A Grounded Unsupervised Universal Part-of-Speech Tagger for Low-Resource Languages

Ronald Cardenas[♣] **Ying Lin**[♠] **Heng Ji**[♠] **Jonathan May**[♡]
[♣] Institute of Formal and Applied Linguistics, Charles University in Prague
[♠] Computer Science Department, Rensselaer Polytechnic Institute
[♡] Information Sciences Institute, University of Southern California
ronald.cardenas@matfyz.cz liny9@rpi.edu
jih@rpi.edu jonmay@isi.edu

Abstract

Unsupervised part of speech (POS) tagging is often framed as a clustering problem, but practical taggers need to *ground* their clusters as well. Grounding generally requires reference labeled data, a luxury a low-resource language might not have. In this work, we describe an approach for low-resource unsupervised POS tagging that yields fully grounded output and requires no labeled training data. We find the classic method of Brown et al. (1992) clusters well in our use case and employ a decipherment-based approach to grounding. This approach presumes a sequence of cluster IDs is a 'ciphertext' and seeks a POS tag-to-cluster ID mapping that will reveal the POS sequence. We show intrinsically that, despite the difficulty of the task, we obtain reasonable performance across a variety of languages. We also show extrinsically that incorporating our POS tagger into a name tagger leads to state-of-the-art tagging performance in Sinhalese and Kinyarwanda, two languages with nearly no labeled POS data available. We further demonstrate our tagger's utility by incorporating it into a true 'zero-resource' variant of the MALOPA (Ammar et al., 2016) dependency parser model that removes the current reliance on multilingual resources and gold POS tags for new languages. Experiments show that including our tagger makes up much of the accuracy lost when gold POS tags are unavailable.

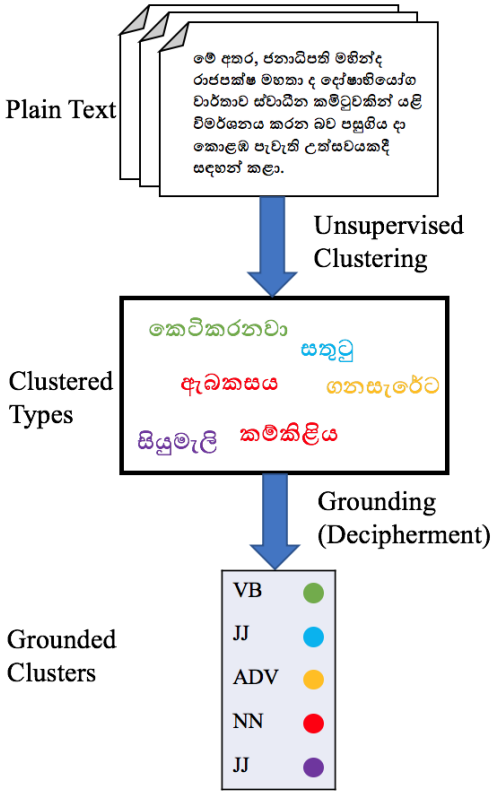

Figure 1: Overview of our approach to grounded POS tagging. We use an unsupervised clustering method (Section 3.2) then reduce and ground the clusters using a decipherment approach informed by POS tag sequence data from many languages (Section 3.3).

1 Introduction

While cellular, satellite, and hardware advances have ensured that sophisticated NLP technology can reach all corners of the earth, the language barrier upon reaching remote locales still remains. As an example, when international aid organizations respond to new disasters, they are often unable to deploy technology to understand local reports detailing specific events (Munro and Manning, 2012; Lewis et al., 2011). An inability to communicate with partner governments or civilian populations in a timely manner leads to preventable casualties.

The lack of adequate labeled training data has been the major obstacle to expanding NLP's outreach more multilingually. Developments in unsupervised techniques that require only monolingual corpora (Lample et al., 2018a; Artetxe et al., 2018) and the ability to leverage labeled resources in other languages have been proposed to address this issue (Das and Petrov, 2011; Duong et al., 2014; Ammar et al., 2016). Unfortunately, these

methods either do not work in practice on true low-resource cases or unrealistically assume the availability of some amount of supervision.

Consider syntactic parsing as a prime example. Past editions of the CoNLL Shared Task on Multilingual Parsing (Zeman et al., 2017, 2018) featured a category of target languages for which either little or no training data was provided. However, even in the 'no-resource' scenario that most closely matches our use case, gold part-of-speech (POS) tags for test data were provided for the participants to use. Prior to these shared tasks, Ammar et al. (2016) proposed a variant of their main model, MALOPA, that was meant to produce reasonable parses for languages under "zero-resource" conditions. In order to function, however, the model requires users to provide gold POS tags and word mappings from these languages into a common semantic space, using approaches that require parallel data (Guo et al., 2015).

Indeed, the compulsion to use POS tag-labeled data in zero-resource circumstances extends to the vast, varied lines of research in unsupervised POS tagging itself! Every approach explored so far ultimately requires POS-annotated resources for the language being studied in order to produce a final, grounded output. Even the most conservative strategies (Goldwater and Griffiths, 2007; Berg-Kirkpatrick et al., 2010; Stratos et al., 2016) that do not require any supervised signal during training still ultimately produce only ungrounded clusters, and require a reference annotated corpus to map the inferred clusters or states to actual POS tags.

Making matters worse, evaluation is generally offered in terms of the 'many-to-one' or 'one-to-one' analyses Johnson (2007). These metrics use a reference corpus to determine the *optimal* mapping of clusters to tags. While this evaluation approach is intuitively sensible for measuring cluster purity, to actually *use* such an output, an entire annotated training corpus is required.[1] It is not enough to simply rely on ungrounded clusters in real-world systems; grounded labels offer a sort of universal API between other resources such as rule-based modules that operate on certain word types or between resources built from other annotated high-resource language data.

Since POS tag and parallel data resources for new languages are often unavailable or unreliable, we make the following contributions to ensure the surprise of a new language does not immobilize us:

- We introduce a decipherment-based approach to POS grounding, which yields fully grounded output and does not require any annotated data or parallel corpora in the language to be analyzed. The approach uses pre-existing human-labeled POS tag sequences from high-resource *parent languages* (PL) but no labeled data or sequences for the target, or *child language* (CL). An overview of the approach is shown in Figure 1.
- We demonstrate our approach by evaluating over a variety of languages spanning 4 families and 8 genera (Germanic, Romance, Slavic, Japanese, Semitic, Iranian, Indic, and Bantoid), and show across-the-board reasonable intrinsic performance, given the difficulty of the task and the stringency (straightforward accuracy) in comparison to other unsupervised evaluation strategies.
- We test the utility of our grounded tags in a name tagging task, obtaining state-of-the-art performance for Sinhalese and Kiryarwanda, two languages with nearly no labeled POS or named entity resources.
- We further pare down the annotated resources required in an existing 'zero-resource' dependency parser model and show that our unsupervised and grounded tags are helpful at closing the gap between a nihilistic tag-free setting and an unrealistic gold tag setting.
- We release our code so that others may create zero-resource syntactic analysis and information extraction systems at the onset of the next new emergency.[2]

2 POS Grounding as Decipherment

We consider the task of POS induction as a two-step pipeline: from word sequence w to POS tag sequence p via cluster sequence c. Formally, our conditional probability model is

[1] Additionally, Headden III et al. (2008) demonstrated that these metrics are not indicative of downstream performance.

[2] https://github.com/isi-nlp/universal-cipher-pos-tagging.git

$$\begin{aligned}
&\operatorname*{argmax}_p P_\theta(p|w) \\
&\quad = \operatorname*{argmax}_p \sum_{c \in C^{|w|}} P_\theta(p, c|w) \\
&\quad = \operatorname*{argmax}_p \sum_{c \in C^{|w|}} P_\theta(p|c, w) P_\theta(c|w)
\end{aligned}$$

where C is the cluster vocabulary and θ parameterizes our probability model. If we assume a deterministic pipelined clustering of words and a tag labeling model that does not depend on words, then for chosen $\hat{c}$, this becomes

$$\begin{aligned}
&\operatorname*{argmax}_p \sum_{c \in C^{|w|}} P_\theta(p|c, w) P_\theta(c|w) \\
&\quad = \operatorname*{argmax}_p P_\theta(p|\hat{c}) \\
&\quad = \operatorname*{argmax}_p P_\theta(\hat{c}|p) P_\theta(p) \qquad (1)
\end{aligned}$$

We call this model the *cipher grounder*. As presented it requires an estimate for $P_\theta(p)$ for the CL, which requires POS training data. Under the zero-resource scenario, we instead approximate $P_\theta(p)$ by the tag distribution of a PL. Then, the *cipher table* $P_\theta(\hat{c}|p)$ can be trained using a noisy-channel, expectation-maximization (EM)-based approach as in Ravi and Knight (2011).

3 POS Tagger construction

We approach the search for optimal components in the two-step pipeline outlined in Section 2 in a cascaded manner. First, an optimal word clustering is determined by means of the many-to-one evaluation method. This method is explained well by Johnson (2007):

> " ...deterministically map each hidden state to the POS tag it co-occurs most frequently with, and return the proportion of the resulting POS tags that are the same as the POS tags of the gold-standard corpus."

While unrealistic for POS tagger performance purposes, many-to-one is a good choice for determining cluster 'purity' and provides a reasonable grounding upper bound. As the calculation of many-to-one does require labeled data, we constrain the use of these labels for development and will evaluate extrinsically using languages for which we do not have any training data; see Section 5.2.

Secondly, we search for the best approach to ground the chosen clusters, given several possible PL options.

After the optimal components and parameters are determined, we validate POS tag quality intrinsically via tag accuracy on reference data where it exists, and then extrinsically on two downstream tasks. We investigate a simulated no-resource scenarios in the task of dependency parsing, and a real low-resource scenario in name tagging.

3.1 Datasets

For intrinsic evaluation and optimization of the tagging pipeline, including all preliminary experiments, we use annotated corpora from Universal Dependencies (UD) v2.2[3] for the following languages: English (en), German (de), French (fr), Italian (it), Spanish (es), Japanese (ja), Czech (cs), Russian (ru), Arabic (ar), and Farsi (fa). For Swahili (sw), we use the Helsinki Corpus of Swahili 2.0.[4] Overall in these experiments we cover 11 languages and 4 language families.

In our dependency parsing experiments, we use the Universal Treebank v2.0 (McDonald et al., 2013) for en, de, fr, es, it, Portuguese (pt), and Swedish (sv). This set of treebanks is chosen instead of UD in order to obtain results comparable to those of previous work on simulated zero-resource parsing scenarios (Ammar et al., 2016; Zhang and Barzilay, 2015; Rasooli and Collins, 2015).

In our name tagging experiments, we use monolingual texts for Sinhalese (si) and Kinyarwanda (rw) provided by DARPA's Low Resource Languages for Emergent Incidents (LORELEI) Program during the 2018 Low Resource Human Languages Technologies (LoReHLT) evaluation.

3.2 Unsupervised Clustering

In this step we compare two approaches to unsupervised ungrounded labeling. The first strategy is to cluster by word types and thus label each token with its cluster ID independently of its context.[5] We consider Brown's hierarchical clustering algo-

[3] http://universaldependencies.org/

[4] http://urn.fi/urn:nbn:fi:lb-2016011301

[5] We refer to ungrounded POS tag labels as 'clusters' even though not all methods induce a clustering.

rithm, (Brown et al., 1992)[6] BROWN; Brown's exchange algorithm,[7] (Martin et al., 1998) MARLIN; and k-means clustering of monolingual word embeddings of dimension size 100, trained using *fastText* (Joulin et al., 2016), E-KMEANS. The second labeling strategy is context-sensitive; it uses the Bayesian HMM tagger proposed by Stratos et al. (2016), which we call A-HMM. As noted previously, we evaluate unsupervised labeling extrinsically, via the many-to-one approach, and use the best performing labeling in the complete two-step grounded tagging pipeline.

In preliminary experiments, we vary the number of clusters and hidden states ($|C|$) between 17 and 500. We initially sought to create one cluster per UD POS tag and then choose the proper 1:1 assignment of cluster to tag, following the approach of Stratos et al. (2016). However, cluster purity is low when only 17 clusters are allowed (i.e. each cluster has words with a variety of POS tags). Naturally, as the number of clusters is raised, the purity of each cluster improves. We ultimately fix the cluster limit at 500, which gives a good tradeoff between overall cluster quality for all the ungrounded tagging methods, and size small enough to allow EM-based decipherment to be tractable.

Given this setting, we evaluate our four labeling strategies using the many-to-one approach, as presented in Table 1. Due to the larger number of clusters, the results presented here are higher than and not comparable to the original literature describing the methods.[8] We can, nevertheless, make relative judgements. In all cases, clustering by type with Brown-based algorithms works better than using a sophisticated tagger such as A-HMM. Since BROWN and MARLIN obtain similar results, with no consistently dominant model, in all subsequent experiments we use the BROWN labeler with 500 clusters.

3.3 Grounding via Decipherment

We now seek an appropriate method for grounding the clusters generated in Section 3.2. We experiment with en, fr, fa, and sw as CLs. For each CL t, we instantiate our model following Equation 1, using the Carmel toolkit (Graehl, 1997) and forming the cipher table as a one-state transducer. We train these models using EM for 500 iterations or until convergence, and we select the model with the lowest perplexity from among 70 random restarts.

Yet unspecified is the nature of the POS language model $P_\theta(p)$. We begin by training bigram models of POS tag sequences with additive smoothing using the SRILM toolkit (Stolcke, 2002) for each PL $s \in \mathcal{S} = \{$en, de, fr, it, es, ja, ar, cs, ru, sw$\}$. But which PL's POS tag data to use for each CL? We explore two initial criteria for choosing a single suitable PL s: confidence of the model during decoding (perplexity, PPL), and typological similarity. For the first criterion, the PL whose cipher grounder s-t yields the better performance is chosen. For the second criterion, the most similar language to CL t is chosen according to the cosine similarity between typological features vectors. We employ 102 features obtained from WALS[9] related to word order and morphosyntactic alignment, further reduced to 50 dimensions using PCA. However, none these criteria correlates significantly to tagging accuracy, as we elaborate in Section 5.1. We instead try a combined approach.

The likelihood of cluster ID replacement, $P_\theta(\hat{c}_i|p_j), \forall \hat{c}_i \in C, \forall p_j$ in the tagset, is replaced by

$$P_{avg}(\hat{c}_i|p_j) \sim \frac{\sum_{s \in \mathcal{S}, s \neq t} P_\theta(\hat{c}_i|p_j^s)}{|S| - 1}$$

where $P_\theta(\hat{c}_i|p_j^s)$ is the likelihood of POS tag p_j being represented by cluster $\hat{c}_i$ after training with the language s tag distribution. Note that the CL is excluded from $\mathcal{S}$ for the combination. The combined cipher grounder is then defined by

$$\operatorname*{argmax}_p P_{all}(p) P_{avg}(\hat{c}|p) \qquad (2)$$

where $P_{all}(p)$ is a language model trained over the concatenation of POS sequences of all parent languages in $\mathcal{S}$. We call this approach CIPHER-AVG.

4 Downstream Tasks

4.1 Name Tagging

We experiment with the LSTM-CNN model proposed by Chiu and Nichols (2016), one of the

[6] https://github.com/percyliang/brown-cluster

[7] Optimized and implemented by Müller and Schuetze (2015). Available at http://cistern.cis.lmu.de/marlin/

[8] As noted by Clark (2003) and Johnson (2007), in the limit, keeping each type (or, in the case of A-HMM, TOKEN in its own cluster will result in the maximum possible many-to-one (polysemic types prevent perfect accuracy when type clustering).

[9] https://wals.info/

Seq. Tagger	en	de	fr	ru	fa	sw
BROWN	81.37	**81.28**	84.81	**79.78**	**86.94**	87.35
MARLIN	**81.53**	81.25	**85.4**	79.14	86.64	**88.81**
A-HMM	77.12	74.85	81.48	73.88	80.25	76.69
E-KMEANS	63.01	65.14	68.68	70.80	76.94	65.08

Table 1: Comparison of labeling strategies using many-to-one mapping for target languages with available test data, using 500 clusters or number of states. Accuracy is shown in percentage points.

state-of-the-art name tagging models, as our baseline model. To incorporate POS features, we extend the token representation (word and character embeddings) with a one-hot vector representation of the POS tag. Figure 2 presents an outline of the architecture.

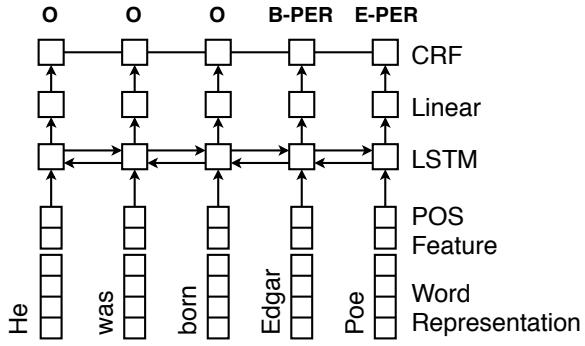

Figure 2: Name tagging model evaluated in Section 5.2. This is an extension of the model of Chiu and Nichols (2016) with POS tag features added.

4.2 Multilingual Dependency Parsing

We base our experiments on the no-treebank setup of MALOPA (Ammar et al., 2016), but change the underlying transition-based parser to the graph-based parser proposed by Dozat and Manning (2017) for implementation convenience. Following this setup, for each CL except en, we train the parser on the concatenation of treebanks of the other 6 languages as PLs.

The original MALOPA work enriches the input representation by concatenating pretrained multilingual word embeddings (Guo et al., 2016), multilingual Brown cluster IDs, and POS tag information. However, these representations are obtained using parallel corpora and gold POS tags are required for parsing at test time.

In contrast, we are interested in the realistic scenario in which no resource is available in the child language but raw text. It is important to note, however, that our objective is not to beat the state-of-the-art on this benchmark but to investigate parsing performance fluctuation when cross-lingual components (gold POS annotations and supervised multilingual embeddings) are replaced by those obtained in an unsupervised manner.

We investigate the following variations to each component of the input representation.

- **Multilingual word and cluster embeddings.** The original work of Ammar et al. (2016) uses 'robustly projected' pre-trained embeddings (Guo et al., 2015) for word embeddings and embeddings learned from English Brown cluster IDs projected through word alignments (Guo et al., 2016) for cluster embeddings; both of these rely on parallel data and we refer to them collectively as GUO. We replace these with monolingual *fastText* embeddings (Bojanowski et al., 2017) projected to a common space using MUSE, the unsupervised method of Lample et al. (2018b). For cluster embeddings we start with *fastText* monolingual embeddings trained over Brown cluster ID sequences instead of word tokens ($|C| = 256$, the same as in Guo et al. (2016)). Then, unsupervised multilingual embeddings are derived, again using MUSE.[10] Note that this approach, which we refer to collectively as MUSE, requires no parallel data. We compare both MUSE and GUO approaches in Section 5.2 and Table 5.
- **POS tag scheme.** The original work uses gold POS tag data at both train and test time. While realistic to have gold POS info from PLs for training, it is unrealistic to have this data available for new CLs at test time. We thus compare the original GOLD scenario with the realistic CIPHER scenario, where the training data is still gold, but the test POS tags use the method presented in this work. Another realistic scenario dispenses

[10] Both cluster and word MUSE embeddings are projected to the corresponding English space.

with POS disambiguation except for the trivial distinction of punctuation; for compatibility purposes this is done in both train and test data and is labeled NONE.

We investigate all combinations of {GUO, MUSE}-{GOLD, CIPHER, NONE}.

5 Results and Discussion

5.1 Labeling and Cipher Grounding

The results in Table 1 are somewhat at odds with those presented in Stratos et al. (2016), but these are done at different operating points; we use different data, the UD-17 tag set instead of the Universal Treebank 12 tag set, and, perhaps most importantly, generate more clusters. We further note that to some degree, choosing Brown clusters based on the results in Table 1 compromises claims of our approach being fully 'unsupervised' for those six languages, however our subsequent experiments on additional languages are truly unsupervised.

Table 2 presents the intrinsic performance of the cipher grounder over all PL-CL pairs considered. The difference between the best and the worst performing PL for each CL ranges from 24.62 percentage points for Swahili to 48.34 points for French, and an average difference of 34.5 points among all languages. The case when PL=CL is also presented in Table 2 as a reference and provides a reliable upper-bound under zero-resource conditions. It is worth noting the difference in accuracy when comparing the best performing PL for each CL with its corresponding PL=CL upper-bound. Among all CLs, the best cipher grounder for French (es-fr) gets the closest to its upper-bound with just 4.81 percentage points of difference, followed by the English grounder (fr-en) with 13.53 points of difference. On the other hand, the best Swahili grounder (ar-sw) is the most distant from its upper-bound with 30.45 points of difference.

Given such wide performance gaps in the CL set, the choice of a suitable PL becomes crucial for performance; therein the cipher model confidence and typological similarity are explored as possible choice criteria. With regards to model confidence, the Pearson correlation between accuracy scores and PPL, expected to be negative, ranges from -0.71 for English to 0.40 for Farsi. Since the PPL values for different PLs are not comparable, we first z-normalize PPL per CL and then concatenate the results for all CLs. The Pearson correlation of the resulting PPL-accuracy values is -0.13. This last result indicates that the most confident model might not be the most accurate, hence this criterion is not suitable for choosing a suitable PL.

With regards to typological similarity, we find that the Pearson correlation between accuracy scores and cosine similarity of typological feature vectors, expected to be positive, ranges from 0.44 for English to -0.14 for Farsi. The total correlation is found to be 0.18. Again, we find that the most typologically similar s might not be the the most accurate, hence this criterion is not suitable either.

Hence, it becomes obvious that choosing a single PL is an inefficient strategy that does not leverage the contribution that other PLs could bring. In this situation, the combination of cipher grounders for several PLs represents a sound strategy when no prior linguistic information of a certain CL is available. As shown in Table 2, this model, CIPHER-AVG, obtains accuracy scores of 56.4, 58.6, 37.4, and 37.8 % for en, fr, fa, and sw, respectively. When compared to the best performing PL for each CL (see bold cells in Table 2), it can be noticed that the performance gap ranges from just 1.2 percentage points for Swahili to 13.3 points for French, with an average of 6.1 points among all target languages.

Let us now compare the performance of CIPHER-AVG with that of a vanilla supervised neural model.[11] Table 3 shows precision, recall, and F1 scores for the NOUN tag. Even though CIPHER-AVG achieved mixed results (mid to low accuracy), the model robustly achieves mid-range performance according to F1-score for all CLs. The results are even more optimistic in terms of recall for English and French, and in terms of precision for Farsi and Swahili. This gives us hope that CIPHER-AVG can provide a useful, if noisy, signal to downstream tasks that depend on non-trivial performance over specific POS tags, such as name tagging, as exposed in the next section.

[11] We use UDPipe v1.2.0 (Straka and Straková, 2017) to train the models.

CL	Parent Language (PL)										CIPHER-AVG	PL=CL
	en	de	fr	it	es	ja	cs	ru	ar	sw		
en	-	57.1	**60.4**	59.9	59.4	25.1	52.8	49.0	30.7	28.4	56.4	73.9
fr	58.1	56.0	-	68.6	**71.9**	23.6	48.3	47.8	35.0	26.7	58.6	76.7
fa	13.8	32.3	29.7	22.7	33.3	19.7	33.3	**43.5**	37.0	38.2	37.4	73.3
sw	24.9	14.3	37.3	21.2	35.9	21.3	25.8	27.9	**38.96**	-	37.8	69.4

Table 2: Performance of cipher grounder using BROWN ($|C| = 500$) as labeler. The best PL for each CL besides itself, is shown in bold. The artificial case where we have CL POS data (PL=CL) is shown for comparison, as is the ultimately used CIPHER-AVG method.

CL	CIPHER-AVG			Supervised		
	P	R	F1	P	R	F1
en	47.70	64.4	54.81	94.04	90.44	92.20
fr	56.26	78.82	65.65	96.15	93.72	94.92
fa	64.94	51.23	57.27	96.48	97.77	97.12
sw	53.46	51.82	52.63	98.88	97.50	98.18

Table 3: Comparison of performance over the NOUN tag, as measured by precision (P), recall (R), and F1 scores, between our combined cipher grounder (CIPHER-AVG) and a supervised tagger.

5.2 Extrinsic evaluation

In the name tagging task, our LSTM-CNN baseline obtains 78.76% and 70.76% F1 score for Kinyarwanda and Sinhalese, respectively. When enriching the input representation with CIPHER-AVG tags, the performance goes up to 80.16% and 71.71% respectively. These results suggest that the signal provided by the combined cipher grounder is significant enough for relevant tags such as common, proper nouns and noun modifiers. As an example, consider the sentence `Kwizera Peace Ndaruhutse , wari wambaye nomero 11.` The baseline model fails to recognize `Kwizera Peace Ndaruhutse` as a person name. In contrast, with the PROPN tag assigned by CIPHER-AVG to `Kwizera`, `Peace`, and `Ndaruhutse`, our model is able to identify this name.

Likewise, the utility of CIPHER-AVG tags for dependency parsing under zero-resource scenarios is summarized in Table 4 and Table 5. It is important to point out that, even though the MALOPA setup follows the no-treebank setup of Ammar et al. (2016), parsing scores in the first row of Table 4 differ from those reported by them (Table 8 in Ammar et al. (2016)). Such difference is to be expected since the underlying parser used in our experiments is a graph-based neural parser (Dozat and Manning, 2017) instead of a transition-based one (Dyer et al., 2015).[12] As mentioned earlier, our objective is to analyze the effect of our tagger's signal on parsing performance under no-resource scenarios, instead of pushing the state-of-the-art for the task.

We first analyze the effect of POS tag information at test time for the MALOPA setup in Table 4. First we remove all POS signal except trivial punctuation information (NONE row), and, predictably, the scores drop significantly across all target languages. Then, we use our cipher tags (CIPHER row) and see improvements for all languages in LAS and for all but one language in UAS (de). This demonstrates the value of our cipher approach.

We then take the next logical step and remove the parallel data-grounded embeddings, replacing them with fully unsupervised MUSE embeddings. Table 5 summarizes these results. Let us compare MUSE-NONE setup (no POS signal at train or test time) with MUSE-GOLD (gold POS signal at train and test time). It can be observed that POS signal improves performance greatly for all languages when using MUSE embeddings. However, consider GUO-GOLD and MUSE-NONE. Here we note a mixed result: whilst de, sv, and it do benefit from POS information, the other languages do not, obtaining great improvements from MUSE embed-

[12] Due to time constraints, we could not experiment with longer training regimes possibly needed given the high block dropout rates in Dozat and Manning (2017).

dings instead. Finally, consider MUSE-CIPHER (gold POS tags during training, cipher tags during testing). When compared to MUSE-NONE setup, it can be observed that, unfortunately, the heuristic POS tagger is too noisy and gets in MUSE's way.

6 Related Work

Our proposed tagging pipeline can be interpreted as first reducing the vocabulary size to a fixed number of clusters, and then finding a cluster–POS tag mapping table that best explains the data without any path constraint (a cluster ID could be mapped to any POS tag). In this sense, our approach applies EM to simplify the task (e.g. when using Brown clustering (Brown et al., 1992)), followed by another EM run to optimize cipher table parameters.

Under this lens, the methods closest to our approach are those which attempt to reduce or constrain the parameter search space prior to running EM. For instance, Ravi and Knight (2009) explicitly search for the smallest model that explains the data using Integer Programming, and then use EM to set parameter values. In a different approach, Goldberg et al. (2008) obtain competitive performance with a classic HMM model by initializing the emission probability distribution with a mixture of language-specific, linguistically constrained distributions. However, both of these approaches are framed around the task of unsupervised POS *disambiguation* with a full dictionary (Merialdo, 1994). Previous work relaxes the full dictionary constraint by leveraging monolingual lexicons (Haghighi and Klein, 2006; Smith and Eisner, 2005; Merialdo, 1994; Ravi and Knight, 2009), multilingual tagged dictionaries (Li et al., 2012; Fang and Cohn, 2017), and parallel corpora (Duong et al., 2014; Täckström et al., 2013; Das and Petrov, 2011).

In addition, previous work includes sequence models that do not rely on any resource besides raw text during training, namely unsupervised POS *induction* models. These models are based, with few exceptions, on extensions to the standard HMM; most, in the form of appropriate priors over the HMM multinomial parameters (Goldwater and Griffiths, 2007; Johnson, 2007; Ganchev et al., 2009); others, by using logistic distributions instead of multinomial ones (Berg-Kirkpatrick et al., 2010; Stratos et al., 2016). However, these models still need to ground or map hidden states to actual POS tags to evaluate, and they inevitably resort to many-to-one or one-to-one accuracy scoring. Some previous work has been cautious in pointing out this ill-defined setting (Ravi and Knight, 2009; Christodoulopoulos et al., 2010), and we argue its inappropriateness for scenarios in which the test set is extremely small or even when no annotated reference corpus exists.

Therefore, the problem of grounding the sequence of states or cluster IDs to POS tags without using any linguistic resource remains unsolved. We formulate this task as a decipherment problem. Decipherment aims to find a substitution table between alphabets or tokens of an encrypted code and a known language without the need of parallel corpora. The task has been successfully applied in alphabet mapping for lost languages (Snyder et al., 2010), and machine translation at the character (Pourdamghani and Knight, 2017) and token level (Ravi and Knight, 2011; Dou et al., 2015). For the task of POS tag grounding, the sequence of states or cluster IDs is modeled as an encrypted code to be deciphered back to a POS sequence. Furthermore, we tackle the problem from a 'universal' perspective by allowing the cipher learn from POS sequences from a varied pool of languages.

Other recent work has declared a 'radically universal' mantra to language inclusivity. Hermjakob et al. (2018) presents a Romanizer that covers all writing systems known to Unicode. Pan et al. (2017) extends name tagging and linking capability to hundreds of languages by leveraging Wikipedia. Kirov et al. (2016) has semi-automatically built inflectional paradigms for hundreds of languages.

7 Conclusion

We present a POS tag grounding strategy based on decipherment that does not require human-labeled data to map states or clusters to actual POS tags and thus can be used in real-world situations requiring grounded POS tags. The decipherment model considers state or word cluster IDs of a CL as a cipher text to be deciphered back to a POS sequence.

The model operates on top of Brown cluster IDs and requires a POS language model trained on annotated corpora of one or more PLs. Experimental results over a large and linguistically varied set of PLs show that the choice of which PL to decipher

Test Tags	de		fr		es		it		pt		sv	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
GOLD	65.57	52.37	71.27	59.80	73.26	63.13	71.46	59.66	63.28	54.93	77.50	64.90
NONE	40.90	18.61	51.14	30.91	43.82	17.67	48.22	33.29	37.89	16.72	38.15	17.96
CIPHER (this work)	38.31	**24.72**	**54.46**	**41.04**	**55.56**	**41.16**	**54.05**	**39.78**	**46.97**	**36.07**	**55.06**	**36.51**

Table 4: Impact of grounded unsupervised POS tagging on MALOPA's 'zero-resource' condition. Bold entries indicate an improvement over the baseline condition of having no POS tag information (beyond punctuation)

Embeddings	Test Tags	de		fr		es		it		pt		sv	
		UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
GUO	GOLD	65.57	52.37	71.27	59.80	73.26	63.13	71.46	59.66	63.28	54.93	**77.50**	**64.90**
MUSE	GOLD	**66.19**	**56.28**	**80.86**	**72.65**	**81.06**	**73.62**	**82.08**	**72.40**	**81.17**	**76.17**	72.46	61.71
MUSE	NONE	57.26	45.10	*73.84*	*63.09*	*77.01*	*67.06*	71.36	*60.48*	*75.31*	*68.36*	60.82	45.25
MUSE	CIPHER	48.56	37.13	69.94	59.22	73.86	61.68	69.30	56.85	73.41	65.23	57.39	41.49

Table 5: Changing to unsupervised MUSE embeddings boosts MALOPA's zero-resource performance significantly (**bold** entries), in many cases doing so even without any POS tag information (*italic* entries), however noisy decipherment-based POS tags are no longer helpful.

POS tags from is crucial for performance. We explore model confidence, as measured by perplexity and typological similarities, as intuitive criteria for PL choice. However, both criteria prove to be not correlated with tagging accuracy scores. Thus, we propose a cipher model combination strategy in order to leverage the word-order patterns in several PLs, at the cost of an accuracy drop ranging from just 1.15 percentage points to 13.33 points.

The resulting combined grounder is completely language agnostic, making it attractive for the analysis of languages new to the academic community. Furthermore, analysis over the tasks of name tagging and dependency parsing demonstrate that the tags induced by the combined grounder provide a non-trivial signal for improvement of the downstream task. We obtain state-of-the-art results for name tagging in Kinyarwanda and Sinhalese, languages for which POS annotated corpora is nearly non-existent.

Acknowledgments

Thanks to Xusen Yin, Nima Pourdamghani, Thamme Gowda, and Nanyun Peng for fruitful discussions. This work was sponsored by DARPA LORELEI (HR0011-15-C-0115).

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4):467–480.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4(1):357–370.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 575–584, Cambridge, MA. Association for Computational Linguistics.

Alexander Clark. 2003. Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 59–66. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 600–609, Portland, Oregon, USA. Association for Computational Linguistics.

Qing Dou, Ashish Vaswani, Kevin Knight, and Chris Dyer. 2015. Unifying bayesian inference and vector space models for improved decipherment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 836–845, Beijing, China. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Long Duong, Trevor Cohn, Karin Verspoor, Steven Bird, and Paul Cook. 2014. What can we get from 1000 tokens? a case study of multilingual POS tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897, Doha, Qatar. Association for Computational Linguistics.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China. Association for Computational Linguistics.

Meng Fang and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada. Association for Computational Linguistics.

Kuzman Ganchev, Ben Taskar, Fernando Pereira, and Joao Graca. 2009. Posterior vs parameter sparsity in latent variable models. In *Advances in Neural Information Processing Systems*, pages 664–672.

Yoav Goldberg, Meni Adler, and Michael Elhadad. 2008. EM can find pretty good HMM POS-taggers (when given a good start). In *Proceedings of ACL-08: HLT*, pages 746–754, Columbus, Ohio. Association for Computational Linguistics.

Sharon Goldwater and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, Prague, Czech Republic. Association for Computational Linguistics.

Jonathan Graehl. 1997. Carmel finite-state toolkit. *ISI/USC*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China. Association for Computational Linguistics.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *AAAI Conference on Artificial Intelligence*.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 320–327, New York City, USA. Association for Computational Linguistics.

William P. Headden III, David McClosky, and Eugene Charniak. 2008. Evaluating unsupervised part-of-speech tagging for grammar induction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 329–336, Manchester, UK. Coling 2008 Organizing Committee.

Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018. Out-of-the-box universal romanization tool uroman. In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.

Mark Johnson. 2007. Why doesn't EM find good HMM POS-taggers? In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 296–305, Prague, Czech Republic. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Will Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398, Jeju Island, Korea. Association for Computational Linguistics.

Sven Martin, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech Communication*, 24(1):19–37.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–171.

Thomas Müller and Hinrich Schuetze. 2015. Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536, Denver, Colorado. Association for Computational Linguistics.

Robert Munro and Christopher Manning. 2012. Short message communications: users, topics, and in-language processing. In *Proceedings of the Second Annual Symposium on Computing for Development*, Atlanta.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Nima Pourdamghani and Kevin Knight. 2017. Deciphering related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2513–2518, Copenhagen, Denmark. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2015. Density-driven cross-lingual transfer of dependency parsers. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 328–338, Lisbon, Portugal. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2009. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 504–512, Suntec, Singapore. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Portland, Oregon, USA. Association for Computational Linguistics.

Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, Michigan. Association for Computational Linguistics.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057, Uppsala, Sweden. Association for Computational Linguistics.

Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden Markov models. *Transactions of the Association for Computational Linguistics*, 4(1):245–257.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti,

Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Yuan Zhang and Regina Barzilay. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1857–1867, Lisbon, Portugal. Association for Computational Linguistics.