Zero Resource Cross-Lingual Part Of Speech Tagging

Sahil Chopra
Saarland University
sach00002@uni-saarland.de

Abstract

Part of speech tagging in zero-resource settings can be an effective approach for low-resource languages when no labeled training data is available. Existing systems use two main techniques for POS tagging i.e. pretrained multilingual large language models(LLM) or project the source language labels into the zero resource target language and train a sequence labeling model on it. We explore the latter approach using the off-the-shelf alignment module and train a hidden Markov model(HMM) to predict the POS tags. We evaluate transfer learning setup with English as a source language and French, German, and Spanish as target languages for part-of-speech tagging. Our conclusion is that projected alignment data in zero-resource language can be beneficial to predict POS tags.

1 Introduction

Over the last few years, supervised machine learning methods have set higher benchmarks for numerous NLP tasks (Wang et al., 2018). However, their success relies heavily on annotated data specific to the field, which is not always readily available. As a result, for various application domains and less-resourced languages, other Machine Learning techniques must be created to handle un-annotated or partially annotated data. Currently, the most effective method for a significant portion of natural language processing tasks is to fine-tune pre-existing models using labeled data that is tailored to the specific task. Regrettably, this task-specific labeled data is often unavailable, particularly for low-resource language. One of the possible solution is fine-tuning a cross-lingual multilingual pre-trained language models (Conneau et al., 2020; Devlin et al., 2019), using available data from some source language to model the phenomenon in a different target language for which labeled data does not exist. However, the similarity between source and target language impacts performance, therefore cross-lingual transfer should not be evaluated using only a single presupposed source language, especially if training sets in multiple languages are available. In order, to limit the scope of this project, we restrict are experiments to a single source language to establish the validity of our hypothesis.

To address the issue of insufficient annotated data, another commonly used approach is to utilize parallel data. This involves pairing a text in a language with abundant resources with its equivalent text in a less-resourced language. By transferring labels from the resource-rich language to the less-resourced language, it is possible to acquire imperfect, but still valuable, annotations that can be used to train a model for the less-resourced language in a distant supervised manner (Ganchev et al., 2009; Yarowsky et al., 2001).

The aim of this project is to investigate HMM performance on a part of speech tagging trained using an artificially generated corpus of language B using language A where A is a labeled resource-rich language and B is a low-resource or unannotated language.

The rest of this paper is organised as follows: Sections 2 explains the three main components of this project, which are constructing translation, the projection labels by implementing word alignments and training HMM. Section 3 describes in detail the data used. Section 4 reports the HMM performance trained on the generated data(GD) and annotated data(AD). Section 5 presents a discussion of this project.

2 Methodology

This section presents the experimental setup and functions in the script utilized to obtain the various models in our experiment. To create the corpus in the target language, and perform POS Tagging, we first have to translate and transfer the labels of the source corpus to the target as shown in the Figure 1.

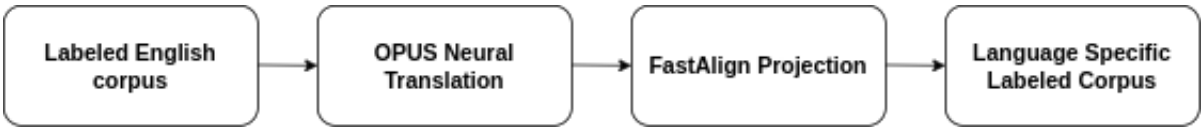

Figure 1: The pipeline for creating a data from English

We follow the same procedure as in (García-Ferrero et al., 2022). The implementation of García-Ferrero et al. (2022) only works for Name Entity Recognition. We had to modify the projection and alignment model to fit our use case.

First, the corpus is translated with the selected machine translation systems and then annotated tags are projected from the original to translated data.

OPUS Translation

The first step in the generation of the corpus in French, German and Spanish is to translate the data in hand. In order to do so, the main objective was to find a good-quality machine translation (MT) system. The corpus was translated by freely available OPUS-MT (Tiedemann and Thottingal, 2020) MT systems. The main problems with other MT systems were that some expressions were not translated at all and it was comparably slow or came with resource quota. In total, 12543 sentence pairs from the English train set were translated.

Corpus Alignment

Word alignment is a method in machine translation widely utilized for annotation projection. It is the natural language processing task of localising translation association among the words in a bitext, resulting in a many to many graph structure between the two sides of the bitext, with an link between two words if and only if they are translations of one another. It is used as a step to transfer labels of gold-annotated data to its translation. There exist numerous methods for word alignments. With the advent of deep learning more precise word alignment techniques have been developed, such as aligners for cross-lingual sequence tagging fastAlign (Dyer et al., 2013) and SimAlign (Jalili Sabet et al., 2020). For SimAlign we ran the model with 1.0 null align rate , no distortion rate and the itermax as a matching method. In our analysis , fastAlign creates better alignment with least number of NULL assignment. We drop the words that are not aligned to any POS tags to reduce the noisy data while learning the hidden representation.

Easy Projection

Taking inspiration from (García-Ferrero et al., 2022), whenever a word from the source sentence is aligned with a word from the target sentence, the target word is assigned the same category as the corresponding word in the source sentence. There are two special cases handled in projection, split annotations and annotation collision. In the first scenario, when a word in the source sentence lacks alignment, it may cause the labeled sequence to be split into multiple sequences in the target sentence. If the gap between these sequences is only one word, they are merged together. In the annotation collision case, a word in the target sentence is aligned to two different labelled sequences in the source language. We diverge our approach in second scenario where we take first the first occurrence of projected target label if they are of different category instead of just consider the one with the longest length. This helped in retaining aligned next sub sequence.

Hidden Markov Model

An HMM is a probabilistic sequence model with a given a sequence of units (words,letters, morphemes, sentences, whatever), it computes a probability distribution over possible sequences of labels and chooses the required label sequence. The use of a Markov chain involves a significant assumption, which is that only the current state is relevant for predicting future outcomes in the sequence. Previous states have no effect on the future, except through their influence on the current state.

Consider a sequence of state variables $q_1, q_2, \ldots, q_i$. A Markov model entails the assumption on Markov probabilities of this sequence that when we are predicting the next sequence , the already occurred sequence doesn't matter, only the present.

Markov Assumption: $P\left(q_i = a \mid q_1 \ldots q_{i-1}\right) = P\left(q_i = a \mid q_{i-1}\right)$

A Markov chain is specified by the following components:

$Q = q_1 q_2 \ldots q_N$

$A = a_{11} a_{12} \ldots a_{N1} \ldots a_{NN}$

$O = o_1 o_2 \ldots o_T$
$B = b_i\,(o_t)$
$\pi = \pi_i \ldots \pi_N$

where Q is a set of N states, A is a transition probability matrix , each a_{ij} representing the probability of moving from state i to state j, O a sequence of observations, B is a sequence of observation likelihoods, also called emission probabilities, each expressing the probability of an observation o_t being generated from a state q_t and π an initial probability distribution over states.

Viterbi

We use the Viterbi algorithm for decoding HMMs as given in the Figure 2. Given an observation sequence and an HMM , the algorithm returns the state path through the HMM that assigns maximum likelihood to the observation sequence. The intial parameter for viterbi algorithm is calculated as follows:

initial probabilities: the probability that a sentence starts with tag q_j

transition probabilities: probability of a tag q_j given that the previous tag in this sequence was q_i

emission probabilities: probability of a word O given the tag q_j.

Our tagger initially failed to produce output for sentences that contain words it haven't seen during training. In order to implement better unknown word handling, we use a smoothing technique. Whenever there is no emission data for the word, we replace B with the $1/f(q)$ where $f()$ returns the frequency of the state.

3 Data

We conducted experiments using Universal Dependencies 2.8 (Zeman et al., 2021) dataset. The manually annotated data contains for 114 languages; among these all have test data and 75 languages have training data. Our training datasets consist of at least 125 samples. As a result, there 105 languages that can used as a target languages, of which 65 can also serve as source languages since they have training data.

We use English UD_English-EWT dataset to generate cross lingual data in other languages and pud-ud-test dataset in French, German and Spanish languages to test. We choose these languages because they have large overlapping words with English (Dinu et al., 2015). To train HMM on gold AD, (gsd-ud) corpus in Zeman et al. (2021) is used to compare the performance with GD.

# Language	Train	Test
En(ud_english-ewt)	12543	-
Fr(Fr_gsd-ud)	16341	1000
De(De_gsd-ud)	15590	1000
Es(Es_gsd-ud)	16013	1000

Table 1: Number of samples per language

The proposed universal limited POS tagset by Petrov et al. (2012) offers a practical foundation for associating different part-of-speech categories. While it is primarily an empirical approach, this tagset consists of 12 categories: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PART (particles), and X (a catch-all for other categories). These labels have been selected for their common utility across languages and their applicability in various applications with multilingual needs.

Evaluation

We calculate Precision, F1, and Recall scores for all categories of POS tags and average over it. Note that the files for the evaluation need to be in Conll standard format. We compare the HMM model trained over the generated corpus and over the labeled corpus in the same language. To keep the comparison fair, we restrict the number of sentences in the labeled corpus equal to the number of sentences in the generated corpus.

4 Results

Table 2, 3, 4 reports the HMM performance in terms of Precision (Pr.), Recall (Re.) and F1 score. HMM trained for Spanish on generated data achieves a F1 score of 0.70 whereas it achieves a F1 score of 0.82 when a corpus is annotated. Similarly, HMM trained for French language on generated data achieves a F1 score of 0.71 whereas a F1 score of 0.74 when a corpus is annotated. Same is the case with German language. HMM trained on generated data achieves a F1 score of 0.71 whereas a F1 score of 0.79 when a corpus is annotated. The HMM did not manage to outperform the one supervised with gold labels. However, Our underline assumption that the labeled data is unavailable

POS	Generated Data			Annotated Data		
	Pr.	Re.	F1	Pr.	Re.	F1
ADJ	0.67	0.54	0.59	0.85	0.79	0.82
ADP	0.89	0.89	0.89	0.93	0.96	0.94
ADV	0.71	0.47	0.57	0.96	0.83	0.89
AUX	0.60	0.74	0.66	0.92	0.81	0.86
CCONJ	0.95	0.99	0.97	0.96	1.00	0.98
DET	0.88	0.93	0.91	0.88	0.98	0.93
NOUN	0.77	0.82	0.80	0.93	0.88	0.90
NUM	0.84	0.61	0.71	0.97	0.85	0.90
PRON	0.57	0.67	0.62	0.70	0.53	0.60
PROPN	0.50	0.39	0.44	0.63	0.64	0.63
SCONJ	0.40	0.51	0.45	0.35	0.89	0.50
VERB	0.78	0.68	0.73	0.89	0.86	0.88
Overall	0.71	0.70	0.70	0.83	0.83	0.82

Table 2: Test results on Spanish language

POS	Generated Data			Annotated Data		
	Pr.	Re.	F1	Pr.	Re.	F1
ADJ	0.73	0.55	0.63	0.92	0.80	0.86
ADP	0.85	0.83	0.84	0.93	0.95	0.94
ADV	0.83	0.49	0.61	0.94	0.92	0.93
AUX	0.73	0.92	0.81	0.86	0.99	0.92
CCONJ	0.97	0.98	0.98	0.99	1.00	1.00
DET	0.89	0.83	0.86	0.94	0.98	0.96
NOUN	0.75	0.85	0.80	0.94	0.95	0.95
NUM	0.91	0.60	0.72	0.96	0.85	0.90
PRON	0.55	0.77	0.64	0.83	0.94	0.88
PROPN	0.54	0.40	0.46	0.86	0.67	0.75
SCONJ	0.38	0.74	0.50	0.64	0.97	0.77
VERB	0.74	0.66	0.70	0.94	0.86	0.90
Overall	0.74	0.72	0.71	0.90	0.91	0.90

Table 3: Test results on French language

makes the results significant.

The gap in the performance of different training dataset widens in the case of Spanish and French language. PRON,PROPN and SCONJ have relatively low scores compared to the other tags. This observation is consistent in all the languages tested. This can be attributed to the disproportionate number of tags instances in the corpus and the corpus does not have enough instances for HMM to learn its representation.

5 Discussion

Our analysis demonstrates that the pre-training of both the source and target languages, along with the alignment of language families, writing systems, word order systems, and lexical-phonetic distance, have a huge influence on the cross-lingual

POS	Generated Data			Annotated Data		
	Pr.	Re.	F1	Pr.	Re.	F1
ADJ	0.68	0.56	0.61	0.78	0.71	0.74
ADP	0.67	0.89	0.76	0.80	0.98	0.88
ADV	0.83	0.53	0.65	0.80	0.72	0.76
AUX	0.83	0.88	0.85	0.88	0.97	0.92
CCONJ	0.99	0.78	0.87	0.99	0.74	0.85
DET	0.78	0.86	0.82	0.78	0.88	0.83
NOUN	0.78	0.77	0.77	0.91	0.81	0.86
NUM	0.80	0.57	0.66	0.96	0.83	0.89
PART	0.37	0.94	0.53	0.57	0.88	0.69
PRON	0.67	0.72	0.69	0.64	0.73	0.68
PROPN	0.49	0.39	0.43	0.52	0.57	0.54
SCONJ	0.67	0.77	0.72	0.91	0.71	0.80
VERB	0.69	0.63	0.66	0.88	0.76	0.81
Overall	0.71	0.71	0.60	0.80	0.79	0.79

Table 4: Test results on German language

performance. However, even with these considerations, our new approach's performance is not yet on the same level with that of a fully supervised POS tagger. Our findings indicate that the errors occur due to incorrect or missing alignments, particularly with articles and prepositions such as "de" and "la." Large multi-word names such as "Office national de l'immigration et de l'intégration en France" are not tagged properly. Word aligners struggle to correctly align articles in these complex expressions especially when a one-to-many or many-to-one alignment is required. For example, the alignment of the word failed to correctly align "and the" with "et de".

Other primary errors are caused by systematic differences between the tags of test and supervised text. For example in French, , the contraction "du" is formed by combining the preposition "de" (of/from) with the determiner "le" (the). In the Universal Dependency Treebank, "du" is labeled as ADP (preposition) because it functions as a prepositional phrase marker. For example, "du pain" means "of the bread" or "from the bread". However, in Wiktionary and some other linguistic resources, "du" is categorized as a contraction of the determiner "le" (the). It represents the masculine singular form of "de + le". For example, "Je mange du pain" translates to "I'm eating some bread" or "I'm eating bread". This difference in labeling can lead to challenges when aligning "du" with its counterpart in parallel corpora or when considering its part-of-speech category. While certain differences in part-of-speech labeling can be linguistically jus-

tified, stemming from inherent disparities in language structure and usage, others appear to be the result of arbitrary annotation conventions.

6 Conclusion

In conclusion, part-of-speech tagging in zero-resource settings can be achieved through the use of projected alignment data, which can be an effective approach for low-resource languages where labeled training data is not available. Existing systems rely on either pretrained multilingual large language models or projecting source language labels into the zero-resource target language and training a sequence labeling model on it. This paper explores the latter approach using an off-the-shelf alignment module and training a hidden Markov model to predict POS tags. Through evaluation of transfer learning setups with English as a source language and French, German, and Spanish as target languages for POS tagging, the authors conclude that projected alignment data in zero-resource languages can be beneficial for predicting POS tags.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Anca Dinu, Liviu P. Dinu, and Ana Sabina Uban. 2015. Cross-lingual synonymy overlap. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 147–152, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *North American Chapter of the Association for Computational Linguistics*.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. 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 369–377, Suntec, Singapore. Association for Computational Linguistics.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Daniel Zeman et al. 2021. Universal dependencies 2.8.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Appendix

function VITERBI(*observations* of len T,*state-graph* of len N) **returns** *best-path*, *path-prob*

create a path probability matrix *viterbi[N,T]*
for each state s **from** 1 **to** N **do** ; initialization step
  $viterbi[s,1] \leftarrow \pi_s * b_s(o_1)$
  $backpointer[s,1] \leftarrow 0$
for each time step t **from** 2 **to** T **do** ; recursion step
  for each state s **from** 1 **to** N **do**
    $viterbi[s,t] \leftarrow \max_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
    $backpointer[s,t] \leftarrow \operatorname{argmax}_{s'=1}^{N} viterbi[s',t-1] * a_{s',s} * b_s(o_t)$
$bestpathprob \leftarrow \max_{s=1}^{N} viterbi[s,T]$; termination step
$bestpathpointer \leftarrow \operatorname{argmax}_{s=1}^{N} viterbi[s,T]$; termination step
bestpath $\leftarrow$ the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

Figure 2: Viterbi Algorithm