An Ontology-based Approach to Automatic Part-of-Speech Tagging Using Heterogeneously Annotated Corpora

Maria Sukhareva[1,2] **and Christian Chiarcos**[1]
[1] Goethe University Frankfurt am Main, Germany
[2] Technical University Darmstadt, Germany
{sukharev|chiarcos}@informatik.uni-frankfurt.de

Abstract

In the LOD era, the conceptual interoperability of language resources is established by using modular architectures like the Ontologies of Linguistic Annotations (Chiarcos, 2008a, OLiA). Available as a part of the Linguistic Linked Open Data (LLOD) cloud,[1] OLiA provides ontological representations of annotation schemes for over 70 languages, as well as their linking to a reference model. We successfully train an ontology-based POS tagger on corpora with different tag sets of divergent granularity and partially compatible annotations. Making use of OLiA, we achieve interoperability of annotation schemes, and, despite sparse training data, we do not only outperform state-of-the-art POS taggers in concept coverage, but also show how traing on heterogeneously annotated data produces richer morphosyntactic annotation with no or only marginal loss of precision.

1 Introduction

Ontologies have long been recognized as a primary device for interoperability among annotations and linguistic descriptions (Farrar and Langendoen, 2003; Ide and Romary, 2004; Saulwick et al., 2005), and they have been applied to facilitate querying (Saulwick et al., 2005; Rehm et al., 2007), interoperability among modules in NLP pipelines (Buyko et al., 2008; Hellmann, 2010), or for post-processing (i.e., merging, enriching or disambiguating) the output of NLP tools (Pareja-Lora and Aguado de Cea, 2010; Chiarcos, 2010a; Hellmann et al., 2013). In this paper, we describe a novel approach towards the next challenge along this trajectory, i.e., the development of NLP tools that can directly produce and consume ontological descriptions.

In comparison with classical, string-based annotation, key advantages include a detailed assessment of classification accuracy for different annotation concepts (rather than for opaque strings representing bundles of these), a freely scalable degree of granularity (the system produces statements at all levels of granularity), and interoperability with state-of-the-art technologies from NLP and the Semantic Web. Another advantage is that annotations from different sources become interoperable, and tools can be trained on annotations from multiple corpora annotated according to different schemes.

In this regard, this paper describes a novel approach toward automatic part-of-speech (POS) annotation, and investigates the extent to which ontology-based annotations allow us to train NLP tools on corpora with divergent, but conceptually related annotations, and whether the increase in the granularity of analysis outweighs possible losses in precision arising from the heterogeneity of the training data.

2 Corpora

For reasons of interpretability, we use English corpora for this experiment, but we consider the approach to be language-independent, and (in the longer perspective) particularly relevant to less-resourced languages with a lower degree of *de facto* standardization in annotated corpora than English. Historical and modern less-resourced languages are often annotated according to a great variety of annotation schemes which can not be trivially mapped to a generalization without substantial loss of information. In order to emulate the conditions for less-resource languages, we use two heterogeneously annotated, but deliberately small corpora. Even though the amount of annotated training data is much lower than in traditional ap-

[1] http://linguistic-lod.org

Proceedings of the Second Workshop on Natural Language Processing and Linked Open Data, pages 23–32,
Hissar, Bulgaria, 11 September 2015.

	training	test	total	tag set
EWT	50,767	4,767	55,534	51
Susanne	54,109	4,886	58,995	270

Table 1: Corpus statistics: tokens , tagsets with number of POS tags

proaches, we outperform state-of-the-art taggers in concept coverage and precision (Sect. 6).

We conduct our experiments on two manually annotated corpora with different annotation schemes, namely, *Susanne* (Sampson, 1995), and the *English Web Treebank* (Silveira et al., 2014, EWB), Tab. 1.

Susanne contains annotations of 130,000 words of literary prose, drawn from the (unannotated) Brown corpus. Its hallmark is the Susanne-specific tagset (further Susa) with its high granularity and detailization of POS tags (270 unique tags). In addition, the Penn Treebank (Taylor et al., 2003, PTB) includes an independent annotation of the Susanne corpus, which enabled us to conduct the evaluation on the data annotated with both PTB and Susanne tags.

The EWT is a corpus of online reviews manually annotated with the PTB tag set. In comparison with Susanne, the lexical diversity of the EWT reviews is lower which can easily be explained by the peculiarities of the genre. Here, we use a subsection of Susanne proportional to the size of the EWT reviews and a 90:10 split into training and test corpora, respectively.

3 Ontologies of Linguistic Annotations

The **Ontologies of Linguistic Annotations** (Chiarcos, 2008a)[2] represent an architecture of OWL2/DL ontologies that formalize the mapping between annotations, a 'Reference Model' and existing terminology repositories ('External Reference Models'): OLiA solves the problem of different heterogeneous schemes by a modularized representation of annotation schemes and its declarative linking with an overarching Reference Model. Unlike a tag set, whose string-based annotations require disjoint categories at a fixed level of granularity, this ontology-based approach allows to decompose the semantics of annotations and consider all aspects independently.

The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al., 2006), and within the LLOD cloud, OLiA serves as a vocabulary hub for linguistic terminology for various phenomena and resources. It currently provides ontological representations for over 70 languages with morphological, morphosyntactic, syntactic and discourse levels of annotation.

3.1 OLiA Architecture

In the OLiA architecture, four different types of ontologies are distinguished (cf. Fig. 1):

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.
- Multiple OLiA ANNOTATION MODELs formalize annotation schemes and tag sets. Annotation Models are based on the original documentation, so that they provide an interpretation-independent representation of the annotation scheme.
- For every Annotation Model, a LINKING MODEL defines $\sqsubseteq$ relationships between concepts in the respective Annotation Model and the Reference Model. Linking Models are interpretations of the Annotation Model in terms of the Reference Model.
- Community-maintained terminology repositories in OWL2/DL (Farrar and Langendoen, 2003; Saulwick et al., 2005, etc.), are integrated as EXTERNAL REFERENCE MODELS: Linking Models specify $\sqsubseteq$ relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g., *olia:Determiner*) and grammatical features (e.g., *olia:Accusative*), as well as properties that define relations between these (e.g., *olia:hasCase*).

Conceptually, Annotation Models differ from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model.

Figure 1 gives the ontological representation of the Susanne tag `APPGf` as an example, used for

[2] `http://purl.org/olia/`, includes PTB and Susa models

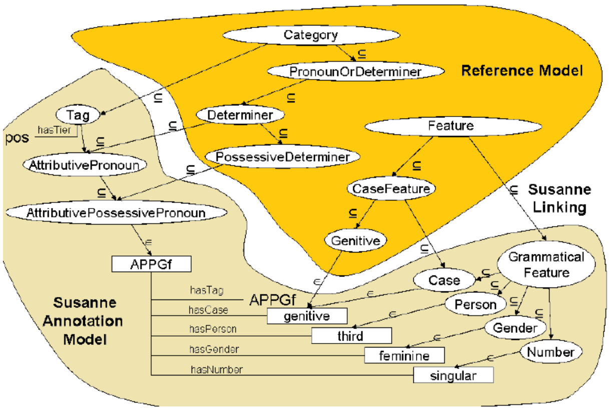

Figure 1: The Susanne tag `APPGf`, its representation in the Annotation Model and (partial) linking with the Reference Model, cf. Chiarcos (2008a)

her as a possessive determiner, the corresponding inheritance structure of the word class and the case property. Using the inheritance structures in the Linking Model, the tag can be rendered in terms of the Reference Model by the following OWL2 class description

$$PossessiveDeterminer \sqcap \exists hasCase.Genitive \sqcap \exists hasPerson.Third \sqcap \exists hasGender.Feminine \sqcap \exists hasNumber.Singular$$

Through ontological inheritance within the Reference Model, we can further infer that `APPGf` is also an instance of *Determiner* and *PronounOrDeterminer* (superconcepts of *PossessiveDeterminer*).

One important difference between this description and the (similar) description in terms of the Annotation Model is that this description is tagset neutral, and does not only apply to the English *her* as a possessive, but also to the corresponding tags in other annotation schemes (even if from different languages), e.g., the PTB tag for *her*, `PRP$`. Although this does provide a partial description only (*PossessiveDeterminer* ⊓ *Determiner* ⊓ *PronounOrDeterminer*), we can generalize over both tags by referring to atomic statements found in both ontological renderings (i.e., their intersection).

3.2 Related Research

Using OLiA for processing of heterogeneously annotated corpora has several benefits in comparison with other approaches. As such, we would like to emphasize that the ontology-based approach is *lossless*. Instead of simplifying heterogeneous tag sets to a common meta tag set or creating a mapping between the tag sets, we decompose tag sets into statements (triples) grounded in an ontology. This is a major difference as compared to radically reductionist approaches like Petrov et al. (2012) which inevitably lead to an extensive information loss, especially for highly detailed annotation schemes such as Susanne. A different kind of information loss frequently occurs with approaches based on a meta tag set as 'interlingua' (Leech and Wilson, 1996; Zeman, 2008): Here, a taxonomy tags is enforced from one set of languages (that the taxonomy was developed for) to another, where the pressure to stay within the pre-defined model frequently leads to 'tag abuse', see Chiarcos and Erjavec (2011) for the corresponding analysis of MULTEXT-East (Erjavec, 2004). But it also differs from more flexible, bottom-up-grown meta tag sets (Zeman, 2008), because without the implicit disjointness assumption of tags (categories) in classical tagsets, it is possible to preserve divergent, but compatible analyses, e.g., *enduring* in *capable of enduring friendships* is both a verb (morphologically) and an adjective (syntactically).

As being lossless, OLiA ensures that the information contained in the original schemes will be preserved to a maximal extent by its conceptual representation.

3.3 From OLiA to neural networks

Originally, the OLiA ontologies were conceived for conceptually interoperable information retrieval and tag set independent corpus querying (Saulwick et al., 2005; Rehm et al., 2007), but also have found a use case in NLP, so far, however, only to *represent* the output of modules in an NLP pipeline in a tool-independent fashion (Buyko et al., 2008; Hellmann, 2010), or to *merge* the output of different NLP tools in an ensemble combination architectures, where information from different sources (say, NLP tools) was integrated on the basis of the Reference Model and disambiguated using ontological axioms (Chiarcos, 2010a).

Here, we describe the first approach on directly produce ontology-based descriptions, with an ontology-based POS tagger, opening the field for future applications of ontology-based NLP which raises the current string-based state of the art of annotation in NLP to conceptual annotation processing. In order to do so, we employ a neural network architecture, as its output vector is capable to represent and to predict probability/confidence scores for all concepts and features in the ontology simultaneously, regardless of whether these are compatible with each other.

Then, for encoding and decoding annotations, *MorphosyntacticCategory*s from the OLiA Reference Model are employed. Note that for the experiments described here, we only consider these and leave morphosyntactic (and other) features for subsequent research.

4 Configuring and Training Neural Networks

We trained neural networks on EWT reviews, an equally sized subset of the Susanne corpus (Sect. 2), and on both training sets combined. The core of the algorithm is a feed-forward neural network with resilient backpropagation with the following structure:

1. 75 input neurons that correspond to three 25-dimensional word embeddings (Turian et al., 2010)[3] of the target word, its predecessor and its successor from its immediate context;

2. one hidden layer with the tanh activation function. The number of neurons in the hidden layer is heuristically set to the average length of input and output layers, thus, a natural geometric (pyramidal) design;

3. a layer of output neurons that represent OLiA *MorphosyntacticCategory*s, again with tanh normalization. The activations of these neurons represent the output vector.

The first step of our algorithm is generation of OLiA triples from heterogeneously annotated corpora using existing Susa and PTB annotation and linking models. Instead of a POS tag, every word is annotated with a set of triples, each assigning the word a *MorphosyntacticCategory* as its associated class (concept). For example, the Susanne tag `AT` for the definite article *the* is now annotated with RDF triples like *_:word*$_i$ *a olia:DefiniteArticle*. For the sake of simplicity, we abbreviate OLiA type assignment triples for any given word here by the assigned concept, here *DefiniteArticle*. Through subsumption inference over the ontology, every *DefiniteArticle* is also an *Article*, the full set of classes for `AT` can thus be given as {*DefiniteArticle, Article, Determiner, PronounOrDeterminer*}, for the Susanne tag `AT1` (indefinite article *a*) as {*IndefiniteArticle, Article, Determiner, PronounOrDeterminer*}, etc.

The PTB tag set is not as rich as Susa and does not distinguish between definite and indefinite articles, assigning to both *the* and *a* the tag `DT`. It conversion to OLiA thus yields the set {*Determiner, PronounOrDeterminer*}.

In the training data, the target vector is then populated with ternary values for assigned triples ($+1$), underspecified/non-predictable triples (0, i.e., not predictable from the given tag set), and non-assigned (but predictable) triples (-1) for a given gold annotation. For a tag set $X \in$ {PTB, Susa}, T_X is the set of unique OLiA concepts predictable from any tag in X. Every cell in the output layer $\vec{y}$ thus corresponds to an assignment of a unique concept from $T = T_{ptb} \cup T_{susa}$. For a given word w_i with PTB annotation and its concept set $s \subseteq T_{ptb}$, every output node y_k with $k \in \{1..|T|\}$ is assigned as follows:

$$y_k = \begin{cases} 1, & \text{if }, t_k \in s \\ 0, & \text{if}, t_k \in T \setminus T_{ptb} \\ -1, & \text{if}, t_k \in T_{ptb} \setminus s \end{cases}$$

For, say, training on EWT, all the output values that corresponded to the concepts generated only from Susa (e.g., *DefiniteArticle* and *Article* from the example above) are thus set to 0.

Training against this data is a regression problem whose application of the neural networks to the unseen data will produce values normalized between -1 and 1 for every output node y_n, resp., its associated *MorphosyntacticCategory* concept.

5 Decoding the Output Layer

For decoding the output vectors produced by the neural network, we interpret the value of an output node y_n as a confidence score for the associated concept, with positive scores indicating high probabilities, lower scores indicating low probability (or underspecification in/lack of evidence from the training data) and negative scores indicating counter-evidence for the corresponding triple.

[3] Note that we aim to study whether neural classifiers trained over different corpora – which will show an increase in coverage (or annotation granularity) by design – will suffer in their precision. This research question is independent from the dimensionality of the embeddings, so that we chose the minimal embeddings available from `http://metaoptimize.com/projects/wordreprs`. With higher-dimensional embeddings, better results are likely to be obtained.

These scores provide a *ranking* of concepts which forms the basis to decode an output vector into a set of OLiA triples (concept assignments).

In an ideal world, the ontology provides us with consistency constraints, e.g., regarding the disjointness of two classes. At present, however, no publicly available ontology of linguistic annotation is fully axiomatized. Therefore, we employ and evaluate pruning heuristics to infer consistency axioms: *structural (path) pruning* (exploiting the hierarchical structure of the ontology), and two variants of *corpus pruning* (exploiting concept combinations observed in the training set).

5.1 Structural (Path) Pruning

In an ontology, conecpt assignments are dependent on each other: assigning class C necessarily entails assigning of its superclass C'. From all concepts with positive activation, we calculate the set P of all possible paths along the ancestor (superclass) axis in the ontology, represented as a list, e.g., $p_1 = \langle$*Determiner, PronounOrDeterminer*$\rangle$ for the PTB tag `DT`.

This set is reduced by *eliminating partial* paths: If any path p is a sublist of another path q, it is removed from P. For example, p_1 is a sublist of the path $p_2 = \langle$*DefiniteArticle, Article, Determiner, PronounOrDeterminer*$\rangle$ (for Susa `AT`) and thus to be removed if p_2 is a possible solution.

From the reduced set of non-redundant, and maximal paths P', we select the path with the highest confidence, i.e.,

$$p_{best} = \arg\max_{p \in P'} \left(\frac{\sum_{n=0}^{|p|} y_{p(n)}}{|p|} \right)$$

Here, $y_{p(n)}$ is the activation of the output neuron y_i that corresponds to the nth element in the path p. In order to prevent any bias towards longer paths, the sum of activation scores is divided by the length of the path $|p|$. Concepts that are compatible with the path but have values less than 0 (= negative evidence) are skipped.

Path-based pruning follows Chiarcos (2010b) who also assumed that classes along the subclass-superclass axis are compatible with each other, whereas siblings (and their descendants) are incompatible.

5.2 Corpus Pruning

As an alternative to structural pruning, we estimate path consistency directly out of the tags of the training corpus: Given a particular training corpus, we consider any pair of concepts compatible with each other for which co-occurrence is observed. For well-attested, frequent concepts, this is a very elegant way to enable an assignment to multiple classes. For example, an adjectival participle like *enduring* in the example above is analyzed as a verb in Susanne (`VBD`, concepts {*Ing, Participle, NonFiniteVerb, Verb*}), while in PTB, it is analyzed as an adjective (`JJ`, concepts {*Adjective*}). With a corpus having both Susa and PTB annotations, such systematic double analyses can be observed and thus, tolerated, but would be ruled out by structural Pruning.

A drawback of this method is that concepts not sufficiently attested in the training corpus may be regarded incompatible with other tags – although their occurrence would be possible, they were just too rare to be observed in the training set.

With two heterogeneously annotated corpora, we employ two variants of corpus pruning: *Disjoint corpus pruning* on each corpus individually, and *joint corpus pruning* on the merged annotations of texts in the intersection of both corpora.

With the disjoint corpus pruning strategy, concepts generated by either tagset A or B are compatible with each other if they co-occur in A- *or* B-annotations, any concept generated only by tagset A (or B) is compatible with every concept generated only by tagset B (resp., A).

This strategy may be too permissive, so that if A- and B-annotations for the same stretch of text are available (or can be produced using automatic tools), we merge the triple sets for every word before the corpus pruning routine applies. By doing so, we are able to learn that systematic correspondences between Susa *Participle* and PTB *Adjective* exist. This joint corpus pruning strategy, however, presupposes that a considerable body of text is annotated according to both schemes, a situation that, fortunately, we face for the intersection of PTB and Susanne (PTB$\cup$Susa referring to the Susanne corpus with both annotations merged).

6 Experimental Results

Three neural networks were trained on the training sets: EWT/PTB data only, Susanne/Susa data only, and both training sets combined. Several state-of-the-art POS taggers have been trained on this data as baseline: TreeTagger (Schmid, 1999), Lapos (Tsuruoka et al., 2011) and Stanford

(Toutanova et al., 2003), all trained and tested on the same (non-combined) data as the neural networks.

Training these on PTB annotations was straightforward. On Susa, however, TreeTagger could not accomodate 270 unique tags and was thus skipped, and Lapos could be trained but showed very low performance on the full tagset. The Stanford tagger was successfully trained using state-of-the-art MaxEnt (left3words) models for EWT and Susanne, respectively.

Like the training data for the neural network, the output of each tagger was mapped to OLiA Reference Model concepts by means of the corresponding Annotation and Linking Models. This is the basis for comparative evaluation with the neural networks.

tagset	corpus/	coverage	
	tool	%concepts	%triples
PTB	EWT	64.9% (50)	81.2%
Susa	Susanne	85.7% (66)	85.7%
PTB∪Susa	NN:Combined	100% (77)	100%

Table 2: Evaluation: Coverage/granularity
%concepts: number of predictable concepts per tagset, relative to the number of concepts predictable from PTB∪Susa
%triples: number of NN:Combined-predicted triples interpretable against the gold tagset

Table 2 shows how NN:Combined yields a gain of informativity in comparison to the original annotations (and tools trained on that basis). Neither of both original tagsets is a proper subset of (the ontological representation of) the other one (%concepts), and accordingly, NN:Combined (with structural pruning) predicts *more triples* than can actually be evaluated against the gold annotation (1-%triples). We refer to this evaluation metric as *(OLiA) concept coverage*.

While NN:Combined trivially a gain in concept coverage over tag-based tools by design, this is logically independent from accuracy, and it may be suspected that training over heterogeneous annotations adds additional noise. Yet, as we eventually observed, it reaches the precision of state-of-the-art string-based POS taggers.

In order to evaluate this aspect, we employ two precision metrics. *Concept precision* is calculated in the conventional way with the following definitions: A predicted concept is a true positive if also generated from the gold annotation, e.g., *Noun* from both predicted tag NNP and observed tag NN. Otherwise, it is a false positive, e.g., *ProperNoun* from predicted NNP but not from observed NN (common noun).

For *path precision*, a path is considered to be a true positive only if *all* the concepts in the path are also generated from the gold tag. In the example above, the predicted tag NNP yields the path ⟨*ProperNoun, Noun*⟩, while the gold tag NN yields ⟨*CommonNoun, Noun*⟩, hence, a false positive. For conventional taggers, path precision corresponds to standard tag precision.

As shown in Tab. 2, Susa generates 66 unique concepts while 50 concepts are generated by PTB, the union of both is 77 unique concepts. To calculate concept and path precision for tag set-specific taggers (Tab. 3), concepts not predictable by the gold data are excluded from the evaluation. Thus, 18.8% of the concepts predicted by NN:Combined for the EWT test set and 14.3% predicted for the Susanne test set are ignored in the evaluation, as they could not have been generated from the original gold annotation, but only from the 'other' tag set (Tab. 2) Yet, the precision of these 'alien' concepts can evaluated on the (test set of the) PTB/Susanne intersection with double annotations (PTB∪Susa). The gold data in the test set is the union of PTB and Susa triples for the same word.

Table 3 provides overall evaluation results for the conventional taggers as well as the different neural network configurations in terms of concept and path precision on triple-represented annotations of EWT, Susanne and the merged PTB-Susanne annotations on the PTB∪Susa test set.

In general, path precision is lower than concept precision (Tab. 3). A likely reason is that tagging errors tend to occur between related POS. For example, proper nouns are frequently erroneously tagged as common nouns, but concept precision still rewards the common superconcept. Thus, the higher the granularity of a tag set, the greater the discrepancy between path and concept precision. The neural network trained only on EWT achieves the best path precision on the EWT test set, outperforming Lapos by almost 3%. The neural network trained only on Susanne outperforms the Stanford tagger both by path and concept precision. The neural network trained on both Susanne and EWT fell slightly short of the best tagger in path and concept precision on EWT, but still outperforms the best tagger (Stanford) on the Susanne test set. Furthermore, concept precision of the combined neural network on the Susanne data is only 0.3%

lower than the precision of the neural network trained on Susanne only.

Statistics over the most frequent[4] false predictions are given in Tab. 4. The first column of Tab. 4 contains the gold concept, the second column the predicted concept, the third column is the error $e_{g,t}$ for the concept pair $\langle g, t \rangle$, counted as

$$e_{g,t} = \frac{\text{freq}(concept_g, concept_t)}{\text{freq}(concept_g)}$$

The fourth column of Tab. 4 shows the contribution of $e_{g,t}$ to the total error.

For NN:Combined, the key result is that we achieve a substantial increase in coverage (18.8%, resp. 14.3%, Tab. 2) while facing only a marginal drop of precision (around 1%, Tab. 3) between individually trained neural networks and NN:Combined. The precision neural network predictions against individual corpora remains constantly high, and also for the merged test set. Furthermore, neural networks in any configuration reach state-of-the-art tagger performance; neural networks with structural pruning even outperform it, for both path and concept precision.

Tab. 3 shows little – if any – decay of precision if the neural network is trained over heterogeneous annotations of different corpora: In comparison to the best-performing conventional tagger considered (Lapos), NN:Combined (with structural pruning) loses 0.2% in path precision and 0.6% in concept precision, but yields a gain of 18.8%, resp. 14.3% in coverage.

To our surprise we found that structural pruning – which we initially regarded as being too restrictive – outperforms other decoding strategies, whereas joint corpus pruning showed the lowest precision. One reason is probably that not all deviations in annotation were eventually compatible, but that some of those mismatches were actual tagging errors, thus propagated into the neural learning algorithm. Such original annotation errors in the linguistic analyses are possibly the main reason why the performance of the combined network is slightly lower than the performance of networks trained on homogeneous data. The disjoint corpus pruning suffered less from annotation inconsistency, but its poor performance can probably be attributed to sparsity issues, i.e., rarely attested concept were incorrectly regarded as inconsistent with possible other concepts.

$concept_g$	$concept_t$	$e_{g,t}$	total(e)
ProperNoun	*CommonNoun*	30.2%	1.8%
ProperNoun	*Adjective*	16.3%	1%
AuxiliaryVerb	*Indicative(Full)Verb*	8.2%	2.5%
AuxiliaryVerb	*Finite(Full)Verb*	8.2%	2.5%
Participle	*Adjective*	5.8%	3.6%
PersReflPronoun	*DemonstrativeDeterminer*	5.8%	5.6%
PersonalPronoun	*DemonstrativeDeterminer*	21.1%	5.4%

Table 4: Confusion matrix
$concept_g$ are gold standard concepts, ordered by their percentage of the total error $total(e)$. $e_{g,t}$ is a relative count for $concept_g$ erroneously predicted as $concept_t$ to the total count of $concept_g$ predictions.

It should be noted that our NN setting was deliberately minimalistic: We used minimal context information with the smallest-dimensional word embeddings available, and trivial backpropagation without employing any more advanced procedures to improve convergency properties (e.g., deep learning). Also, we did not optimize hyperparameters but followed a simple geometric (pyramidal) structure for their initial assessment. Despite the lack of any such optimization, we were nevertheless able to prove an increase in coverage while maintaining state-of-the-art precision, thereby proving the feasibility and the potential of ontology-based neural learning over multiple heterogeneously annotated corpora.

7 Discussion and Outlook

We presented an ontology-based neural network approach to POS tagging, or, more precisely, predicting morphosyntactic categories underlying part-of-speech annotation.

Unlike other approaches trying to generalize over heterogeneously annotated corpora (Sect. 3.2), our approach is informationally *lossless*. The usefulness of such approach is obvious when dealing with heterogeneous annotations with different granularity. But also comparably-designed annotation schemes can differ in their use of apparently identical categories: POS tag semantics conflate different criteria from morphology, syntax, semantics and lexicon, respectively, but at the same time enforce categories (tags) to be disjoint. As for attributive possessive pronouns, for example, these are both pronouns (semantically) and determiners (syntactically). (Other examples for English are numerals vs. determiners, participles vs. adjectives, subordinating conjunctions vs. prepositions, various functions of TO, lexical vs. syntactic definition of auxiliary verbs, etc., so this is really

[4] concept frequency >1000, excluding punctuation

	path precision			concept precision		
	EWT	Susanne	PTB∪Susa	EWT	Susanne	PTB∪Susa
baseline taggers						
TreeTagger	77.3%	-	-	85.6%	-	-
Lapos	92.0%	16.9%	-	**95.4%**	31.0%	-
Stanford	91.4%	82.5%	-	94.8%	89.4%	-
disjoint corpus pruning						
NN:EWT Only	93.4%	-	-	95.0%	-	-
NN:Susanne Only	-	88.7%	-	-	91.4%	-
NN:Combined	91.9%	87.0%	82.1%	94.7%	90.6%	89.9%
joint corpus pruning						
NN:EWT Only	92.1%	-	-	93.9%	-	-
NN:Susanne Only	-	87.5%	-	-	90.3%	-
NN:Combined	91.2%	86.2%	76.5%	94.3%	90.0%	86.7%
structural (path) pruning						
NN:EWT Only	**94.9%**	-	-	95.2%	-	-
NN:Susanne Only	-	**90.1%**	-	-	**91.8%**	-
NN:Combined	91.8%	88.7%	**86.3%**	94.8%	91.5%	**91.4%**

Table 3: Evaluation: Path and concept precision

wide-spread even for English as the “prototypical” NLP language.) Tagset designers do not have the expressive means to state if categories overlap, so an ad hoc decision has to be made, thus naturally leading to incompatibilities between tagsets both cross-lingually and monolingually.

Using an ontology, no implicit disjointness criterion applies, but instead, every tag can be decomposed into a set of triples. This has been elaborated before by Chiarcos (2008b) and Chiarcos and Erjavec (2011). In our setting, we learn concept (and feature) assignments for every possible statement independently (and simultaneously), together with a confidence score (activation of the output layer=, and then employ *pruning* strategies to extract ontologically consistent descriptions of maximum granularity and confidence. This approach does not only guarantee consistent results, but it also is way more flexible than any string-based annotation and tools trained on that basis, whereas tags – given the likely sources of deviation in the use and interpretation of near-equivalent categories mentioned above – represent more or less opaque bundles of features.

Moreover, this allows us to combine the advantages of coarse-grained tagsets (more training data, robust categories) and fine-grained tagsets (fine-grained categories and features, but less reliably trainable on limited amounts of data). More general concepts and features higher in the hierarchy occur more frequently, and like in a small tagset that can be more robustly trained against limited training data, these can be reliably learned. Using a confidence-based ranking, this means that these concepts are *first* selected during the pruning. That is, more general concepts/features guide the choice among more fine-grained concepts/features (whose reliability is likely to improve as a result).

Also, this was an experiment in preparation for research on low-resource languages: By using pretrained word embeddings as input vectors, we reduced the need for large POS-annotated corpora, and achieved state-of-the-art results even limited amounts of labeled training data. This scenario particularly beneficial for less-researched major languages such as Hausa or Farsi for which only sparse data annotated with different tagsets is available, but where it is rather unproblematic to acquire large amounts of unannotated texts (e.g. by web crawling) to compute word vector representations.

Our findings indicate the viability of ontological models for part of speech tags: Even with overly restrictive consistency constraints applied, these guarantee consistent results. Future research will focus on optimizing parameters and explore applications of this technique to less-resourced languages and cross-lingual applications: The OLiA ontologies employed here are both cross-lingual and cross-tagset, and therefore, our monolingual use case can be easily extended to multi-lingual scenarios projection, where the output of annotations originating from difference source languages is to be combined.

References

E. Buyko, C. Chiarcos, and A. Pareja-Lora. 2008. Ontology-based interface specifications for a NLP pipeline architecture. In *Proc. LREC 2008*, Marrakech, Morocco.

C. Chiarcos and T. Erjavec. 2011. OWL/DL formalization of the MULTEXT-East morphosyntactic specifications. In *Proc. 5th Linguistic Annotation Workshop, held in conjunction with ACL-HTL 2011*, pages 11–20, Portland, June.

C. Chiarcos. 2008a. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Christian Chiarcos. 2008b. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.

Christian Chiarcos. 2010a. Towards robust multi-tool tagging. An OWL/DL-based approach. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pages 659–670, Uppsala, Sweden.

Christian Chiarcos. 2010b. Towards robust multi-tool tagging. an owl/dl-based approach. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 659–670. Association for Computational Linguistics.

T. Erjavec. 2004. MULTEXT-East version 3. In *Proc. LREC 2004*, pages 1535–1538, Lisboa, Portugal.

S. Farrar and D.T. Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International*, 7(3):97–100.

S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. 2013. Integrating NLP using Linked Data. In *Proc. International Semantic Web Conference (ISWC-2013)*, pages 98–113, Heraklion, Crete. Springer.

S. Hellmann. 2010. The semantic gap of formalized meaning. In *Proc. 7th Extended Semantic Web Conference (ESWC 2010)*, Heraklion, Greece.

N. Ide and L. Romary. 2004. A registry of standard data categories for linguistic annotation. In *Proc. LREC 2004*, pages 135–39, Lisboa, Portugal.

Geoffrey Leech and Andrew Wilson. 1996. EAGLES guidelines: Recommendations for the morphosyntactic annotation of corpora.

A. Pareja-Lora and G. Aguado de Cea. 2010. Ontology-based interoperation of linguistic tools for an improved lemma annotation in Spanish. In *Proc. LREC 2010*, Valetta, Malta.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey.

G. Rehm, R. Eckart, and C. Chiarcos. 2007. An OWL- and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. RANLP 2007*, Borovets, Bulgaria.

G. Sampson. 1995. *English for the computer: The SUSANNE corpus and analytic scheme*. Oxford University Press.

A. Saulwick, M. Windhouwer, A. Dimitriadis, and R. Goedemans. 2005. Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proc. 17th Conf. on Advanced Information Systems Engineering (CAiSE'05)*, Porto.

H. Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 13–25. Springer Netherlands.

T. Schmidt, C. Chiarcos, T. Lehmberg, et al. 2006. Avoiding data graveyards. In *Proc. E-MELD Workshop 2006*, Ypsilanti.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for english. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Ann Taylor, Mitchell Marcus, and Beatrice Santorini. 2003. The Penn Treebank: An overview. In Anne Abeill, editor, *Treebanks*, pages 5–22. Springer, Dordrecht.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Kazama. 2011. Learning with lookahead: Can history-based models rival globally optimized models? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 238–246, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, Uppsala, Sweden, July. Association for Computational Linguistics.

D. Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proc. LREC 2008*, Marrakech, Morocco.