboratory overview. In Forner, P., Müller, H., Paredes, R., Rosso, P., and Stein, B., editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, pages 353–367, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition.
- Sido, J., Pražák, O., Příbáň, P., Pašek, J., Seják, M., and Konopík, M. (2021). Czert – Czech BERT-like Model for Language Representation. In Mitkov, R. and Angelova, G., editors, *Proc. of the International Conference on Recent Advances in NLP (RANLP 2021)*, pages 1326–1338, Held Online. INCOMA Ltd.
- Stanković, R., Ikonić Nešić, M., Perisic, O., Škorić, M., and Kitanović, O. (2024). Towards semantic interoperability: Parallel corpora as linked data incorporating named entity linking. In Chiarcos, C., Gkirtzou, K., Ionov, M., Khan, F., McCrae, J. P., Ponsoda, E. M., and Chozas, P. M., editors, *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024*, pages 115–125, Torino, Italia. ELRA and ICCL.
- Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.