Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1907–1916
Marseille, 20-25 June 2022

NerKor+Cars-OntoNotes++

Attila Novák, 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 an upgraded version of the Hungarian NYTK-NerKor named entity corpus, which contains about twice as many annotated spans and 7 times as many distinct entity types as the original version. We used an extended version of the OntoNotes 5 annotation scheme including time and numerical expressions. NerKor is the newest and biggest NER corpus for Hungarian containing diverse domains. We applied cross-lingual transfer of NER models trained for other languages based on multilingual contextual language models to preannotate the corpus. We corrected the annotation semi-automatically and manually. Zero-shot preannotation was very effective with about 0.82 F_1 score for the best model. We also added a 12000-token subcorpus on cars and other motor vehicles. We trained and release a transformer-based NER tagger for Hungarian using the annotation in the new corpus version, which provides similar performance to an identical model trained on the original version of the corpus.

Keywords: named entity recognition, cross-lingual transfer, annotated corpus, machine-generated annotation, multilingual contextual language models

1. Introduction

In this paper, we present a new version of the Hungarian NYTK-NerKor named entity corpus upgraded to an extended version of the OntoNotes 5 annotation scheme doubling the annotated spans and introducing a 7-fold increase in entity types.[1] We describe the annotation procedure applying cross-lingual transfer followed by semi-automatic and manual correction. This is followed by evaluation of the transfer models and comparison of models trained on the original[2] and the new version of the corpus[3].

1.1. Resources

Named entity recognition is a fundamental NLP task that plays 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 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), many legacy named entity corpora contain an annotation distinguishing four entity types: in addition to organizations (ORG), persons (PER), locations (LOC) there is a general entity category covering all the rest (MISC). This was the case for all named entity corpora available for Hungarian, the *Szeged NER corpus* (Szarvas et al., 2006), the Hungarian *Criminal NE corpus*[4], the silver-standard Hungarian *hunNERwiki* corpus (Simon and Nemeskey, 2012) automatically derived from Wikipedia, and even the most recent *NYTK-NerKor corpus* (Simon and Vadász, 2021) published in 2021.[5]

In contrast, some corpora containing named entity annotation, like the English *OntoNotes 5* corpus (Weischedel et al., 2013), contain a richer set of entities. The OntoNotes 5 annotation differentiates geopolitical entities (GPE: countries, settlements, etc.) and facilities (FAC: buildings, roads, airports etc.) from geographical locations like mountains or bodies of waters. Within the very generic MISC category, products (PROD), laws, directives and other legal 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 because English orthography happens to prescribe capitalization for words (in the case of NORP: adjectives) belonging to these categories. See also Table 3 on the tagset used in NerKor+Cars-OntoNotes++ for further details.

[1] The corpus is available at https://github.com/ppke-nlpg/NYTK-NerKor-Cars-OntoNotesPP with the same license as the original NerKor.

[2] The model trained on the original NerKor is available at https://huggingface.co/novakat/nerkor-hubert for comparison.

[3] The model trained on NerKor+Cars-OntoNotes++ is available at https://huggingface.co/novakat/nerkor-cars-onpp-hubert.

[4] https://rgai.inf.u-szeged.hu/node/130

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

There are more than a dozen named entity resources of English biomedical text covering entities completely different from those annotated in resources for generic text. In most of them, genes and proteins play a major role. Other entity types covered include names of species, diseases, cell lines, anatomic structures and chemical compounds. The most important milestone of biomedical named entity research was the creation of the *GENIA* corpus (Kim et al., 2003) covering 47 different entity types.

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 get special annotation. The same annotation scheme was applied when creating the Danish *DaN+* corpus (Plank et al., 2020). In addition, these corpora, similarly to other resources like GENIA, the Spanish and Catalan newspaper text corpus *AnCora* (Taulé et al., 2008) and one of the most richly annotated generic NER corpora, the *Czech Named Entity Corpus (CNEC 2)* (Ševčíková et al., 2007), feature nested named entities.

The *ACE task* datasets of LDC for English, Chinese and Arabic[6] also have nested entities (and also a relatively fine-grained entity subtype taxonomy) with not only proper names but also NP's headed by common nouns annotated with entity types and relations also marked. Unfortunately, these resources are not freely available. The same applies to the *NNE (Nested Named Entities)* dataset[7] (Ringland et al., 2019), which is based on Penn Treebank data.

The resources mentioned before involve a (word)-token-level annotation. A fairly recent resource for Hebrew, *NEMO* (Bareket and Tsarfaty, 2021), (although relatively small at 93,500 tokens) features not only a moderately detailed entity classification (the genuine NE tags of the OntoNotes tagset), but also nested entities and (what is the real novelty) subtoken-(morpheme)-level entity annotation, which again may turn out to be a breakthrough innovation for morphologically complex languages.

There are dozens of NER datasets in other languages, which we cannot review here. An overview of the datasets mentioned above is shown in Table 1.

2. The NerKor Corpus

NYTK-NerKor (NerKor) (Simon and Vadász, 2021), with a size of over one million tokens, is much bigger than any of the previous gold standard Hungarian named entity corpora. Moreover, in contrast to its predecessors, which covered only a single domain, NerKor has a broad coverage of domains and topics, and it is thus an important milestone in Hungarian NLP research. Another property that distinguishes it from some earlier Hungarian NER corpora is that it has a permissive licence.

NerKor consists of five 200,000-token subcorpora. The *fiction subcorpus* contains literary work from the beginning of the 20th century, whose copyright has already expired. These texts deviate from present-day orthography and language use quite significantly. The other part of this subcorpus contains movie subtitles from the Opus Opensubtitles[8] Corpus[9]. The *legal subcorpus* consists of sentences taken from EU sources: parts of the 2004 EU constitution, European Economic and Social Committee documents, and portions of JRC-Acquis and DGT-Acquis, all downloaded from the Opus corpus website. A conspicuous feature of these texts is the high prevalence of references to laws, regulations, standards and some difficult-to-categorize EU entities like specific sections of budget. The *web subcorpus* contains content downloaded from the Hungarian portion of Common Crawl. One source of content in the *news subcorpus* is the Hungarian edition of Global Voices[10] containing articles translated by non-professional translators. A hallmark of these texts is the prevalence of references to social media sites and content. The rest of this subcorpus comes from the NewsCrawl 2019 corpus[11] created by the team organizing the 2019 WMT machine translation conference.[12] Finally, the *wiki subcorpus* is a subset of the Hungarian part of the hunNERwiki corpus (Simon and Nemeskey, 2012), which contained a selection of sentences from a dump of Hungarian Wikipedia that contained at least one internal Wikipedia link to a Wikipedia entry that belongs to a named entity type according to DBpedia. The automatically generated silver-standard annotation from HunNERwiki was turned into gold standard annotation during the creation of NerKor.

About one third of the corpus (the fiction subcorpus, Global Voices and a small portion of the Wikipedia content) contains coherent text, the rest is a shuffled collection of unrelated sentences or sentence fragments. In this respect, the corpus differs from earlier Hungarian named entity corpora, which consist predominantly of coherent text. In addition, about half of the corpus contains standard proof-read text that conforms orthographic norms. Unfortunately, this is the part which consists mostly of shuffled sentences.

Shuffling sentences has been a method applied to

[6] https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

[7] https://github.com/nickyringland/nested_named_entities

[8] http://www.opensubtitles.org/

[9] https://opus.nlpl.eu/OpenSubtitles-v2018.php,

[10] https://hu.globalvoices.org/

[11] http://data.statmt.org/news-crawl/

[12] http://www.statmt.org/wmt19/

corpus	granul.	non-ent	der/cmpd	nest	cmn N	subtok	language	domain	gold
CONLL 2002/2003	-						en, ne, es	news	+
Szeged NER	-						hu	business	+
Criminal NER	-						hu	news	+
hunNERwiki	-						hu	wiki	-
NYTK-NerKor	-						hu	multi	+
OntoNotes 5	+	+					en	multi	+
Genia	-			+			en	biomed	+
NoSta-D	-		+	+			de	wiki+news	+
DaN+	-		+				dk	multi	+
AnCora	-	+		+			es, ca	news	+
CNEC 2	++	+		+			cz	?	+
ACE datasets	++	+		+	+		en, ar, zh, es	news	+
NNE	+++			+			en	news	+
NEMO	+		+	+		+	he	news	+

Table 1: Properties of NER datasets mentioned in Section 1.1

legacy corpora to avoid copyright problems.[13] Not very long ago, when most NLP models did not try to handle text-level dependencies, shuffling was not too high a price to pay for avoiding legal affairs in order to ensure that corpora can be freely used.

The task of named entity recognition per se, is not in general considered to require a context wider than a sentence. However, Schweter and Akbik (2020) report improvement when incorporating some extrasentential context in their models. Omitting parts of the corpus not containing relevant entities can also be justified by reducing the workload on annotators (we also applied this method when constructing the *car subcorpus,* see below), although it may significantly affect named entity density and thus possibly also the performance of models trained on the resource. E.g. NE density is 1.7–3.75 times higher in the filtered wiki subcorpus than in other subcorpora of NerKor.

Moreover, considering the effort put into the annotation process, it would be desirable that the corpus could be further annotated for other tasks for which a wider context might be necessary (e.g. co-reference resolution). One third of the NerKor corpus can be used for such purposes: it is indeed a positive aspect of the corpus that some parts of these texts originate from the already coreference-annotated KorKor(pusz) pilot corpus (of 31492 tokens) (Vadász, 2020).

20% of the corpus includes morphological annotation as well, so that traditional machine learning algorithms can also be trained on these portions. In the present adaptation of the corpus, we ignored these morphological annotations.

3. Annotation Method

Our reannotation workflow for NerKor followed in most aspects the method described in Novák and Novák (2021). We relied on zero-shot application of transformer-based named entity recognition models trained on resources in other languages: English and Czech, algorithmic merging with the original annotation, and semi-automatic and manual correction.

3.1. Zero-shot Preannotation

First we applied two models trained on the English OntoNotes 5 corpus to the Hungarian corpus. The first model was created by the DeepPavlov team (Burtsev et al., 2018) fine-tuning multilingual BERT (Devlin et al., 2019). The other model is based on XLM-RoBERTa (Conneau et al., 2019), 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 perform different tokenization following the tokenization scheme of the underlying contextual language model: XLM-RoBERTa is based on a SentencePiece tokenizer (Kudo and Richardson, 2018), while the BERT model applies legacy tokenization first, which is followed by WordPiece subword tokenization. Thus the difference between the token sequence in the output of the models and the original input token sequence had to be taken into account when merging the annotation from the models with the original annotation. The merging algorithm considered the spans in the input annotations gold standard in the case of overlapping entity spans, and if the generated annotation contains a compatible entity subtype, the entity type is updated accordingly. 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 was introduced based on the output of the models.

3.2. Error Analysis and Automatic Error Correction

Zero-shot model application resulted in typical errors. E.g., in the case of transfer from English to Hungar-

[13] Note, however, that copyright issues do not arise concerning most parts of the corpus containing non-coherent text: the legal, the web or the Wikipedia subcorpus.

ian, a typical problem is that for some named entity types, like names of organizations, journals, titles of works of art etc., a definite article is present in Hungarian when the name is incorporated in the sentence structure (but not in parentheticals), while there is no article in English. This made the models include definite articles in the span for these entity types, an error that could be easily eliminated from the output using regular-expression-based patterns. Similar automatic correction patterns were applied to fix certain types of anomalies concerning numerical and quantity expressions.

3.3. Benefits and dilemmas arising from cross-lingual mapping

While cross-lingual mapping resulted in some anomalies like inclusion of definite articles, it had other side-effects that we found useful. E.g., since English prepositional phrases of names (which are obviously annotated as named entities) often correspond to adjectives derived from the given name in Hungarian, the output of the models also included entity annotation for these adjectives. In contrast to the German NoSta-D or the Danish DaN+ corpus, words like this remained unannotated in all legacy Hungarian named entity corpora. However, identifying these words as references to named entities is desirable in practical applications like information retrieval or data de-identification. We thus decided to keep this kind of annotation as part of our annotation enrichment effort.

This decision, however, also raises dilemmas not complicating one's life if one refrains from annotating derived adjectives. While in the original OntoNotes corpus, adjectives labeled as NORP, such as *Chinese* can be easily distinguished from cases where elements with similar meanings appear as possessive or prepositional constructions like *of/to/from China* (a GPE or ORG[14] annotation is used in these cases), their Hungarian equivalent is often neutralized, and this may lead to ambiguities that are difficult to resolve. In some cases the equivalents differ, like for *English* vs. *of/to England: angol* vs. *angliai,* and it is relatively clear in such cases that the former should be labeled as NORP and the latter as GPE. However, in most analog cases there is no lexical difference, e.g. *Chinese* and *of/to/from China* both neutralize to *kínai.* In these cases, the guideline for manual disambiguation or annotation was to use the *angol* vs. *angliai* duality as an analogy, but it is easy to make mistakes when making these annotation decisions.

When extending annotation to non-name entities, such as time expressions and quantities, the exact range of terms to be annotated is also relatively difficult to define coherently and to consistently adhere to. In addition, the emergence of new entity types leads to a proliferation of elements increasing the need to annotate nested entities, which we have so far refrained from.

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.

In the annotation generated by the OntoNotes models, quantities form a separate class: these are composed of a number and a unit of measure. However, expressions of time durations, such as *két napra* 'for two days', *hároméves* 'three-year-long/three-year-old' were annotated by the models as dates. We have checked these expressions manually and transformed them either to spans of type *time duration* or *age*.

3.4. NameTag 2

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). This model is based on a fine-grained hierarchy of entity classes having many subclasses within the broader categories like a distinction of companies vs. governmental/political institutions vs. academic/educational/cultural/sports 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. Although this fine-grained model seemed attractive, at least in the zero-shot cross-lingual setting, the annotation generated by this model turned out to be much less accurate than those generated by OntoNotes-based models. This may not only be due to the higher number of distinct classes and the more complicated algorithm but also the limited size of the CNEC 2 training data (consisting of about 200 thousand tokens vs. the 1.5 million tokens in OntoNotes 5).

Since there is no definite article in Czech, we expected this model to have a problem with definite articles for the entity types the English models struggled with, but it turned out to have this problem only in the case of sentence-initial capitalized definite articles (probably due to a constraint on capitalization that might be included in the algorithm). A more prevalent problem with this model was that it often assigned different classes to different occurrences of the same entity (and usually this was an error rather than real ambiguity) and often left the same entity unannotated. Identification of the span of the entities was also less accurate than what the English-based models generated.

3.5. A lemmatized named entity list and automated correction patterns

In spite of its weaker performance, annotation generated by the Czech model proved to be useful. We lemmatized the entity annotations generated by all models (only the head, i.e. the last token of names were lemmatized) and created a list of all the lemmatized named entities occurring in the corpus grouped by named entity types, resulting in a gazetteer-like resource. This

[14] e.g. for affiliations with political parties

list contained all alternative analyses for each entity along with their corpus frequencies. Reviewing these lists, we have found that the Czech model frequently fails to assign the correct classification, thus we did not adopt the taxonomy of CNEC2.

Nevertheless, using the automatically generated named entity list, we were able to identify elements frequently misclassified in the OntoNotes model, and entities that should be assigned to distinct classes, such as journals and social media sites. These were assigned the MISC annotation in the original NerKor annotation when they denoted the newspaper or the social media site itself, and ORG if they denoted the corresponding company, publisher, or editorial board, i.e. in NerKor, the annotation generally follows the tag-for-meaning principle. The OntoNotes models, however, assigned an ORG annotation to all of these entities, which is not an optimal solution as the two entities are not identical. We assigned new entity types (MEDIA and SMEDIA) to these MISC types, inspired by the solution used in the CNEC corpus. We identified other entity subtypes, like awards and projects/programs in a similar fashion. Elements to be reclassified were marked manually in the list, and we automatically generated regular-expression-based correction scripts from these that automatically corrected the annotation of all inflected forms of the listed entities in the corpus. Using these patterns, we bulk-corrected the annotation.

3.6. Manual error correction

In the whole corpus, we have performed manual error correction starting from three points. First, we have reviewed and corrected anomalies in the lemmatized named entity list, as described in the previous section, applying mainly automatic patterns, but also checking manually. Second, we have resolved contradictions of the annotations in the original corpus and those generated by the transfer models (and also the differences of the annotations generated by the transfer models) by manually correcting each case. Third, we have noticed that the generated annotation for references to legislation in the law subcorpus (e.g. *[1260/2001/EK rendelet] [1. cikke (1) bekezdésének a) pontjában]* ‘in point (a) of Article 1(1) of Regulation (EC) No 1260/2001.’), was often fragmented and produced various annotation patterns. We normalized these in the form that the name of the law was considered a single entity, and the exact location/document part within the legal reference was another one following it. We also corrected any errors found around any corrections and if a recurrent error type was discovered, we also looked them up and corrected these in the corpus. Systematic correction of the whole corpus is in progress, we have finished about 25%, including the test set.

3.7. Metonymic language use

In the original NerKor corpus, metonymic uses of names were assigned the tag corresponding to the meaning of the actual occurrence of the entity (tag-for-meaning annotation). This was the case not only for the journal/publisher ambiguity, but also when for example country names are used as agents in the text. In these cases, they were annotated as organization instead of location. This type of metonymy is completely productive for every geopolitical entity, and we think that annotating all of them uniformly as geopolitical entities can handle this frequent type of metonymic language use.[15] We have annotated country names as simple organizations only if they referred to national teams (a case of so called ‘nickname metonymy’), similarly to the annotation of other sport teams. The use of countries as agents, in our opinion, significantly differs from the phenomena of using the same name for a company and its product or other less productive (nickname metonymic) patterns, like using a street name *(Wall Street)* to denote an organization (the New York Stock Exchange) or a city name (*Brussels*) to denote a much more extensive geopolitical entity (the EU).[16] In these cases the distinction is clearly justified, since they are different quite loosely related entities.

In the original corpus, predicative occurrences of names, like *her name was X*, were annotated as MISC. In these cases, the first step was to add a label to the MISC tag that indicates the real class of the name, such as MISC-PER for a person’s name. For practical purposes, however, these cases can also be considered as references to the given person/organization, so we feel it is appropriate to consider them simply as a person/organization mention.

In Table 2 presenting the tag distribution of the corpus and for the evaluation we applied the latter strategy, but in the published version of the corpus we did not simplify these tags, so that the information coded in the original annotation can also be retrieved. The case of author-work metonymy (e.g. *Chomskyt olvas* ‘he is reading Chomsky’) was handled similarly in the original NerKor annotation. However, while it is evident that it is not the author that is read, but his work, for practical purposes these can be considered mentions of the author.

As we have already mentioned, we introduced new types for the annotation of media, social media, project and award entities (subtypes of the umbrella MISC category). There are, however, further borderline entities that fall into the vague region of the organization-media-‘work-of-art’-product continuum. The annota-

[15] The guidelines to the annotations of LDC ACE resources discuss in detail metonymies of GPE’s, and while some metonymy types resembling ORG, LOC or PER entities are identified, many other uses involve more than one aspect of these complex entities. Poibeau (2006) also shows that the distinction of PER vs. ORG metonymies of geopolitical entities is problematic even for humans.

[16] GPE subtypes (state, province, city, etc.) would be needed to distinguish cases of the latter type from simple city mentions.

PER[a]	15266	PER[a]	15234	LOC	2226	FAC	831	PROJ	254
LOC	12988	GPE[b]	13872	WORK_OF_ART	1975	MONEY	681	MISC	117
ORG[a]	12343	DATE	11224	QUANTITY	1918	TIME	661	ID	83
MISC	5751	ORG[a]	9512	CAR	1423	EVENT	627	AWARD	64
		CARDINAL	6710	DUR	1395	LANGUAGE	499		
NYTK-NerKor		NORP	4551	PERCENT	1257	AGE	336	**NerKor+Cars**	
tokens	**entities**	ORDINAL	3258	PROD	1174	MISC-ORG	306	**tokens**	**entities**
1027218	**46348**	LAW	3245	MEDIA	1062	SMEDIA	271	**1038947**	**84831**

[a] We manually corrected erroneous annotation of names of bands from PER to ORG.

[b] A subset of items originally tagged ORG (names of countries) is GPE (GPE-ORG) in the new version, many adjectives derived from names are also of type GPE.

Table 2: Size and label distribution of the original and new corpus versions

tion of these remained MISC-ORG for now as a combination of the original and the OntoNotes-based annotation if the entity type was institution-like, and it remained MISC if it did not fit into any of the above mentioned categories.

4. The *cars* subcorpus

We have performed an experiment on how to introduce a new subtype for an existing type. As the experimental class, we chose vehicles within the *product* class. For training, we selected articles from the archive of the hvg.hu news site using motor-vehicle-related keywords. Then, we chose sentences from this collection that contained car makes and models that were present in the menu structure of a car dealer's website. This method resulted in a 12000-token corpus, which we annotated using the Flair OntoNotes model. We then manually corrected the annotation and replaced *product* tags by *car* tags for car names. We added this special subcorpus as a training/dev/test set for motor vehicles.

5. Features of the corpus

Size and label distribution of the original NerKor corpus and the newly created version is shown in Table 2. The number of distinguished entity types increased 7-fold while the number of entities marked almost doubled from the original. The annotation also features many types not present in the OntoNotes tag set either, e.g. DUR (time duration), AGE, MEDIA (journals, tv stations and news portals) and SMEDIA (social media), PROJ (projects and programs), AWARD and ID. The tags used for entity types in the corpus are described in Table 3.

6. Possible future enhancements

While corpora containing nested entities have existed for some languages and domains, most resources contained only unnested annotation for the lack of competitive nested taggers. Recently, the development of models that can also handle nested entities has gained momentum, and some open source neural nested entity taggers have emerged (Wang et al., 2020; Shibuya and Hovy, 2020; Shen et al., 2021; Tan et al., 2021). However, 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), nevertheless updating the dataset to have nested entities is a possible future enhancement of the corpus. In many cases it would be easier to make annotation decisions if we allowed nested entities, and such an annotation would also be more motivated for information retrieval applications.

7. Models and performance

We evaluated the zero-shot performance of the transfer-based models: the OntoNotes5-based Flair and DeepPavlov models and the Czech NameTag2 tagger on the test set of the corpus. We also performed the evaluation with the tagset normalized to the tags present in the original model (ignoring distinctions the model was not trained to make). We also trained a neural tagger model based on the Hungarian huBERT contextual language model (Nemeskey, 2021) on the training set of the corpus using the HuggingFace Transformers library (Wolf et al., 2020) with an improved Viterbi-like decoding that eliminates invalid tag sequences from the output (Nemeskey, 2020). The performance of these models is shown in Table 4. We report P, R and F_1 scores as percentage.

Models using language transfer performed quite well, but among the English models trained on the same corpus, the XLM-RoBERTa-based Flair model performed significantly (about 10% F-measure) better. The Flair model using a "stronger" language model obtained higher precision and recall values across the board for all named entity types, than the weaker model. The performance of these models increased (by 5-6% F-measure) when the automatic regular-expression-based correction of definite articles was applied to their output. The zero-shot performance of the Flair model on entity types in common with those in the final version (i.e. ignoring the newly introduced MEDIA, SMEDIA, PROJ, etc. tags) is quite convincing. This performance made our re-annotation effort feasible.

The apparently quite weak performance of the Czech model is partly explained by the fact that it works with a much more fine-grained tagset, thus in order to mea-

non-entities	*DATE*	dates and intervals (granularity over 24 hours)
	CARDINAL	cardinal numbers
	NORP	nationalities, religion, political affiliation (adjectives)
	ORDINAL	ordinal numbers
	LAW	references to laws, directives and other norms
	QUANTITY	quantites: cardinal number + unit of measure
	DUR	time durations (time quantites, unanchored to the timeline)
	PERCENT	percentages and ratios (in OntoNotes: only percentages)
	TIME	time and short intervals (granularity below 24 hours)
	LANGUAGE	names of languages
	AGE	age of persons and things (time durations with spec. semantics)
	MONEY	sums of money: cardinal number + monetary unit
organizations	ORG	organizations: companies, parties, institutions, teams etc.
persons	PER	people, fictive persons, families, animals
places	GPE	geopolitical entities: states, settlements, provinces, counties etc.
	LOC	geological locations: mountains, deserts, bodies of water etc.
	FAC	facilities: roads, streets, buildings etc.
other entities	WORK_OF_ART	titles of creative works
	PROD	products (except motor vehicles)
	MEDIA	journals, tv channels, news sites
	CAR	motor vehicles
	SMEDIA	social media
	EVENT	named events (except projects)
	PROJ	projects and programmes
	AWARD	awards
	MISC-ORG	organization-like types of residual entities
	MISC	residual entities

Table 3: Description of tags/entity types used in the corpus. Types in **bold** are not present in the OntoNotes 5 tagset.

version	original			Det fixed			only labels in common			com. labels, Det fixed		
model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
CZ	15.82	11.39	13.25	15.89	11.44	13.30	64.57	52.92	58.16	64.63	52.97	58.22
DP	66.32	60.41	63.23	71.66	65.27	68.31	68.79	63.42	65.99	74.63	68.81	71.60
FL	74.81	70.73	72.71	80.59	76.19	78.33	77.68	74.34	75.97	**83.90**	**80.29**	**82.06**
NKC	**91.07**	**88.12**	**89.57**				**91.64**	**89.18**	**90.39**			
test	91.92	87.65	89.73									

Table 4: Performance of models on the test set, CZ: Czech model NameTag2, DP: DeepPavlov OntoNotes/m-BERT, FL: Flair-OntoNotes-Large/XLM-RoBERTa, NKC: NerKor+Cars/huBERT, test: precision of the test set before manual correction.

sure its performance, the normalization of tags was unavoidable. Its performance, however, lags far behind the other models after normalization, too. The Czech training set is much smaller than that of the other models, and the more complex algorithm allowing embedded entities might also play a role in its weaker performance.

Nevertheless, the models based on the Hungarian language model trained on the corpus performed significantly better on the complete final tagset than the best transfer-based model considering only the common tags, so creating the corpus 'made sense'.

7.1. Comparison with performance on the original NerKor annotation

We also compared the performance of the best tagger model with that of the same algorithm trained on the original NerKor annotation to see how the division of some entity classes (especially MISC) into several subclasses impacts performance. For the sake of comparability, we partitioned entity types into non-entities, which are not part of the original NerKor annotation (numerical and time expressions, language names, NORP adjectives and law references), and named entities. The results are shown in Table 5 : non-entities at the top half (ordered by decreasing frequency in the test set), entities at the bottom half, with aggregate scores on named entities at the bottom row.

The huBERT-based model with the Viterbi-based decoding performed similarly on named entities to a similar model (emBERT) (Simon et al., 2022) trained on the original version of the corpus.

The F-score on locations is lower than in the case of the model trained on the original corpus partly due to missed adjectival GPE entities (not present in the origi-

NerKor	F_1	NerKor+Cars	F_1
		DATE	**88.85**
		CARDINAL	**83.78**
		NORP	**87.12**
		ORDINAL	**94.67**
		LAW	**82.12**
		QUANTITY	**91.11**
		DUR	**74.67**
		PERCENT	**84.21**
		TIME	**66.67**
		LANGUAGE	**83.33**
		AGE	**100.00**
		MONEY	**87.50**
ORG	88.45	ORG	**93.33**
PER	95.32	PER	**97.11**
LOC	**92.28**	GPE	91.98
		LOC	76.60
		FAC	80.00
MISC	81.85	WORK_OF_ART	**90.27**
		PROD	79.37
		MEDIA	**91.53**
		CAR	**92.86**
		SMEDIA	73.33
		EVENT	72.73
		MISC-ORG	47.06
		PROJ	66.67
		AWARD	**100.00**
		MISC	66.67
	91.02		89.57/**92.05**

Table 5: Performance of the best model trained on NerKor+Cars on each entity type compared to performance a similar model on the original NerKor annotation. Tags ordered within the categories with descending frequency in the test set top to bottom.

nal annotation), massive ambiguity of *Europe* as a continent or a reference to the EU (which was not consistently marked in the not yet checked portion of the training corpus), ORG vs. FAC ambiguity of institutions (universities) and obscure place names (GPE vs. LOC ambiguity). Most confusion is within subtypes of locations.

Performance on frequent and easier-to-distinguish subtypes of MISC (WORK OF ART, MEDIA and CAR) is better than on the generic MISC category, while for rare and difficult-to-categorize entities (as well as for products) we got worse-than-average performance. Nevertheless, the division of the MISC class to several subclasses (even with mainly automatic methods) did not result in a substantial drop in the performance of the system, the aggregate F_1 score for named entities turned out to be even better than for the same type of model trained on the original four-class annotation (92.05 vs. 91.02).

8. Conclusion

In this paper, we have presented our research concerning the automatic enhancement of the annotation of the large NYTK-NerKor named entity corpus containing Hungarian texts of various genres. We have almost doubled the number of annotated elements in the corpus and made the number of distinguished classes 7 times bigger. For this, we applied cross-lingual transfer, which proved to be efficient according to our evaluation, but the performance of various transfer models showed significant differences. We have corrected the annotation using semi-automatic methods. 25% of the corpus has been manually checked, including the test set. Further manual checking of the corpus is in progress. The test set was used for the evaluation of transfer models and a monolingual huBERT-based model trained on the training set of the corpus. The latter had a performance similar to that measured on the original corpus in spite of the much more detailed entity classification in the new version.

9. Acknowledgements

The research presented in this paper 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.

10. Bibliographical References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (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, June. Association for Computational Linguistics.

Bareket, D. and Tsarfaty, R. (2021). Neural modeling for named entities and morphology (NEMO2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Benikova, D., Biemann, Ch., and Reznicek, M. (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, May. European Language Resources Association (ELRA).

Burtsev, M., Seliverstov, A., Airapetyan, R., Arkhipov, M., Baymurzina, D., Bushkov, N., Gureenkova, O., Khakhulin, T., Kuratov, Y., Kuznetsov, D., Litinsky, A., Logacheva, V., Lymar, A., Malykh, V., Petrov, M., Polulyakh, V., Pugachev, L., Sorokin, A., Vikhreva, M., and Zaynutdinov, M. (2018). DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Kim, J., Ohta, T., Tateisi, Y., and Tsujii, J. (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.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Nemeskey, D. M. (2020). Egy `emBERT` próbáló feladat [A task testing emBERT]. In *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2020) [16th Hungarian Conference on Computational Linguistics]*, pages 409–418, Szeged.

Nemeskey, D. M. (2021). Introducing `huBERT`. In Gábor Berend, et al., editors, *XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021)*, pages 3–14, Szeged. Szegedi Tudományegyetem, Informatikai Tanszékcsoport.

Novák, A. and Novák, B. (2021). Transfer-based enrichment of a Hungarian named entity dataset. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2021*, pages 1060–1067.

Plank, B., Jensen, K. N., and van der Goot, R. (2020). DaN+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Poibeau, T. (2006). Dealing with metonymic readings of named entities. *CoRR*, abs/cs/0607052.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., and Curran, J. R. (2019). NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy, July. Association for Computational Linguistics.

Schweter, S. and Akbik, A. (2020). FLERT: document-level features for named entity recognition. *CoRR*, abs/2011.06993.

Ševčíková, M., Žabokrtský, Z., and Krůza, O. (2007). Named entities in Czech: annotating data and developing NE tagger. In Václav Matoušek et al., editors, *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.

Shen, Y., Ma, X., Tan, Z., Zhang, S., Wang, W., and Lu, W. (2021). Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online, August. Association for Computational Linguistics.

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

Simon, E. and Nemeskey, D. M. (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, July. Association for Computational Linguistics.

Simon, E. and Vadász, N. (2021). Introducing NYTK-NerKor, a gold standard Hungarian named entity annotated corpus. In Kamil Ekstein, et al., editors, *Text, Speech, and Dialogue - 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6-9, 2021, Proceedings*, volume 12848 of *Lecture Notes in Computer Science*, pages 222–234. Springer.

Simon, E., Vadász, N., Lévai, D., Nemeskey, D., Orosz, Gy., and Szántó, Zs. (2022). Az NYTK-NerKor több szempontú kiértékelése [A multi-faceted evaluation of NYTK-NerKor]. In Gábor Berend, et al., editors, *XVIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2022)*, Szeged. Szegedi Tudományegyetem, TTIK, Informatikai Intézet.

Straková, J., Straka, M., and Hajič, J. (2019). Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Stroudsburg, PA, USA. Association for Computational Linguistics.

Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., and Csirik, J. (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, May. European Language Resources Association (ELRA).

Tan, Z., Shen, Y., Zhang, S., Lu, W., and Zhuang, Y. (2021). A sequence-to-set network for nested named entity recognition. In *Proceedings of the 30th Inter-*

national Joint Conference on Artificial Intelligence, IJCAI-21.

Taulé, M., Martí, M. A., and Recasens, M. (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, May. European Language Resources Association (ELRA).

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.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Vadász, N. (2020). KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése [Building KorKorpusz, a manually annotated multi-layer pilot corpus]. In Gábor Berend, et al., editors, *XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020)*, pages 141–154, Szeged. Szegedi Tudományegyetem, TTIK, Informatikai Intézet.

Wang, J., Shou, L., Chen, K., and Chen, G. (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, July. Association for Computational Linguistics.

Weischedel, Ralph and Palmer, Martha and Marcus, Mitchell and Hovy, Eduard and Pradhan, Sameer and Ramshaw, Lance and Xue, Nianwen and Taylor, Ann and Kaufman, Jeff and Franchini, Michelle and El-Bachouti, Mohammed and Belvin, Robert and Houston, Ann. (2013). *OntoNotes Release 5.0*. Linguistic Data Consortium LDC2013T19, ISLRN 151-738-649-048-2.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (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, October. Association for Computational Linguistics.