Named Entity Alignment in Czech-English Parallel Data

Zuzana Nevěřilová[1] and Hana Žižková[2]

[1] Natural Language Processing Centre
Faculty of Informatics
Botanická 68a, Brno, Czech Republic 3839@mail.muni.cz
[2] Department of Czech Language
Faculty of Arts
Arna Nováka 1, Brno, Czech Republic zizkova@phil.muni.cz

Abstract. The paper introduces an approach to named entity alignment for Czech-English parallel data. We enriched the Parallel Global Voices corpus with named entity recognition (NER) and named entity linking (NEL) annotations. The new annotation layer employs sentence transformers and cosine similarity to identify NE translations across English, Czech, and possibly other languages, considering atypical entity pairings and achieving an F1 score of 0.94 on evaluated samples.

Keywords: named entity recognition, named entity linking, named entity translation, sentence transformers.

1 Introduction

In our previous work [10], we introduced an efficient method for creating parallel named entity (NE) datasets. We benefited from an existing resource, the Parallel Global Voices [11], and existing models for named entity recognition (NER) annotation. For English, we used the `dslim/bert-large-NER` model from HuggingFace [5,13]. For Czech, we used the Czert-B multi-purpose model [12]. In the project, we aimed to perform named entity linking (NEL) to Wikidata. Finally, we published a dataset where the parallel sentences have another two layers of annotation: 1. NER for classes: PERson, LOCation, ORGanization, and MISCellaneous 2. NEL with links of some English NEs into Wikidata QNames

For the disambiguation of English NEs, we used the OpenTapioca platform [4] with a re-ranking method that uses sentence transformers[3]. In this work, we use the latter for cross-language named entity linking. The goal is to find new links to Wikidata, even for Czech NEs. The ultimate goal is to propose an efficient method for building datasets with NER and NEL annotations from parallel data.

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 59–64, 2024. © Tribun EU 2024

2 Related Work

Named entity translation has been e.g., in [1], which proposes an algorithm for translating NEs between English and Arabic. In [2], the authors propose a system for NE translation between English and Arabic. A chunk symmetry strategy and English–Chinese transliteration model are used in [8].

In our case, the translation candidates were found in previous work. The task is more straightforward – to establish NE alignment between appropriate candidates. A similar (but harder) task is aligning English NE annotations with other languages. Transformer models are used, e.g., in [7], for alignment of NEs in German, Spanish, Dutch, and Chinese, with F1 from 0.71 to 0.81.

3 PGV NER Dataset

Parallel Global Voices (PGV [11]) 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' guide[4]. It contains recommendations to "localize" whenever possible. Also, it mentions English as the most significant source language. However, according to authors of the PGV [11], 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" [11].

The NER annotation with the existing tools performed well in terms of precision: BERT-large-NER achieved 0.77 precision in the MUC-5 exact evaluation scheme [3], and Czert-B achieved 0.79 precision in the same evaluation scheme. On the other hand, the recall of English and especially Czech models was low: 0.45 and 0.2, respectively, in the MUC-5 exact evaluation scheme. In [10], we set up the annotation task with detailed instructions following the Universal-NER [9] annotation scheme. We proved that manual annotation could be performed relatively efficiently using high-precision/low-recall pre-annotations.

The dataset of NE annotations in the parallel data (the NER ground truth) was published in the LINDAT/CLARIAH-CZ repository[5], together with links to Wikidata (the NEL ground truth) for English NEs. In addition, we published the NER model for Czech based on RobeCzech and trained on CNEC[6].

4 Alignment Implementation

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

[4] https://community.globalvoices.org/guide/lingua-guides/lingua-translators-guide/

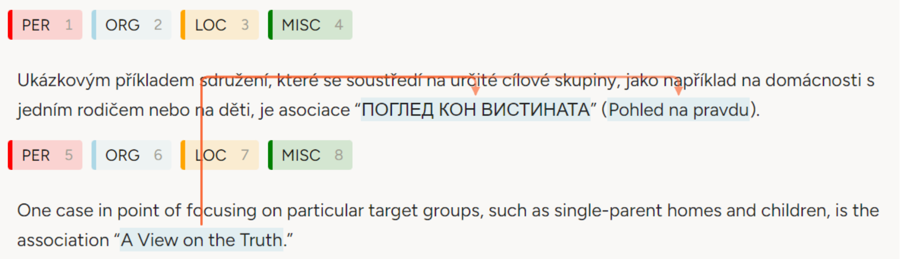
[5] http://hdl.handle.net/11234/1-5533

[6] https://huggingface.co/popelucha/robeczech-NER

1. encode all NEs into embeddings using sentence transformers[7]
2. calculate cosine similarities matrix for NEs in source and target languages
3. iterate from the most similar pairs:
   (a) if the NEs at positions (i, j) share the same class, establish an alignment $R(i, j)$
   (b) set all similarities in row i and column j with a similarity lower than a threshold to 0
   (c) repeat until the similarity matrix is not a zero matrix

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.

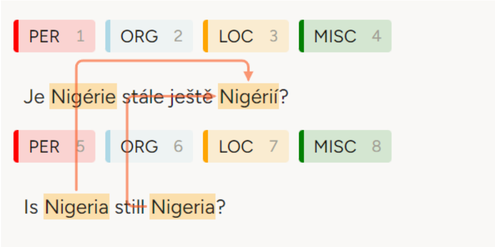
The downside of this approach is that the algorithm cannot distinguish the order of the NEs. Figure 2 shows the annotation corrected manually. The model proposed relations between all mentions of *Nigeria*.



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).

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.”

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



Je Nigérie stále ještě Nigérií?

Is Nigeria still Nigeria?

Fig. 2: Example with multiple NE occurrences: *Nigeria* is mentioned two times. The model can find the translation; however, it finds relations between all occurrences.

[7] `sentence-transformers/distiluse-base-multilingual-cased-v2`

5 Results

We evaluated the proposed alignments against manual annotation. We selected the MUC-5 evaluation scheme[3], although it was not considered to be used for relations. 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

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

We selected a subset of 20 documents for manual annotation. The subset contains 590 sentence pairs, 373 of which contain entities, 320 contain 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 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 how to align the NEs.

Value	$t = 0.2$	$t = 0.3$	$t = 0.4$	$t = 0.5$	$t = 0.6$	$t = 0.7$	$t = 0.8$	$t = 0.9$
CORrect	569	568	568	568	568	568	567	564
MISsed	43	44	44	44	44	44	45	48
SPUrious	32	33	34	35	37	38	39	40
precision	0.95	0.95	0.94	0.94	0.94	0.94	0.94	0.93
recall	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.92
F1	0.94	0.94	0.94	0.93	0.93	0.93	0.93	0.93

Table 1: No. of missed and spurious alignments for different similarity thresholds t.

5.1 Discussion

Some translation errors originate in the algorithm and its inability to calculate the correct order of NE pairs. This leads to a lower F1 score but does not affect the quality of the translation pairs.

The similarity threshold can be the subject of further experiments. When set too high, similar NEs with 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 examined, with best values around 0.3. We could see that with lower threshold, the model incorrectly translated similar NEs, such as *Adidas* to *Nike*.

On the other hand, with a reasonable threshold, the model can find NE translations, even if they are incomplete or contain typos. This can be a benefit for the planned task where the NEs in the target language can be found a priori using the embedding similarities between all candidates (see Section 6.1).

5.2 Known Issues

The UniversalNER community does not agree 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 [6], 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 we translate nouns to adjectives.

A related issue is in translation pairs that contain a noun in English but an adjective in Czech. For example, *pekingská policie* is 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 Conclusion and Future Work

The result of our task is multifold: by another manual annotation, we can identify and correct annotation errors. We plan to release a new version of the NER-NEL dataset. We also plan to incorporate Czech NEL via the English NEL and alignments.

6.1 Finding NEs in Unannotated Data

The procedure (embedding similarity) can be used for unknown data. We plan to experiment with English-Czech sentences without the Czech NER annotation. The question is how well the model can find Czech NEs. Instead of known NEs, we can use all n-grams from the sentence as NE candidates and let the model select the best ones.

If successful, this method could be used for language pairs with no NER model for the target language.

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

1. Al-Onaizan, Y., Knight, K.: Translating named entities using monolingual and bilingual resources. In: ACL. pp. 400–408 (2002), http://www.aclweb.org/anthology/P02-1051.pdf
2. Awadallah, A., Fahmy, H., Hassan Awadalla, H.: Improving named entity translation by exploiting comparable and parallel corpora (2007), https://www.microsoft.com/en-us/research/publication/improving-named-entity-translation-exploiting-comparable-parallel-corpora/
3. Chinchor, N., Sundheim, B.: MUC-5 Evaluation Metrics. In: Fifth Message Understanding Conference (August 1993), https://aclanthology.org/M93-1007
4. Delpeuch, A.: OpenTapioca: Lightweight Entity Linking for Wikidata (2020)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR **abs/1810.04805** (2018), http://arxiv.org/abs/1810.04805
6. Janda, L., Townsend, C.: Czech. Languages of the world: Materials, Lincom Europa (2000), https://books.google.cz/books?id=6VkbAQAAIAAJ
7. Li, B., He, Y., Xu, W.: Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment (2021), https://arxiv.org/abs/2101.11112
8. Li, P., Wang, M., Wang, J.: Named entity translation method based on machine translation lexicon. Neural Computing and Applications **33**(9), 3977–3985 (May 2021)
9. 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., Pinter, Y.: Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark (2024)
10. Nevěřilová, Z., Žižková, H.: Named Entity Linking in English-Czech Parallel Corpus. In: E. Nöth, A. Horák, P.S. (ed.) TSD 2024. pp. 147–158. Springer International Publishing, Switzerland (2024). https://doi.org/http://dx.doi.org/10.1007/978-3-031-70563-2_12
11. Prokopidis, P., Papavassiliou, V., Piperidis, S.: 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). pp. 900–905. European Language Resources Association (ELRA), Portorož, SI (May 2016), https://aclanthology.org/L16-1144
12. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., Konopík, M.: Czert – Czech BERT-like Model for Language Representation. In: Mitkov, R., Angelova, G. (eds.) Proc. of the International Conference on Recent Advances in NLP (RANLP 2021). pp. 1326–1338. INCOMA Ltd., Held Online (Sep 2021), https://aclanthology.org/2021.ranlp-1.149
13. Tjong Kim Sang, E.F., De Meulder, F.: 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. pp. 142–147 (2003), https://www.aclweb.org/anthology/W03-0419