Transfer-based Enrichment of a Hungarian Named Entity Dataset

Attila Novák and **Borbála Novák**
MTA-PPKE Hungarian Language Technology Research Group,
Pázmány Péter Catholic University
Faculty of Information Technology and Bionics
Práter u. 50/a, 1083 Budapest, Hungary
{surname.firstname}@itk.ppke.hu

Abstract

In this paper, we present a major update to the first Hungarian named entity dataset, the Szeged NER corpus. We used zero-shot cross-lingual transfer to initialize the enrichment of entity types annotated in the corpus using three neural NER models: two of them based on the English OntoNotes corpus and one based on the Czech Named Entity Corpus fine-tuned from multilingual neural language models. The output of the models was automatically merged with the original NER annotation, and automatically and manually corrected and further enriched with additional annotation, like qualifiers for various entity types. We present the evaluation of the zero-shot performance of the two OntoNotes-based models and a transformer-based new NER model trained on the training part of the final corpus. We release the corpus and the trained model.

1 Introduction

1.1 Resources

Named entity recognition is a fundamental NLP task that has played an important role in tasks like information extraction, document deidentification, conversational models, etc. Following the annotation scheme used in the CoNLL 2002/2003 NER annotation tasks, legacy named entity corpora usually contain annotation of four entity types: organizations (ORG), persons (PER), locations (LOC) and general entity category covering all the rest (MISC). This is the case for all named entity corpora available for Hungarian, the Szeged NER corpus (Szarvas et al., 2006), the silver-standard Hungarian hunNERwiki corpus (Simon and Nemeskey, 2012) automatically derived from Wikipedia, and the recently published NerKor corpus.[1] The English OntoNotes 5 corpus (Weischedel et al., 2013), on the other hand, contains a richer set of entities. Geopolitical entities (GPE: countries, settlements, etc.) and facilities (FAC: buildings, roads, airports etc.) are differentiated from geographical locations like continents or bodies of waters. Within the MISC category, products (PROD), laws and other norms (LAW), events (EVENT) and titles of works of art (WORK_OF_ART) are differentiated. In addition, the OntoNotes NER tagset also encompasses time and numerical expressions distinguishing dates and times, cardinal and ordinal numbers, quantities, percentages and amounts of money. In addition, other categories covering non-entities like languages (LANGUAGE) and nationalities, religions and political affiliations (NORP 'nationality/other/religion/political') are covered, presumably just because English orthography happens to prescribe capitalization for words (adjectives in the case of NORP) belonging to this category.

Some resources in languages other than English also use NER tagsets richer than the basic four-class tagset. Although the NoSta-D resource used in the GermEval2014 shared task targeting German NER (Benikova et al., 2014) maintains a four-class distinction, words (especially adjectives) derived from names as well as compounds containing them are marked as such. This corpus, similarly to other resources like the GENIA corpus (Kim et al., 2003) containing biomedical entities and the Spanish and Catalan newspaper text corpus AnCora (Taulé et al., 2008), also features nested named entities. One of the most richly annotated NER corpora is the Czech Named Entity Corpus (Ševčíková et al., 2007). It contains both a rather rich set of entity types and nested entities.

1.2 Architectures for Sequence Tagging and Cross-lingual Transfer

Legacy data-driven statistical machine learning algorithms based on Hidden Markov Models (Baum

[1] https://github.com/nytud/NYTK-NerKor

Proceedings of Recent Advances in Natural Language Processing, pages 1060–1067
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_119

and Petrie, 1966), Maximum Entropy models (Ratnaparkhi, 1996) and CRF (Lafferty et al., 2001) provided then state-of-the-art performance for sequence tagging, however they relied on data and features pertaining strictly to the target language. This meant that a significant amount of annotated training data in the target language was required to attain acceptable performance using these models.

The paradigm shift to neural models offered the possibility of changing this situation. Already the simplest non-contextual distributional word embedding models like word2vec (Mikolov et al., 2013a,c) were discovered to have some kind of inherent language-independent property. It was found that models trained on different languages independently can be mapped to each other with high accuracy using a rather limited bilingual vocabulary (Luong et al., 2015) or even in an unsupervised manner (Mikolov et al., 2013b).

It was also discovered that, with neural machine translation models, it is possible to improve performance in specific lower-resource languages simply by training the encoder and the decoder of the model in a shared manner on multiple languages. This resource-sharing also made direct translation between all of the represented languages possible, and resulted in savings in resources concerning both training, storage and inference, i.e. using the model in production. The models offering state-of-the-art performance in machine translation performed similarly well in other NLP tasks. Pretraining the encoder of models used in NMT (especially the now-ubiquitous transformer architecture (Vaswani et al., 2017)) for simple mask-filling tasks on high amounts of (monolingual) plain text resulted in contextual language models that could be fine-tuned for specific tasks in a much more efficient manner than training similar models from scratch (Devlin et al., 2019). These models significantly improved the state-of-the-art for nearly all NLP-related tasks even for high-resource languages like English. The improvement is even more significant in the case of lower-resource languages.

Multilingual training turned out to be fruitful not only in the domain of machine translation. The publication of multilingual contextual language models like multilingual BERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2019) made cross-lingual knowledge transfer efficient for other NLP tasks as well. It is possible to fine-tune the language model for e.g a token classification task, like named entity recognition in one language and apply it to another language. Even in a zero-shot scenario, where the token classification model has not seen any training data in the target language, it can provide a reasonable performance, especially if the target language is included among the languages covered by the underlying language model.

On the other hand, models trained on multiple languages were found not to provide state-of-the-art performance if a significant amount of training data is available for the given task in the target language. Fine-tuning a monolingual language model for a specific task usually results in better performance than using a heavily multilingual model like multilingual BERT, because the target language is usually relatively underrepresented in the underlying multilingual model (Martin et al., 2020). In this paper, we present a data annotation scenario in which we used zero-shot transfer to preannotate a language resource, which was then manually corrected and enriched to create a resource that can then be used to train a monolingual model to optimize performance.

2 Method

In the project presented in this paper, we significantly enriched the annotation in the first Hungarian named entity dataset, the Szeged NER corpus (Szarvas et al., 2006).

2.1 Zero-shot Preannotation

When preannotating the corpus, we fed the tokens to two models trained on the English OntoNotes 5 NER corpus. The first model was created by the DeepPavlov team fine-tuning multilingual BERT (Burtsev et al., 2018). The other model is based on XLM-RoBERTa, a multilingual contextual language model trained on a significantly bigger multilingual corpus than multi-BERT. The latter model is part of the FLAIR tool set (Akbik et al., 2019).

The two models use different tokenization following the tokenization scheme of the underlying contextual language model. The token sequence in the output of the models was thus different from the original input token sequence. This had to be taken into account when merging the annotation from the models with the original annotation. The merging procedure was automatic. While merging the annotations, in the case of overlapping entity spans, we considered the spans in the input anno-

tations gold standard, and if the zero shot model suggested a compatible entity subtype, we updated the entity type. E.g. an entity of type location (LOC) in the original annotation is compatible with any of geographical location (LOC), facility (FAC) and geopolitical entity (GPE). Annotation of non-entities, like dates, quantities and nationalities not present in the original annotation were introduced based on the output of the models.

2.2 Error Analysis and Automatic Error Correction

We identified typical errors of the zero-shot models that could be corrected automatically using regular-expression-based patterns. We have found that, in the case of transfer from English to Hungarian, a typical problem is that for many named entity types like names of organizations, bodies of waters, titles of works of art etc., a definite article is present in most but not all cases in Hungarian, while there is no article in English. This resulted in the model including definite articles for these types of entities in the annotation, an error that could be easily eliminated from the output.

While cross-lingual mapping resulted in some anomalies like inclusion of definite articles, it had other side-effects that we found to be useful. The output of the models also included annotation for adjectives derived from named entities like *londoni* 'of London'. In contrast to the German NoSta-D corpus, words like this remained unannotated in all legacy Hungarian named entity corpora in spite of the fact that the identification of these words as references to named entities would be desirable in practical applications like information retrieval. We thus decided to keep this kind of annotation as part of our annotation enrichment effort.

After automatic correction of entity spans and types, we manually merged the outputs of the two models by checking the differences of the two annotations.

2.3 Considering a Third Model

We also applied a third model to the corpus. We used he Czech model of the NameTag 2 neural named entity tagger (Straková et al., 2019) trained on the Czech Named Entity Corpus CNEC 2 (Ševčíková et al., 2007). The underlying corpus and thus also the model contains a very fine-grained set of entity classes offering many subclasses within the broader categories like a distinction of companies vs. governmental/political institutions vs. academic/educational/cultural institutions and conferences/contests (the latter are also considered a subclass of organizations). NameTag 2 is capable of returning nested annotations (with a maximal depth of two overlapping entities). The model can be accessed via a web service. However, at least in the zero-shot cross-lingual setting, the annotation generated by this model seemed to be less accurate than those generated by OntoNotes-based models. Since there are no definite articles in Czech, this model had a similar problem including definite articles for the types of entities (e.g. organizations) that often appear with a definite article in Hungarian. It often generated two overlapping annotations for these types of entities differing only in whether the article is included. More importantly, the different occurrences of the same entity were often assigned different classes (usually this was an error rather than real ambiguity due to metonymic use). Also the extent of the span of the entities was less accurate than in the annotation generated by the English-based models. The subclassification itself also introduces problems of its own. It is not clear where sports clubs or central banks like the *Bank of England* should belong in this taxonomy.

2.4 Introduction of New Entity Types

Nevertheless, we found good use of the annotation generated by NameTag 2. As Hungarian is an agglutinating language and thus words appear in many different suffixed forms in the corpus, we applied lemmatization to the entity annotations generated by all models and aggregated the results listing the frequency of alternative annotations for the same entity. Tags generated by the FLAIR OntoNotes model and NameTag2 for the most frequent organizations in the corpus are shown in Table 1. Tags containing a hyphen in columns 3 to 8 were assigned by NameTag2, the rest by the FLAIR OntoNotes model. The list features the Budapest Stock Exchange *(Budapesti Értéktőzsde = BÉT)*, the Budapest Commodity Exchange *(Budapesti Árutőzsde)*, *Nasdaq, Wall Street*, the Hungarian Central Bank *(Magyar Nemzeti Bank)* and news agencies *(MTI, MTI-ECO, Reuters)*. It is obvious that NameTag 2 struggles trying to assign them the right category. We thus refrained from adopting the taxonomy in CNEC 2.

On the other hand, the automatically generated gazetteer helped us identify entities really belonging to certain well-distinguishable entity classes

entity	#	tag	#	tag	#	tag	#	tag	#	tag	#	tag
Budapesti Értéktőzsde	766	ORG	332	MEDIA-JOUR	251	ORG-COM	37	MEDIA-RTV	1	GPE-SETL		
MTI	453	ORG	231	ORG-COM	112	MEDIA-JOUR	73	ORG-GOV	2	PROD	1	ORG-CONF
Budapesti Árutőzsde	177	ORG	64	ORG-COM	1	GPE-SETL						
MTI-ECO	143	ORG	72	ORG-COM	65	MEDIA-JOUR	1	ORG-GOV				
Magyar Nemzeti Bank	138	ORG	84	ORG-COM	38	ORG-GOV	12	MEDIA-JOUR	1	LAW		
Reuters	132	ORG	122	MEDIA-JOUR	8	ORG-COM						
Wall Street	121	ORG	69	ORG-COM	25	GPE-URB	21	FAC-STR	9	FAC	4	MEDIA-JOUR
Nasdaq	104	ORG	65	ORG-COM	5	MISC	2	MISC-ORG	2	ORG-EDU		
BÉT	88	ORG	46	MEDIA-JOUR	27	ORG-COM	6	ORG-GOV	1	PROD	1	MISC

#	tag
1	ORG-EDU
4	ORG-EDU
1	MEDIA-RTV

Table 1: Most frequent organizations in the corpus with several different annotations generated by the NameTag 2 tagger (labels containing a hyphen).

(like media) and also often mistagged entities. We generated regular-expression-based automatic correction patterns using manually marked entries from this automatically generated gazetteer (covering also suffixed forms), and mass-corrected annotations using these patterns.

We also discovered that certain types of expressions are mistagged by the zero-shot models due to lack of distinction in the original underlying OntoNotes annotation. One such example was expressions referring to time durations like 'for five minutes' or 'six-day-long'. While other types of quantities are annotated in the OntoNotes NER resource as quantities resulting in sensible annotation also for the Hungarian input, the model mistagged duration expressions as time or date, which should only refer to expressions anchored to the timeline. We thus introduced a new entity type DUR to annotate unanchored duration expressions, and the annotation of many occurrences of this type of expressions could also be automatically introduced/corrected. We also annotated relative date expressions like days of week.

The Szeged NER corpus consists of business news, and due to its genre, it contains many occurrences of certain entity types not covered by the OntoNotes NER tagset: e.g. names of securities and stock exchange indexes. We introduced new tags for these entity types. They were also easy to identify in the generated gazetteer. We also introduced a tag for media like newspapers, broadcasting services and online news portals but refrained from distinguishing subtypes. This type of entities are somewhat similar to but can easily be distinguished from books (covered by the work of art category). On the other hand, they also involve an entity of the organization type (the publisher/redaction).

2.4.1 Metonymic Use of Names

Metonymic use of entities like referring to countries or other geopolitical entities as actors is usually annotated according to the actual metonymic sense. In the recently published NerKor corpus annotated with the coarse-grained ORG-PER-LOC-MISC tagset, references to countries as actors like *Germany invaded France* are annotated as ORG rather than LOC. This kind of metonymy is completely productive for all types of geopolitical entities, and annotating them as such solves the problem in a more elegant way than what the coarse-grained tagset makes possible. Incidentally, this

specific sort of metonymy is less prevalent in the business news genre than in certain other genres. On the other hand, *Wall Street,* one of the top entries on the organizations list in Table 1, is an example of an expression typically used in a metonymic sense referring to the New York Stock Exchange (and related financial institutions).

2.4.2 Qualifiers and Relations

In addition to the annotation mentioned above, we introduced two more tags that could be used to annotate qualifiers of named entities: QUAL and REL. These were used in situations where the named entity had a nominal modifier like in *Ante Vulin*$_{PER}$ *építész*$_{QUAL}$ 'architect Ante Vulin' or it was part of an appositive structure like in *Jurij Lvov*$_{PER}$ *pénzügyminiszterhelyettes*$_{QUAL}$*, Vlagyimir Putyin*$_{PER}$ *elnök*$_{QUAL}$ *bizalmasa*$_{REL}$ 'Yuri Lvov$_{PER}$, Deputy Minister of Finance$_{QUAL}$, confidant$_{REL}$ of President$_{QUAL}$ Vladimir Putin$_{PER}$'. REL was used in situations where the phrase expresses a relation between named entities (if both are part of the same noun phrase), QUAL for other modifiers. We used this type of annotation to mark nominal phrases that are not named entities but as qualifiers of named entities have the same type of reference as the related named entity. We pregenerated and later refined this annotation using syntactic dependency parses of the sentences. The dependency structure could be used to identify the modified entities and thus their type, e.g. that *elnök* 'president' is a qualifier of persons and being a confidant is a relation between persons.

2.5 Filtering and Manual Annotation

We also filtered the corpus for repetitive boilerplate-like content: we removed identical sentences and ones differing only in numerical/date expressions. After creating the preannotation using the zero-shot models, automatically merging them with the original annotation and applying pattern-based corrections, we used the INCEpTION annotation framework (Klie et al., 2018) to correct and augment the annotations. Two researchers and five MA students of theoretical linguistics participated in the manual annotation process. Each document was revised by at least two annotators. Curation and final processing of the results was performed by a single researcher. In the current version, we refrained from generating nested annotation although currently the development of nested entity classifiers gained momentum, and some open-source neural nested entity taggers are available e.g. (Shibuya and Hovy, 2020) or (Wang et al., 2020). Although these models have sub-SOTA performance on flat NER datasets (we also found Nametag 2 to generate much less accurate annotation than the OntoNotes 5-based models), we will consider updating the dataset to have nested entities in a possible future release of the corpus. On the other hand, this resulted in ambiguities concerning the extent of entity spans and types, especially with the introduction of tags that mark non-names like qualifiers, nationalities etc.

225972 tokens	14467 annotations
LOC	1294
MISC	1662
ORG	10529
PER	982

Table 2: The distribution of entity types in the original corpus

2.6 Properties of the Corpus

The size of the corpus is 206722 tokens, smaller than the original corpus due to the removal of repeated boilerplate content. On the other hand, it contains 40158 annotated spans, almost 2.8 times as much as in the original version. The distribution of entity types in the original Szeged NER corpus and the final version is shown in Tables 2 and 3. We conflated relational qualifiers with non-relational qualifiers of the same entity type. The MISC category in the final corpus covers only abbreviations annotated in the original corpus as MISC not referring to named entities. ORG-inf denotes entity mentions that refer to unique entities like names but use some informal reference instead of a name. This includes e.g. references to US departments, the FED and other governmental organizations which are referred to in Hungarian as ministries, offices etc., to central banks, stock exchanges etc. Note that there are more GPE entities than there were LOC entities in the original version due to adjectival forms also annotated and EU annotated as GPE rather than ORG. Many MEDIA entities were also ORG in the original corpus.

Entity types have an obviously skewed distribution with organizations, dates, money, nationalities and percentages (including ratios) dominating due to the genre of the corpus, while some tags are rather underrepresented. We plan to address this issue by adding text in other genres to balance the

206722 tokens	40158 annotations
AGE	7
CARDINAL	1264
DATE	9230
DUR	301
EVENT	58
FAC	182
FEAST	4
GPE	2608
LANGUAGE	5
LAW	36
LOC	381
MEDIA	1517
MISC	56
MONEY	2485
NORP	2325
ORDINAL	408
ORG	9060
ORG-inf	483
PER	985
PERCENT	1883
PRIZE	7
PROD	306
PROJ	29
QUAL_EVENT	12
QUAL_FAC	15
QUAL_GPE	12
QUAL_MEDIA	114
QUAL_ORG	1427
QUAL_PER	947
QUAL_PROD	165
QUAL_PROJ	5
QUAL_SEC	2
QUAL_STX	228
QUAL_URL	2
QUAL_WORK_OF_ART	15
QUANTITY	1599
SEC	345
STX	643
TIME	967
URL	2
WORK_OF_ART	38

Table 3: The distribution of entity types in the final corpus

corpus.

3 Models and Performance

We tested the zero-shot performance of the original OntoNotes-based models on the final corpus disregarding entity types not covered by the OntoNotes annotation. The FLAIR model achieves $F_1 = 75.20$ with a great proportion of the errors coming from erroneously included definite articles. Considering all tag types, the performance is $F_1 = 67.46$. A simple fix of the definite article problem boosts performance on common tags to $F_1 = 87.91$, and $F_1 = 80.63$ on the full tagset.

The DeepPavlov model trained on the same dataset fared much worse achieving only $F_1 = 58.26$ on common tags and $F_1 = 53$ on the full tagset.

We trained a vanilla neural sequence tagger using the HuggingFace Transformers library (Wolf et al., 2020) fine-tuning the monolingual Hungarian huBERT language model (Nemeskey, 2021) using a 9:1 train:test split of the corpus. It achieved $F_1 = 92.69$, performing significantly better than the zero-shot models.

4 Conclusion

In this paper, we presented the procedure we followed to enrich the annotation in a legacy Hungarian NER resource by applying NER models based on multilingual language models and fine-tuned on NER corpora in other languages. We then made a significant effort identifying errors and correcting the annotation using automatic and semi-automatic methods, providing a solid base for the final manual annotation correction. We trained a neural sequence tagger on the final corpus achieving a solid $F_1 = 92.69$ performance.

Acknowledgments

This research was implemented with support provided by grants FK 125217 and PD 125216 of the National Research, Development and Innovation Office of Hungary financed under the FK 17 and PD 17 funding schemes as well as through the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008).

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Leonard E. Baum and Ted Petrie. 1966. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554 – 1563.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology, June 29 - July 3, 2003, Brisbane, Australia*, pages 180–182.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Dávid Márk Nemeskey. 2021. Introducing `huBERT`. In *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14, Szeged.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named entities in Czech: annotating data and developing NE tagger. In *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 188–195, Berlin / Heidelberg. Springer.

Takashi Shibuya and Eduard Hovy. 2020. Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

Eszter Simon and Dávid Márk Nemeskey. 2012. Automatically generated NE tagged corpora for English and Hungarian. In *Proceedings of the 4th Named Entity Workshop (NEWS) 2012*, pages 38–46, Jeju, Korea. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

György Szarvas, Richárd Farkas, László Felföldi, András Kocsor, and János Csirik. 2006. A highly accurate named entity corpus for Hungarian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes Release 5.0.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.