Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4547–4556
Marseille, 11–16 May 2020

NorNE: Annotating Named Entities for Norwegian

Fredrik Jørgensen,[†] Tobias Aasmoe,[‡] Anne-Stine Ruud Husevåg,[◇] Lilja Øvrelid,[‡] Erik Velldal[‡]
Schibsted Media Group,[†] Oslo Metropolitan University,[◇] University of Oslo[‡]
fredrik.jorgensen@schibsted.com,[†] annesh@oslomet.no,[◇]
{tobiaaa,liljao,erikve}@ifi.uio.no[‡]

Abstract

This paper presents NorNE, a manually annotated corpus of named entities which extends the annotation of the existing Norwegian Dependency Treebank. Comprising both of the official standards of written Norwegian (Bokmål and Nynorsk), the corpus contains around 600,000 tokens and annotates a rich set of entity types including persons, organizations, locations, geo-political entities, products, and events, in addition to a class corresponding to nominals derived from names. We here present details on the annotation effort, guidelines, inter-annotator agreement and an experimental analysis of the corpus using a neural sequence labeling architecture.

Keywords: Named Entity Recognition, corpus, annotation, neural sequence labeling

1. Introduction

This paper documents the efforts of creating the first publicly available dataset for named entity recognition (NER) for Norwegian, dubbed NorNE.[1] The dataset adds named entity annotations on top of the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014), containing manually annotated syntactic and morphological information for both varieties of written Norwegian – Bokmål and Nynorsk – comprising roughly 300,000 tokens of each. The corpus contains mostly news texts (around 85% of the corpus), but also other types of texts, such as government reports, parliament transcripts and blogs. The treebank has following its release also been converted to the Universal Dependencies standard (Øvrelid and Hohle, 2016; Velldal et al., 2017). We correspondingly distribute the annotations of NorNE in two versions, mirroring both the original NDT and the UD-version.

The annotations in NorNE include a rich set of entity types. In short, they comprise the following (more details are given in the guideline discussion in Section 3):

- **Person** (PER): Named real or fictional characters.
- **Organization** (ORG): Any collection of people, such as firms, music groups, political parties etc.
- **Location** (LOC): Geographical places and facilities.
- **Geo-political entity** (GPE): Geographical regions defined by political and/or social groups, additionally sub-categorized as either:
    - GPE with a locative sense (GPE_LOC),
    - GPE with an organization sense (GPE_ORG).
- **Product** (PROD): Artificially produced entities, including abstract entities such as radio shows, programming languages, ideas, etc.
- **Event** (EVT): Festivals, weather phenomena, etc.
- **Derived** (DRV): Nominals that are derived from a name, but not a named entity in themselves.

In addition to discussing the annotation process and guidelines, we also provide an exploratory analysis of the resulting dataset through a series of experiments using state-of-the-art neural architectures for named entity recognition (combining a character-level CNN and a word-level BiLSTM, feeding into a CRF inference layer). The purpose of the experimental section is to provide some preliminary baseline results while simultaneously validating the consistency and usability of the annotations, and finally also shedding some light on the consequences of different design choices that can be made in the modeling stage, like the choice of which label set to use or which label encoding to use and so on.

The remainder of the paper is organized as follows. First, Section 2 briefly outlines previous work on NER for Norwegian. Section 3 then describes the NorNE annotation effort in more detail, briefly outlining the annotation guidelines while providing illustrating examples and presenting an analysis of inter-annotator agreement. In Section 4 we summarize the resulting dataset before we in Section 5 turn to an empirically-driven analysis of the data through a series of experiments.

2. Previous Work on NER for Norwegian

While NorNE is the first publicly available dataset for NER in Norwegian, there have been some previous efforts to create similar resources.

Most notably, in a parallel effort, and as part of his doctoral work, Johansen (2019) added named entity annotations to the same underlying corpus, the UD-version of NDT. However, the data was single-annotated and using a reduced label-set with only 4 different categories – locations, organizations, persons, and miscellaneous – and only considering text spans already tagged as PROPN (proper noun) in the original treebank data.

Previously, the Nomen Nescio project (Johannessen et al., 2005), focusing on Scandinavian languages, created a named entity annotated corpus for Norwegian. However, due to copyright restrictions the data was unfortunately not made publicly available. The Nomen Nescio corpus was based on articles from several news papers and

[1] https://github.com/ltgoslo/norne/

magazines, in addition to some works of fiction. It totaled 226,984 tokens, of which 7,590 were part of a named entity annotation, and the following six entity categories were used: person, organization, location, event, work of art, and miscellaneous. For a while, Norwegian named entity recognition based on these categories was supported in the Oslo–Bergen tagger (Johannessen et al., 2012), a rule-based morpho-syntactic tagger and lemmatizer for Norwegian based on constraint grammar, but unfortunately NER is no longer supported in the tagger. For more information on the Nomen Nescio data, including experimental results, see Nøklestad (2009).

The TORCH (Transforming the Organization and Retrieval of Cultural Heritage) project (Hoff and Preminger, 2015; Tallerås et al., 2014) focused on automatic metadata generation to improve access to digitized cultural expressions. As part of this effort, a corpus of subtitles and metadata descriptions from the archive of NRK (the national public broadcaster), was annotated with named entities. To facilitate comparison to previous work on Norwegian NER, the project adopted the categories used in Nomen Nescio, also extending the original annotation guidelines (Jónsdottír, 2003), as documented in (Howard, 2014). However, as some of the data contain confidential information and was only made available to the researchers on the basis of a non-disclosure agreement, the resulting dataset could unfortunately not be made publicly available.

3. Annotation

In this section we discuss in more detail various aspects related to the annotations; we first describe some relevant properties of named entities in Norwegian, and go on to flesh out the entity types in more detail, we highlight important parts of the guidelines, discuss issues that are particular to mentions of named entities in Norwegian and finally present an analysis of inter-annotator agreement.

3.1. Proper Nouns in Norwegian

In Norwegian, proper nouns are generally capitalized. For multi-token names, however, the general rule is that only the first token should be capitalized (Faarlund et al., 1997), e.g. *Oslo rådhus* 'Oslo city hall'. There are a number of exceptions to this general rule as well, for instance in company names, e.g. *Den Norske Bank* 'The Norwegian Bank'. Proper nouns may also be nested to form a larger unit, e.g. *Universitetet i Oslo* 'University of Oslo'. Unlike in English, names for days, months and holidays are not capitalized, e.g. *onsdag* 'Wednesday', *juli* 'July'. Like most Germanic languages, compounding is highly productive in Norwegian and compounds are written as one token, e.g. *designbutikk* 'designer store'. However, when the initial part of a compound is a proper name, these are separated by a hyphen, e.g. *Prada-butikken* 'the Prada store'.

3.2. Entity Types in NorNE

In the NorNE corpus we annotate the six main types of named entities PER, ORG, LOC, GPE, PROD and EVENT, as mentioned in Section 1 above (in addition to a special category for nominals *derived* from proper names). In the following we will present the different categories and the main principles behind their annotation. We will comment on how various aspects of the annotation relates to other well-known datasets, but also discuss entity properties that are specific to Norwegian.

Person (PER) The person name category includes names of real people and fictional characters. Names of other animate beings such as animals are also annotated as PER, e.g., *Lassie* in *den kjente TV-hunden Lassie* 'the famous TV-dog Lassie'. Family names should be annotated as PER even though they refer to several people.

Organization (ORG) This entity category includes any named group of people, such as firms, institutions, organizations, pop groups, political parties etc. ORG also includes names of places when they act as administrative entities, as in (1) below which annotates sport teams associated with a location. Corporate designators like *AS*, *Co.* and *Ltd.* should always be included as part of the named entity, as in (2).

(1)	*Vålerenga*$_{ORG}$	*tapte*	*mot*	*Tromsø*$_{ORG}$.
	Vålerenga	lost	against	Tromsø

'Vålerenga lost against Tromsø'

(2)	*Advokatfirmaet*	*Lie & Co*$_{ORG}$	*representerer*
	Lawyers	Lie & Co	represent
	Hansen$_{PER}$.		
	Hansen		

'The lawyers Lie & Co represent Hansen'

Location (LOC) This entity category denotes geographical places, buildings and various facilities. Examples are airports, churches, restaurants, hotels, hospitals, shops, street addresses, roads, oceans, fjords, mountains and parks. Postal addresses are not annotated, but the building, town, county and country within the address are to be annotated, all as LOC entities, see (3).

(3)	*Øvregaten 2a*$_{LOC}$,	*5003*	*Bergen*$_{LOC}$
	Øvre-street 2a,	5003	Bergen

'Øvre-street 2a, 5003 Bergen'

Geo-political entity (GPE) Similarly to OntoNotes (Weischede et al., 2013), but also the Groningen Meaning Bank (GMB) (Bos et al., 2017), we annotate *geo-political entities* (GPEs). This category was introduced in the annotation scheme of the Automatic Content Extraction program (ACE) (Mitchell et al., 2003), and GPEs will further have either location or organization as its sub-type (these are dubbed 'mention roles' in ACE, where also additional such sub-types are defined).

Following Mitchell et al. (2003), GPE entities denote geographical regions that are defined by political and/or social groups. GPES are describe complex entities that refer both to a population, a government, a location, and possibly also a nation (or province, state, county, city, etc.). A GPE must be one of the following: a nation, city or region with a parliament-like government. Parts of cities and roads are not annotated as GPEs.

As mentioned above, GPE entities are further subtyped, either as GPE_LOC or GPE_ORG. If a sense is mostly locative, it should be annotated as GPE_LOC, otherwise it should be GPE_ORG. Example (4) below shows both these entity types.

(4) *Norge*$_{GPE\_ORG}$ *reagerer på politivolden i*
Norway reacts to police violence in
Catalonia$_{GPE\_LOC}$
Catalonia
'Norway reacts to the police violence in Catalonia'

Sometimes the names of GPE entities may be used to refer to other referents associated with a region besides the government, people, or aggregate contents of the region. The most common examples are sports teams. In our annotation scheme, these entities are marked as teams (ORG), as they do not refer to any geo-political aspect of the entity.

Product (PROD) In line with other recent named entity corpora like OntoNotes 5.0 (Weischede et al., 2013) and MEANTIME (Minard et al., 2016), we also annotate *products* as a separate category. Note that while OntoNotes additionally label *works-of-art* as a separate category, as was also done in the Norwegian Nomen Nescio corpus (Johannessen et al., 2005), this is subsumed by the product category in NorNE.

All entities that refer to artificially produced objects are annotated as products. This includes more abstract entities, such as speeches, radio shows, programming languages, contracts, laws and even ideas (if they are named). Brands are PROD when they refer to a product or a line of products, as in (5), but ORG when they refer to the acting or producing entity (6):

(5) *Audi*$_{PROD}$ *er den beste bilen.*
Audi is the best car.
'Audi is the best car'

(6) *Audi*$_{ORG}$ *lager de beste bilene.*
Audi makes the best cars.
'Audi makes the best cars'

Event (EVT) Again similarly to OntoNotes (Weischede et al., 2013), but also the Groningen Meaning Bank (GMB) (Bos et al., 2017), NorNE also annotates *events*. This category includes names of festivals, cultural events, sports events, weather phenomena and wars. Events always have a time span, and often a location where they take place as well. An event and the organization that arranges the event can share a name, but should be annotated with different categories: ORG, as in (7) vs. EVT, as in (8).

(7) *Quartfestivalen*$_{ORG}$ *gikk konkurs i 2008.*
Quart-festival went bankrupt in 2008.
'The Quart Festival went bankrupt in 2008'

(8) *Rolling Stones*$_{ORG}$ *fikk dessverre aldri*
Rolling Stones got unfortunately never
spilt på Quartfestivalen$_{EVT}$*.*
played at Quart-festival
'The Rolling Stones unfortunately never got to play at the Quart Festival'

Derived (DRV) The DRV category is a special category for compound nominals that contain a proper name, as described in section 3.1 above. The main criteria for this category is that the entity in question (i) contains a full name, (ii) is capitalized, and, (iii) is not itself a name. Examples include for instance *Oslo-mannen* 'the Oslo-man'. Names that are inflected and used as common nouns, are also tagged derived (DRV), e.g. *Nobel-prisene* 'the Nobel-prizes'. The reason for the special treatment of these types of forms is that these words do not have a unique entity as a reference, but rather exploit an entity as part of their semantics. Even so, if a user of the annotated data wishes to extract all information about a particular named entity, these may still be relevant, hence should be marked separately.

Entity types not included Of the categories discussed above, *location*, *person*, and *organization* comprise the core inventory of named entity types in the literature. They formed part of the pioneering shared tasks on NER hosted by CoNLL 2002/2003 (Sang, 2002; Sang and Meulder, 2003), MUC-6 (Grishman and Sundheim, 1995) and MUC-7 (Chinchor, 1998), and have been part of all major NER annotation efforts since. However, the CoNLL shared tasks also included a fourth category for names of *miscellaneous* entities not belonging to the aforementioned three. During the annotation of NorNE we similarly operated with an entity type MISC, but eventually we decided to discard this label in the final release of the data as it was annotated too rarely to be useful in practice (with a total of 8 occurrences in the training data but 0 occurrences in both the development and held-out splits).

The MUC shared tasks, on the other hand, additionally included identification of certain types of *temporal expressions* (date and time) and *number expressions* (monetary expressions and percentages). We did not include these in NorNE as we do not strictly speaking consider them named entities.

3.3. Annotation Guidelines

We now turn to present the most relevant aspects of the annotation guidelines developed in the project. Note that the complete annotation guidelines are distributed with the corpus. The guidelines were partly based on Jónsdottír (2003) and Howard (2014), as well as guidelines created for English corpora, in particular the ACE (Mitchell et al., 2003) and CoNLL (Sang and Meulder, 2003) datasets.

3.3.1. Main Annotation Criteria

The guidelines formulate a number of criteria for the annotations. In general, the annotated entities should have a unique reference which is constant over time. Further, the annotated text spans should adhere to the following annotation criteria:

Proper nouns The text span corresponds to or contains a proper noun, e.g. *Per Hansen* which consists of two consecutive proper nouns with a single reference. Names may include words which are not proper nouns, e.g. prepositions, as in *Universitetet i Oslo* 'University of Oslo'.

Span The maximum span of a name should be annotated, rather than its parts, e.g. in *Høgskolen i Oslo og Akershus* 'the College of Oslo and Akershus', *Oslo* and *Akershus* are not annotated separately as locations.

Capitalization As mentioned above, it is often the case in Norwegian that only the first token in a multi-token named entity is capitalized. We treat the subsequent tokens as part of the name if they in combination denote a unique named entity, e.g. *Oslo rådhus* 'Oslo city-hall'.

Titles Most titles do not have an initial capital letter in Norwegian. Exceptions are some instances of royal titles. We never annotate titles as a name or part of a name, even when they are capitalized. Official work titles like *Fylkesmannen* 'the County official' should be annotated as organizations because they refer to official institutions, as in (9). When the same entity refers to the person/occupation, as in (10), they are not, however, annotated as named entities.

(9) *I går ble det klart at*
in yesterday was it clear that
Fylkesmannen$_{ORG}$ *har vedtatt* ...
County-man has declared ...
'Yesterday, it was clear that the County official has declared ...'

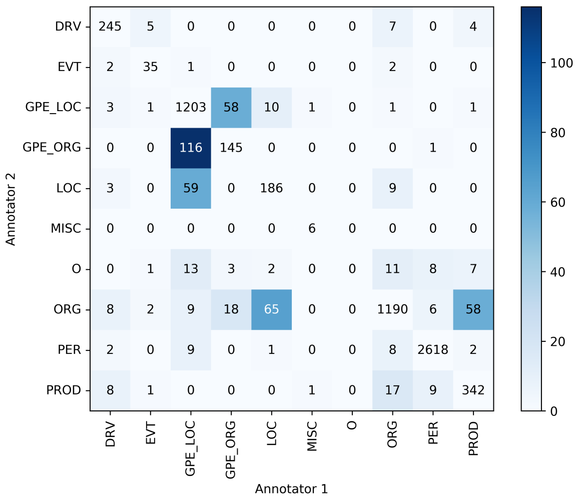
(10) *Fylkesmannen kjørte i grøfta sør for Oslo.*
County-man drove in ditch south of Oslo
'The county official drove into a ditch south of Oslo'

Hyphenation Names that include hyphens should be annotated as a named entity if and only if they constitute a new name, e.g. *Lillehammer-saken*$_{EVT}$ 'the Lillehammer case' which denotes a specific court case, hence has a unique reference. The similar, but generic, *Lillehammer-advokaten* 'the Lillehammer lawyer' is not a named entity, but rather a compound noun, and should therefore receive the special category DRV.

3.3.2. Ambiguity and Metonymy

Ambiguity is a frequent source of doubt when annotating. This is often caused by so-called metonymical usage, where an entity is referred to by the name of another, closely related entity (Lakoff and Johnson, 1980). In the annotation of the NorNE corpus we have tried to resolve the ambiguity and choose the entity type based on the context (the document). We assume that every entity has a base, or literal, meaning and that when there is ambiguity, either genuinely or due to a lack of context, we resort to the literal meaning of the word(s) (Markert and Nissim, 2002). For instance, in the example in (11) below, the context does not clearly indicate whether this is a reference to a geo-political location or organization. We here assume that the location sense is the literal sense of the word *Vietnam* and that the organization sense is by metonymical usage, hence the annotation is GPE_LOC.

(11) *Vietnam*$_{GPE\_LOC}$ *er flott.*
Vietnam is great.
'Vietnam is great.'



	DRV	EVT	GPE_LOC	GPE_ORG	LOC	MISC	O	ORG	PER	PROD
DRV	245	5	0	0	0	0	0	7	0	4
EVT	2	35	1	0	0	0	0	2	0	0
GPE_LOC	3	1	1203	58	10	1	0	1	0	1
GPE_ORG	0	0	116	145	0	0	0	0	1	0
LOC	3	0	59	0	186	0	0	9	0	0
MISC	0	0	0	0	0	6	0	0	0	0
O	0	1	13	3	2	0	0	11	8	7
ORG	8	2	9	18	65	0	0	1190	6	58
PER	2	0	9	0	1	0	0	8	2618	2
PROD	8	1	0	0	0	1	0	17	9	342

Figure 1: Entity confusion matrix for the two annotators.

3.4. Annotation Process

The annotation of the NorNE corpus was performed by two trained linguists, and all documents in the Bokmål section were doubly annotated. As the second phase of the project, the Nynorsk section was annotated by a single annotator. Disagreements in annotations were discussed jointly at regular intervals and the final analysis was agreed upon by both annotators, often followed by an update to the annotation guidelines. All annotation was performed using the Brat web-based annotation tool (Stenetorp et al., 2012).

3.5. Inter-Annotator Agreement

The annotation proceeded in two stages; a first round was used to train the annotators, and hence is not representative of the subsequent annotation process. We only consider the second round of annotation to calculate inter-annotator agreement scores, prior to consolidation. At this stage, the annotation agreement shows a micro F_1-score of 91.5 over a set of 138 documents (of 460 in total), comprising 185k tokens and 8434 entities.

The inter-annotator agreement scores are calculated at the entity-level and the annotation for an entity is considered correct only when: (i) both annotators agree on the exact span of the entity, and (ii) both annotators agree on the entity type. Tokens that are not part of an entity, i.e., where both annotators agree on the O (outside) tag are not considered. In cases where only one of the two annotators has marked an entity, this is still considered as a disagreement.

Figure 1 shows a confusion matrix for the annotations performed in the second round. While the agreement is generally very high, we observe that the main difficulties of the annotators revolve around the distinction of the subcategories of the GPE entity type, in particular confusion between the GPE_ORG and GPE_LOC types themselves. The annotators also have some difficulties in distinguishing these from the main category of LOC and in distinguishing between the general LOC and ORG entity types. Further, we find that the PROD category is often confused for the

Entity	Annotator 1	Annotator 2
Afghanistan	GPE_LOC	GPE_ORG
Norge	GPE_ORG	GPE_LOC
Dagbladet	ORG	PROD
The Economist	ORG	PROD
Facebook	ORG	PROD
Pantera	PROD	ORG

Table 1: Examples of common annotator disagreements.

Standard	Sentences	Tokens	Entities
Bokmål (BM)	16,309	301,897	14,369
Nynorsk (NN)	14,878	292,315	13,912

Table 2: Total number of sentences, tokens and annotated named entities in NorNE. The dataset has a separate section for each of the two official written standards of the Norwegian language; Bokmål and Nynorsk.

Type	Train	Dev	Test	Total	%
PER	4033	607	560	5200	36.18
ORG	2828	400	283	3511	24.43
GPE_LOC	2132	258	257	2647	18.42
PROD	671	162	71	904	6.29
LOC	613	109	103	825	5.74
GPE_ORG	388	55	50	493	3.43
DRV	519	76	48	644	4.48
EVT	131	9	5	145	1.00

Table 3: Entity distributions for Bokmål (BM) in NorNE.

ORG category. Table 1 presents some examples of common annotator disagreements. Common entity sub-types among the PROD/ORG disagreements are newspapers, magazines, web sites and bands.

4. Dataset Overview

The corpus is divided in two parts, one for each of the official written standards of the Norwegian language: Bokmål (BM) and Nynorsk (NN). Note that the two parts contain different texts, not translations. Global counts of sentences, tokens and annotated entities are shown in Table 2. On a more granular level, Tables 3 and 4 summarizes the number of annotations of each entity type for Bokmål and Nynorsk respectively, broken down across the data splits for training, development, and held-out testing. Note that NorNE reuses the same 80-10-10 split previously defined for NDT for both the Bokmål part (Hohle et al., 2017) and the Nynorsk part (Velldal et al., 2017), which aimed to preserve contiguous texts in the various sections while also keeping the splits balanced in terms of genre. Note that, as distributed, NorNE follows the CONLL-U format, with entities labeled according to the IOB2 scheme.

Type	Train	Dev	Test	Total	%
PER	4250	481	397	5128	36.86
ORG	2752	284	236	3272	23.51
GPE_LOC	2086	195	171	2452	17.62
PROD	728	86	60	874	6.28
LOC	893	85	82	1060	7.61
GPE_ORG	367	66	11	444	3.19
DRV	445	50	30	525	3.77
EVT	141	7	9	157	1.12

Table 4: Entity distributions for Nynorsk (NN) in NorNE.

5. Experimental Results and Analysis

In this section we present some preliminary experimental results for named entity recognition using NorNE. We investigate the effects of using different mappings of the label set, different label encodings (IOB2, etc), different embedding dimensionalities, as well as joint modeling of the Bokmål and Nynorsk variants. Apart from the joint modeling, the other experiments will target only the Bokmål section of the dataset. Before moving on to the results, we first briefly outline the experimental setup.

5.1. Experimental Setup

The modeling is performed using NCRF++ (Yang and Zhang, 2018) – a configurable sequence labeling toolkit built upon PyTorch. Following Yang et al. (2018), our particular model configuration is similar to the architecture of Chiu and Nichols (2016) and Lample et al. (2016), achieving results that are close to state-of-the-art for English on the CoNLL-2003 dataset: it combines a character-level CNN and a word-level BiLSTM, finally feeding into a CRF inference layer. The input to the word-level BiLSTM is provided by the concatenation of (1) the character sequence representations from the CNN using max-pooling in addition and (2) pre-trained word embeddings from the NLPL vector repository[2] (Fares et al., 2017). Further details about the latter are provided in the next section.

Across all experiments we fix and re-use the same random seed for initializing the models, as to reduce the effect of non-determinism, and otherwise fix the parameters to their default values.[3]

For model evaluation we follow the scheme defined by the SemEval 2013 task 9.1 (Segura-Bedmar et al., 2013), using the re-implementation offered by David S. Batista.[4] We report F1 for exact match on the entity level, i.e., both the predicted boundary and entity label must be correct. (This measure was dubbed *strict* in SemEval 2013 task 9.1.)

[2] http://vectors.nlpl.eu/repository/

[3] Parameter settings include the following: optimizer=SGD, epochs=50, batch size=10, dropout=0.50, learning rate = 0.015 with a decay of 0.05, L2-norm=1^{-8}, seed=42.

[4] https://github.com/davidsbatista/NER-Evaluation

Pre-train.	Dim.	F_1
none	100	76.54
CBOW	100	84.36
SG	50	87.61
SG	100	89.47
SG	300	90.02
SG	600	**90.75**

Table 5: Evaluating the impact of fastText pre-training, testing on the Bokmål development split of NorNE.

5.2. The Use of Pre-Training

In a preliminary round of experiments, we evaluated the impact of pre-training on the model. The word embeddings are trained on the Norwegian News Corpus (over 1 billion tokens of Norwegian Bokmål) and NoWaC (Norwegian Web as Corpus; approximately 687 million tokens of Bokmål) using fastText (Bojanowski et al., 2017) with a vocabulary of 2.5 million unique words and a window size of 5. We here re-use embeddings that are made available by the NLPL vector repository (Fares et al., 2017); for more details on the training process see Stadsnes (2018) and Stadsnes et al. (2018).

Table 5 shows results for the Bokmål development split both with and without pre-training, and using both the CBOW and SkipGram algorithm as implemented in fastText. Unsurprisingly, we observe that pre-training substantially boosts performance of the NER model. Moreover, we observe that CBOW (here only shown for a 100-dimensional model) is substantially outperformed by the SkipGram model, and that performance steadily increase with increased dimensionality. In all the subsequent experiments reported in the paper we use a SkipGram model with a dimensionality of 600.

5.3. Label set and Label Encoding

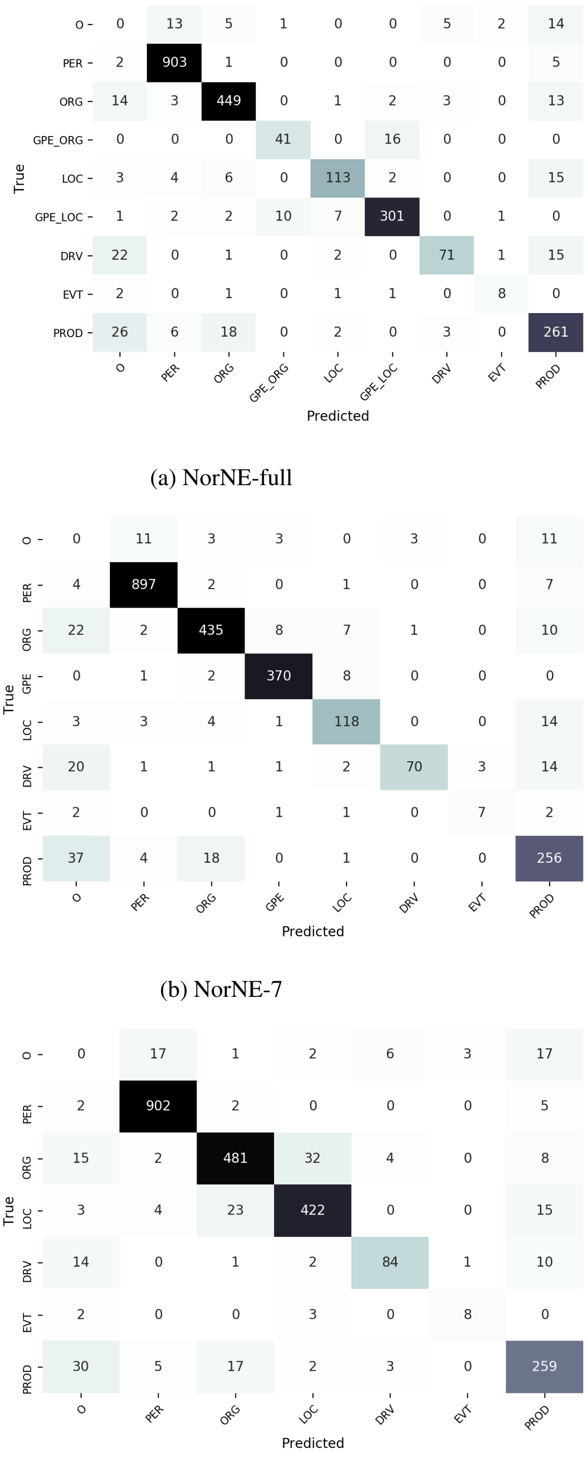
In this section we investigate the interactions between choice of label set and label encoding. On the one hand we experiment with the granularity of the label set or entity types; mapping the original entity types to a smaller set of more general types. On the other hand we experiment with mapping the IOB labels specified in the distributed corpus to variations of the BIOES (BIOLU) label encoding scheme.

5.3.1. Label Set

We consider the following label mappings:

- **NorNE-full**: Using the full set of 8 entity types, as in the experiments above.
- **NorNE-7**: Conflating instances of the geo-political subcategories GPE_ORG and GPE_LOC to the more general type GPE, yielding 7 entity categories.
- **NorNE-6**: Dispensing with the geo-political types entirely, merging GPE_ORG and GPE_LOC into ORG and LOC respectively, yielding 6 entity categories.

The question of what comprises the most suitable level of granularity ultimately depends on the downstream task, but in this section we report experimental results for training and testing with the different labels sets to analyze the learnability of the different granularities.



(a) NorNE-full

(b) NorNE-7

(c) NorNE-6

Figure 2: Aggregated confusion matrices for the different label granularities; NorNE-full, NorNE-7 and NorNE-6.

			B-PER			B-GPE_LOC	
Til	dagleg	jobbar	Sydnes	mykje	utafor	Noregs	landegrenser
Til	daglig	jobber	Sydnes	mye	utenfor	Norges	landegrenser
To	*daily*	*works*	*Sydnes*	*lot*	*outside*	*Norway's*	*borders*

Figure 3: Example sentence in Nynorsk (second row) and Bokmål (third) with English gloss (bottom) and IOB named entity labels (top). A fluent translation to English would be *On a daily basis Sydnes works a lot outside of Norway's national borders.*

	IOB	IOBE	IOBS	IOBES
NorNE-full	90.75	90.45	**90.76**	90.58
NorNE-7	91.86	91.80	91.99	**92.29**
NorNE-6	90.95	90.50	**91.85**	91.38

Table 6: Perfomance comparison of models trained on NorNE-full using IOB, IOBE, IOBS and IOBES encodings, when evaluated on the NorNE development set, using label sets of different granularities.

5.3.2. Label Encoding

The annotations of NorNE are distributed using the standard IOB(2) scheme.[5] However, this can be easily mapped to other variations like IOBES, where the extra E-label indicates the end-token of an entity and the S-label indicates single, unit-length entities. (This latter scheme also goes by other names like BIOLU.) Several studies have reported slight performance increases when using the IOBES encoding compared to IOB (Ratinov and Roth, 2009; Yang et al., 2018; Reimers and Gurevych, 2017). However, it is typically not clear whether the benefits stem from adding the E- or S-labels or both.

5.3.3. Results

Table 6 report experimental results for all of these variations – i.e. isolating the effects of the E- and S-labels – and across all the three different sets of entity types discussed above.

There are several things to notice here. IOBE seems to have a negative performance impact regardless of the chosen label set. Also, compared to the standard IOB encoding, IOBES also has a negative impact paired with NorNE-full, but gives improved results together with NorNE-7 and NorNE-6. It is not easy to pinpoint *exactly* why we get lower results for NorNE-full when used with more fine-grained encodings, but possible explanation could be that the resulting increased class sparsity has more of an effect using the full label-set.

Interestingly, we find that IOBS gives better results for all label sets, giving the highest F1 scores for both NorNE-full and NorNE-6, and also a marginal performance increase for NorNE-7. In other words, no matter the label set it seems beneficial to use a dedicated tag for single-token entities, while the benefits of including an end-tag (or both) are less clear.

Regardless of the chosen encoding, we see that the NorNE-7 label-set yields the highest scores. We also see that reducing the label granularity always leads to higher absolute scores compared to using the full label-set. This is not in itself informative however and is an effect that be expected just from the fact the label ambiguity is reduced.

Figure 2 shows confusion matrices for all label-sets, using models trained with the IOB encoding, making for some interesting observations. First of all, we see that there is little confusion among the entities `ORG`, `GPE_ORG`, `LOC` and `GPE_LOC` in NorNE-full. At the same time, collapsing `GPE_LOC` and `GPE_ORG` to a single category in NorNE-7 does not seem detrimental, with a marginal amount of confusion between `LOC`, `ORG`, and `GPE`. However, with NorNE-6, valuable information appears to have been lost when removing the geo-political category, introducing more confusion between the location and organization category.

In general, we see that products (`PROD`) seems to be a difficult category to classify, likely reflecting the rather heterogeneous character of this category. Moreover, this entity category has the longest average token span, and one might suspect that long entities might have more boundary-related errors, which could explain the high confusion with the *outside-of-entity* class.

Of course, the appropriate choice of label set ultimately depends on the downstream use case. However, unless one really needs to distinguish between the `GPE` sub-categories, our experiments above seem to point to NorNE-7 label set as a good option, possibly in combination with an IOBES encoding.

5.4. Joint Modeling of Bokmål and Nynorsk

As mentioned in the introduction, there are two official written standards of the Norwegian language; Bokmål (literally 'booktongue') and Nynorsk (literally 'new Norwegian'). Focusing on dependency parsing, Velldal et al. (2017) investigated the interactions across the two official standards with respect to parser performance. The study demonstrated that while applying parsing models across standards yields poor performance, combining the training data for both standards yields better results than previously achieved for each of them in isolation. We here aim to investigate similar effects for named entity recognition.

5.4.1. Background: On Nynorsk and Bokmål

While Bokmål is the main variety, roughly 15% of the Norwegian population uses Nynorsk. However, language leg-

[5] In IOB2, the B-label is used at the beginning of every named entity, regardless of its span, while in the IOB1 variant the B-label is only used when a named entity token is followed by a token belonging to the same entity.

islation specifies that minimally 25% of the written public service information should be in Nynorsk. The same minimum ratio applies to the programming of the Norwegian Public Broadcasting Corporation (NRK). The two varieties are so closely related that they may in practice be regarded as 'written dialects'. However, lexically there can be relatively large differences.

Figure 3 shows an example sentence in both Bokmål and Nynorsk. While the word order is identical and many of the words are clearly related, we see that only 3 out of 8 word forms are identical. When quantifying the degree of lexical overlap with respect to NDT – the treebank data that we too will be using – Velldal et al. (2017) find that out of the 6741 non-punctuation word forms in the Nynorsk development set, 4152, or 61.6%, of these are unknown when measured against the Bokmål training set. For comparison, the corresponding proportion of unknown word forms in the Bokmål development set is 36.3%. These lexical differences are largely caused by differences in productive inflectional forms, as well as highly frequent functional words like pronouns and determiners.

5.4.2. A Joint NER Model

For the purpose of training a joint NER model we also train a new version of the fastText SkipGram embeddings on the NNC and NoWaC corpora, using the same parameters as before (and 600 dimensions), but this time including the available Nynorsk data for NNC, amounting to roughly 60 million additional tokens.

Several interesting effects can be observed from the results in Table 7. The first two rows show the results of training single-standard NER models (like before) with the joint embedding model, but this time also testing across standards; training a model on Bokmål and applying it to Nynorsk, or *vice versa*. As can be clearly seen, performance drops sharply in the cross-standard settings (italicized in the table). For example, while the Bokmål NER model achives an F1 of 89.47 on the Bokmål development data, the performance plummets to 82.34 when the same model is applied to Nynorsk.

The last row of Table 7 shows the effects of training a *joint* NER model on the combination of the Bokmål and Nynorsk data (randomly shuffling the sentences in the combined training and validation splits). We see that the joint model substantially outperforms both of the single-standard models on their respective development splits. On the heldout splits, the joint model again has much better performance for Nynorsk, although the single-standard setup shows slightly better results for Bokmål.

The results for the joint modeling setup is a double-win with immediate practical consequences: Not only do we see comparable or increased performance, it also means we only need to maintain a single model when performing NER for Norwegian. The alternative would be to either accept a sharp drop in performance whenever Nynorsk, say, was encountered, *or* to first perform language identification to detect the given variety and then apply the appropriate model.

	Development		Heldout	
Training	BM	NN	BM	NN
BM	89.47	*82.34*	**83.89**	*81.59*
NN	*84.01*	86.53	*76.88*	83.89
BM+NN	**90.92**	**88.03**	83.48	**85.32**

Table 7: Joint and cross-standard training and testing of NER models. The first column indicates the language standard used for training the NER model; either Bokmål (BM), Nynorsk (NN), or both. The Development and Heldout columns shows F1 scores when testing on the respective splits of either standard. The italicized scores correspond to a cross-standard setup where the language variety used for training is different from testing. Bold indicates best performance.

6. Summary

This paper has documented a large-scale annotation effort adding named entities to the Norwegian Dependency Treebank. The resulting dataset – dubbed NorNE – is the first publicly available[6] dataset for named entity recognition (NER) for Norwegian and covers both of the official written standards of the Norwegian language – Bokmål and Nynorsk – comprising roughly 300,000 tokens of each. The annotations include a rich set of entity types including persons, organizations, locations, geo-political entities, products, and events, in addition to a class corresponding to nominals derived from names. In addition to discussing the principles underlying the manual annotations, we provide an in-depth analysis of the new dataset through an extensive series of first benchmark NER experiments using a neural sequence labeling architecture (combining a character-level CNN and a word-level BiLSTM with a CRF inference layer). Among other results we demonstrate that it is possible to train a joint model for recognizing named entities in Nynorsk and Bokmål, eliminating the need for maintaining separate models for the two language varieties.

Acknowledgements

NorNE is the result of a collaborative effort involving the Schibsted Media Group, Språkbanken ('the Language Bank') of the National Library of Norway, and the Language Technology Group at the Department of Informatics, University of Oslo. We would like to thank the annotators, Vilde Reichelt and Cato Dahl, for their meticulous efforts in providing accurate and consistent annotations.

Bibliographical References

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

[6] https://github.com/ltgoslo/norne/

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The Groningen Meaning Bank. In Pustejovsky J. Ide N., editor, *Handbook of Linguistic Annotation*, pages 463–496. Springer, Dordrecht.

Chinchor, N. (1998). Appendix E: MUC-7 Named Entity Task Definition (version 3.5). In *Proceedings of 7th Message Understanding Conference*, San Diego, USA.

Chiu, J. P. and Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). *Norsk Referansegrammatikk*. Universitetsforlaget, Oslo.

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 271–276, Gothenburg, Sweden.

Grishman, R. and Sundheim, B. (1995). Design of the muc-6 evaluation. In *Proceedings of the 6th Message Understanding Conference*, pages 1–11, Columbia, Maryland, USA.

Hoff, K. and Preminger, M. (2015). Usability testing of an annotation tool in a cultural heritage context. *Metadata and Semantics Research*, 544:237–248.

Hohle, P., Øvrelid, L., and Velldal, E. (2017). Optimizing a PoS tagset for Norwegian dependency parsing. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 142–151, Gothenburg, Sweden.

Howard, M.-H. (2014). Manuell annotering i nrk-prosjektet: Hvilke utfordringer må det tas stilling til? Bachelor's thesis at HiOA: Oslo and Akershus University College.

Johannessen, J. B., Hagen, K., Haaland, Å., Jónsdottir, A. B., Nøklestad, A., Kokkinakis, D., Meurer, P., Bick, E., and Haltrup, D. (2005). Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*, 20(1).

Johannessen, J. B., Hagen, K., Lynum, A., and Nøklestad, A. (2012). OBT+stat. A combined rule-based and statistical tagger. In Gisle Andersen, editor, *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*. John Benjamins Publishing Company.

Johansen, B. (2019). *Automated analysis of Norwegian text*. Ph.D. thesis, University of Bergen, Bergen, Norway, 6.

Jónsdottír, A. B. (2003). ARNER: what kind of name is that? Master's thesis, University of Oslo, Norway.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago, IL.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California.

Markert, K. and Nissim, M. (2002). Towards a corpus annotated for metonymies: The case of location names. In *Proceedings of the 3rd international Conference on Language Resources and Evaluation (LREC 2002)*.

Minard, A., Speranza, M., Urizar, R., Altuna, B., van Erp, M., Schoen, A., and van Son, C. (2016). Meantime, the newsreader multilingual event and time corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., and Sundheim, B. (2003). Ace-2 version 1.0. Web Download. LDC Catalog No. LDC2003T11.

Nøklestad, A. (2009). *A Machine Learning Approach to Anaphora Resolution Including Named Entity Recognition, PP Attachment Disambiguation, and Animacy Detection*. Ph.D. thesis, University of Oslo.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, Boulder, Colorado.

Reimers, N. and Gurevych, I. (2017). Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 7th Conference on Computational Natural Language Learning*, Edmonton, Canada.

Sang, E. F. T. K. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the 6th Conference on Computational Natural Language Learning*, Stroudsburg, PA, USA.

Segura-Bedmar, I., Martínez, P., and Herrero Zazo, M. (2013). SemEval-2013 Task 9 : Extraction of Drug–Drug Interactions from Biomedical Texts. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. B. (2014). The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Stadsnes, C., Øvrelid, L., and Velldal, E. (2018). Evaluating Semantic Vectors for Norwegian. In *Proceedings of the 31st Norwegian Informatics Conference (NIK)*, Longyearbyen, Svalbard.

Stadsnes, C. (2018). Evaluating Semantic Vectors for Norwegian. Master's thesis, University of Oslo.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, pages 102 – 107, Avignon, France.

Tallerås, K., Massey, D., Husevåg, A.-S. R., Preminger, M., and Pharo, N. (2014). Evaluating (linked) metadata transformations across cultural heritage domains. *Metadata and Semantics Research*, 478:250–261.

Velldal, E., Øvrelid, L., and Hohle, P. (2017). Joint UD parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, pages 1–10, Gothenburg, Sweden.

Weischede, R., Palmer, M., Marcus, M., Hovy, E., Pradhan, S., Ramshaw, L., Xue, N., Taylor, A., Kaufman, J., Franchini, M., El-Bachouti, M., and Houston, R. B. A. (2013). Ontonotes release 5.0. Web Download. LDC Catalog No. LDC2013T19.

Yang, J. and Zhang, Y. (2018). NCRF++: An Open-source Neural Sequence Labeling Toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics – System Demonstrations*, pages 74–79, Melbourne, Australia.

Yang, J., Liang, S., and Zhang, Y. (2018). Design Challenges and Misconceptions in Neural Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3879–3889, Santa Fe, New Mexico, USA.