Attention-based Domain Adaption Using Transfer Learning for Part-of-Speech Tagging: An Experiment on the Hindi Language

Rajesh Kumar Mundotiya, Vikrant Kumar, Arpit Mehta and **Anil Kumar Singh**
Department of Computer Science and Engineering, IIT(BHU), Varanasi, India
{rajeshkm.rs.cse16, vikrantkumar.cse18}@iitbhu.ac.in,
{arpitmehta.cse18, aksingh.cse}@iitbhu.ac.in

Abstract

Part-of-Speech (POS) tagging is considered a preliminary task for parsing any language, which in turn is required for many Natural Language Processing (NLP) applications. Existing work on the Hindi language for this task reported results on either the General or the News domain from the Hindi-Urdu Treebank that relied on a reasonably large annotated corpus. Since the Hindi datasets of the Disease and the Tourism domain have less annotated corpus, using domain adaptation seems to be a promising approach. In this paper, we describe an attention-based model with self-attention as well as monotonic chunk-wise attention, which successfully leverage syntactic relations through training on a small dataset. The accuracy of the Hindi Disease dataset performed by the attention-based model using transfer learning is 93.86%, an improvement on the baseline model (93.64%). In terms of F_1-score, however, the baseline model (93.65%) seems to do better than the monotonic-chunk-wise attention model (94.05%).

1 Introduction

Deep learning has been consistently providing promising results on a large variety of language processing problems. Textual processing includes diverse applications of NLP such as text classification, dialect identification and classification, sequence labelling problems (such as Named Entity Recognition and Extraction, Chunking and POS tagging) and machine translation.

However, for performance improvement obtained on the preliminary NLP tasks – POS tagging and Chunking – especially under a low resource scenario, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) have been used more. An efficient way of information modeling by Gated Recurrent (GRU) and Long Short Term Memory (LSTM), a variant of RNN, has also been tried.

Earlier work on POS tagger for canonical Hindi text achieved considerable results of about 97.10% on Universal Dependency dataset (Plank et al., 2016), which belongs to a single domain. The performance reduces radically after deploying this existing trained model to a different domain-specific data or out-of-domain data. Domain-specific data such as Tourism and Disease has its own distributions and having a minimal amount of annotated dataset, considered as low resources, which also causes an Out-of-vocabulary (OOV) words issue.

OOV is a major problem in low resources text processing, faced while training a model on one domain of a language and trying it to another domain of the same language. This problem is partly countered by incorporating character level information into the model.

Lately, Transfer Learning has been shown to enhance the performance of the model by transferring learned features (general features as well as domain-specific features) which were obtained during training the model. The general features are transferred to the target domain through an initializer or feature extractor. These methods are beneficial as they benefit from the pre-trained model via neurons (Zennaki et al., 2019). Yang et al. (2017), Meftah et al. (2018) have followed the Transfer Learning approach on English (following Subject-Verb-Object sentence structure), while there is not much work for Hindi (following Subject-Object-Verb sentence structure) using such models.

The proposed architecture of (Ma and Hovy, 2016) is employed as a baseline model for the purposes of our work. It encodes character level information by CNN. Authors have strengthened the baseline model through attention mechanism: self-attention and monotonic chunk-wise attention as the contribution. The motivation behind using these attention mechanisms is that it exhibits adequate improvement on neural machine translation, especially for low resource regime (Chiu and Raffel, 2017; Bahdanau et al., 2014; Goyal et al., 2020). Also, the experimental datasets required can be smaller in size. The improvement in capturing syntactic information is due to the attention mechanisms. The results obtained by the attention mechanism provide an improvement over the original baseline results.

2 Baseline Model

We use as our baseline the above mentioned model using a discriminative tagging model proposed by Ma et al. (2016), together with character-level information encoded by CNN, illustrated in Figure 1.

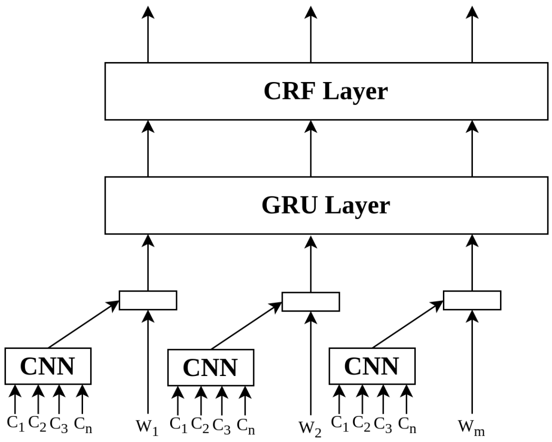

Figure 1: Baseline model for POS Tagging

In this model, the preservation of both syntactic and semantic information of words is achieved by a combination of two vectors obtained at word-level and character-level (Murthy et al., 2018).

The character-level information captures orthographic and morphological features by applying CNN (Murthy et al., 2018), where characters are initially represented by a one-hot encoder and passed to convolution layer. The convolution layer holds n-gram information followed by max-pooling layer, where n is given by filter size. Maximum relevant information over the different features perceived through this layer, which are the distinct features of the word, represented at the character level, are passed to a fully connected layer. This layer used a Rectifier Linear Unit (ReLU) as a non-linear activation function to produce character-level word vector. The word vector is assigned by random initialization which is learnt during model training. The concatenated character and word-level vector is fed to the Bidirectional GRU. The obtained output from forward and backward GRUs at each time-step are combined before being fed to a Conditional Random Fields (CRF) layer. The CRF layer generate a probability score over the labels at each time-step.

3 Attention Based Model

Since the last few years, attention mechanisms have been providing promising results in NLP applications as well, e.g. Machine Translation gets a better alignment between the source and the targets words after applying the attention mechanism (Bahdanau et al., 2014; Chiu and Raffel, 2017). Here, we use two attention mechanisms into the baseline model: self-attention (Cheng et al., 2016) and Monotonic Chunkwise Attention (MOCHA) (Chiu and Raffel, 2017) to enhance the capabilities of capturing syntactic relations from input words.

3.1 Attention Mechanism

Self-attention or intra-attention (Cheng et al., 2016) became popular after a Transformer model came into existence for Neural Machine Translation (Vaswani et al., 2017). The Transformer proposed by Vaswani et al. (2017) completely relied on self-attention, which uses different positions of the input to obtain the attention score. The primary reason for calling self-attention as intra-attention is a dependency on itself for score calculation, which is calculated by applying softmax over the additive or dot product of the current vector with previous attention score. These intra-word dependencies are helpful for capturing the syntactic relations among words during labelling.

Monotonic chunk-wise attention (Chiu and Raffel, 2017) is also an extension of Hard monotonic attention. It provides flexibility to the attention score calculation. In this method, the calculation of energy score is based on the chunk (a particular static word window size) rather than entire word input (usually following soft attention) or a particular time-step of input (generally following Hard monotonic attention). The energy score uses chunk energy (soft attention over a limited window) and monotonic energy (Bahdanau attention (Bahdanau et al., 2014) with a sigmoid function instead of softmax) to calculate the attention score. This attention score is calculated for each time-step input.

3.2 Attention-based Model

The previous extensions to the attention mechanisms are based on the encoder-decoder architecture, prevalent in end-to-end neural machine translation systems. In our work, two Bidirectional GRU layers are exploited for incorporating the attentions in the baseline model for POS tagging. The first GRU layer is treated as an encoder for attention and the remaining layer as a decoder for the attention-based extended baseline model. The dropout layer is also used between attention input and output to prevent overfitting. The rest of the model architecture from the input data by CNN and word vector to predictions by CRF are the same as the baseline model, as shown in Figure 2.

3.3 Domain Adaption model

Domain adaption has been performed with supervised, unsupervised and semi-supervised settings until now for many tasks including POS tagging. We have used relatively little annotated data to build a robust POS tagger for the target domain by using Transfer Learning. Transfer Learning procedure closely follows the Meftah et al. (2018) settings. The attention-based model has been trained on the first domain for POS tagging while performing transfer learning. The optimal learned parameters during this training are passed for the training of another domain. That is a standard procedure of transfer learning where all labels are considered as equal.

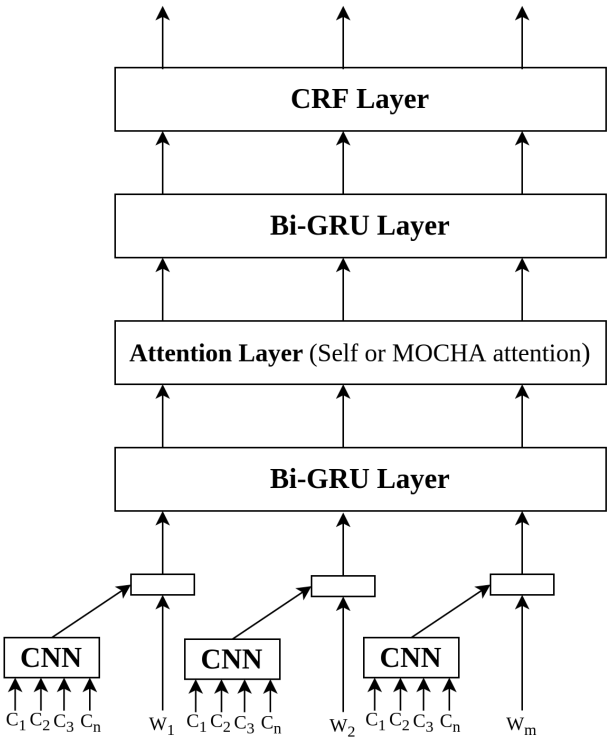

Figure 2: Attention-based extended baseline model for POS Tagging

Here, the optimal parameters θ_s from the training of source domain are used for initialization of the target domain's parameters θ_t. After this initialization ($\theta_s \rightarrow \theta_t$), the model is fine-tuned for the target domain, as shown in Figure 3.

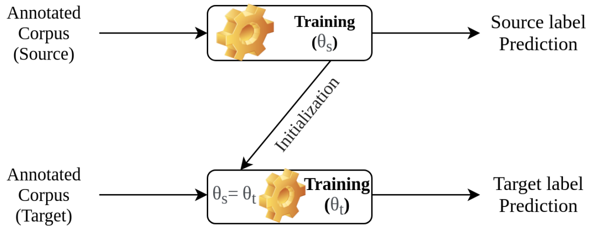

Figure 3: Domain adaption via transfer learning approach

4 Experimental Setup

4.1 Dataset

For performing the experiments of domain adaption, we have used Disease and Tourism domains of the Hindi Treebank dataset[1]. The dataset follows the

[1] http://tdil-dc.in/index.php?lang=en

Bureau of Indian Standards (BIS) tagset. The statistics of the dataset are mentioned in Table 1. As the size of the dataset is small, and out of which overlapped types are 1579, we have extracted Treebank for our experiments.

Domain	**Sentences**	**Types**
Tourism	3022	7100
Disease	1494	4987
Overlapping	-	1579

Table 1: Hindi Treebank data statistics according to domain

Since the size of the Disease domain dataset is smaller compared to Tourism, it is considered a source domain, while the other is considered the target domain for domain adaption.

4.2 Settings

The source and target domain datasets are divided in a 70%–30% ratio for performing validation of the trained model. The maximum length of sentences and words has been fixed for training the model, which is 52 and 22, respectively. However, gradient calculation avoided the padded sentences and words, which in turn prevents overfitting. The character vector size 32 are obtained after applying two filters 64 and 124, each with the size of 3, with a dropout of 30%. The model trained with the word vector and GRU unit of 100 and 128, respectively. As annotation corpus is tiny, the model tends to overfit quickly. Hence, dropout and early stoppage have applied with the value of 50% and 30 as patience, respectively. The parameters and hyper-parameters used in training are briefly mentioned in Table 2.

5 Result and Analysis

The baseline model is also robust towards the POS tagging as the obtained results on the Disease dataset for isolated training has improved by domain adaptions even tough overlapping vocabularies are relatively small (1579 types). The baseline model gets up from 93.64% to 94.29% as in isolation and domain adaption training, respectively, which is the highest accuracy among reported results in Table 3.

The self-attention-based model has degraded the performance due to their nature of attention score calculation and limitation of sentence length. On the other hand, MOCHA-based model has improved the POS tagging system's performance due to the nature of chunk consideration during attention score calculations. We have used a chunk size of 8 in model setup. The MOCHA-based model obtained an accuracy of 93.86%, which has a slight improvement over the baseline model depicted in the Table 3.

(Hyper-)parameter	**Value**
Char. vector	30
Word vector	100
Batch Size	32
Filters	[64, 124]
Filter size	3
CNN Dropout	0.3
GRU unit	128
Dropout	0.5
Early stoppage	20
Optimizer	Adam

Table 2: The value of parameters and hyper-parameters used in model training

Model (%)	**Accuracy**	**F$_1$-score**
Baseline Model	93.64	94.05
Baseline Model + DA	**94.29**	94.20
Self-Attention + DA	91.11	90.46
MOCHA + DA	**93.86**	93.65

Table 3: Obtained results from baseline model and attention based models, where DA indicates Domain adaption settings

The baseline model and monotonic chunk-wise attention model achieved 94.05% and 93.65%, respectively as best F$_1$-score for domain adaption. However, after tuning the hyper-parameter for DA, learning rate (0.01 for Baseline, 0.02 for MOCHA and 0.004 for Self-attention) of these model have improved the performance. We have used Variable length of training size (200, 400, 600 and 900) for DA training by using these models that show Self-attention model performs better (94.63% F$_1$-score on training size of 900) than other models (94.20% and 93.65% F$_1$-score on training size of 900 for Baseline and MOCHA model, respectively), as illustrated in Figure 4.

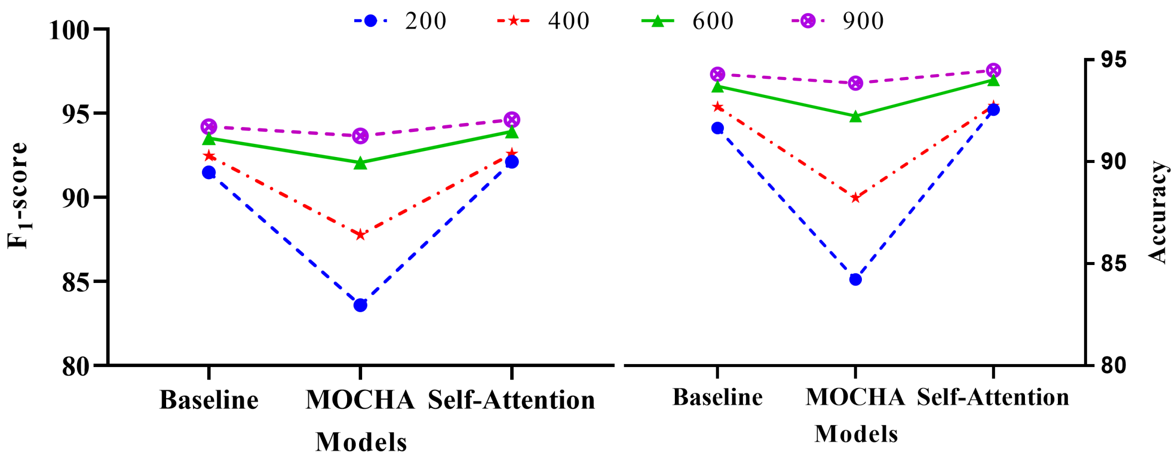

Figure 4: Accuracy and F$_1$-score comparison on Variable length of training data size for DA on the Baseline, Self-attention and MOCHA-based model

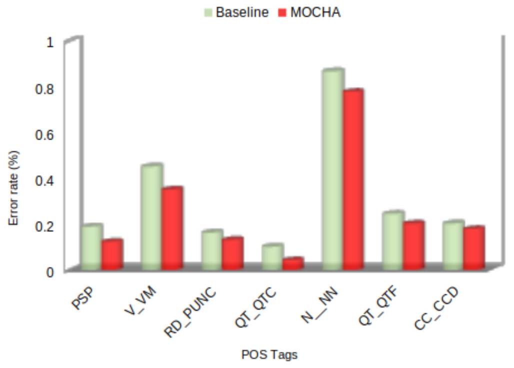

Figure 5: Error-rate comparison between selective most and less frequent POS tags obtained from predictions of baseline and MOCHA-based model

As evident from Table 3, the MOCHA-based model is precise over the baseline model. From the analysis of predictions file of the baseline model and MOCHA-based model, we found that error-rate reduced on the selective tags. Postposition (PSP), Main Verb (V_VM), Punctuation (RD_PUNC), Cardinal Quantifier (QT_QTC) and Co-ordinator Conjunction (CC_CCD), General Quantifier (QT_QTF), Common Noun (N__NN) are selective most and less frequent POS tags, respectively. These tags have reduced error rate, the difference among these are shown in Figure 5. Hence, It shows that MOCHA-based model more accurate to predictions of right POS tags on scarce words as well. On other POS tags, the error rate of the MOCHA-based model found to be comparable to the baseline model.

6 Related Work

A short chronological overview of the related work presented here to provide the context of our work. Blitzer et al. (2006) used Structural Correspondence Learning (SCL) to automatically induce correspondence to the features of a different domain in order to transfer POS tagger from Wall Street Journal (financial news) to MEDLINE (biomedical abstracts).

Collobert et al. (2011) presented a task-independent, a learning algorithm and unified convolutional neural network architecture, pertaining to various NLP tasks as POS tagging, Chunking, Named Entity Recognition and Semantic Role Labelling. They jointly trained models of POS tagging, Chunk and NER tasks with the additional linkage in trainable parameters for transferring knowledge learned in one task to another.

Zhang et al. (2014) showed type-supervised domain adaptation for the Chinese word segmentation and POS tagging, using domain-specific tag dictionaries. Unlabeled target domain dataset has improved target domain accuracy by providing annotated source domain dataset. They have obtained a 33% error reduction on target domain tagging by unlabeled sentences and a lexicon of 3000 words.

Yu et al. (2015) used an effective confidencebased self-training approach to select additional training samples for domain adaptation of a dependency parser and were able to improve parsing accuracy for out-of-domain texts by 1.6% on texts from a chemical domain.

Mishra et al (2017) used unlabeled data for POS tagging applying for feature transfer via transfer learning from resource-rich to resource-poor language across eight Indian languages, each having 25K sentences and gained an average accuracy of 81%.

Yang et al. (2017) explored transfer learning for neural sequence tagging, where source task with large annotated dataset was exploited to enhance the performance of the target task with smaller dataset. They examined the effect of Transfer Learning on recurrent neural networks across domains, applications and languages, and obtained significant improvement.

Meftah et al. (2018) used GRU, CRF and CNN for character level feature representation as model components for POS tagging as a sequence labelling problem. To address the data scarcity, they examined the effectiveness of Cross-Domain and Cross Task Transfer Learning.

Li et al. (2019) proposed a domain embedding approach to merge the source and the target domain training data. The results demonstrated that it is more effective than multi-task learning approaches and both direct corpus concatenation (as traditional approach). Contextualized word representation with fine-tuning is used to utilize unlabeled target-domain data, which further increased its cross-domain parsing accuracy.

We have used a similar CNN architecture as proposed by Meftah et al. (2018), except that we have applied different sizes of stacked convolution layers. We have also used the same transfer settings across the domain for performing domain adaption the Hindi Treebank dataset.

Distributed word representation usually learns semantic and syntactic information about the word and ignores word size and morphological features. Part-of-speech tagging requires intra-word information when dealing with morphologically rich language.

Santos et al. (2014) have demonstrated that CNN is an effective approach for extracting morphological features and encoding it into neural representations. Singh et al. (2018) used CRF and LSTM Recurrent Neural Networks to model POS Tagging on Hindi-English Code Mixed dataset from Twitter and achieved a result of overall F_1-score of 90.20%. These works are related to our use of character level information in the models that we used.

7 Conclusion

The attention-based extended baseline model is a simple model for domain adaption to perform Part-of-Speech (POS) tagging if there is scarcity of annotated corpus. It is an extension of the LSTM-CNN-CRF model by replacing LSTM by GRU and appending attention mechanisms (self-attention and monotonic chunk-wise attention). This model was used to perform domain adaption on the Hindi Treebank dataset, where the Tourism domain was considered as the source domain and Disease as the target domain for the Transfer Learning scenario. The results show the improvement over the baseline model by the monotonic chunk-wise attention mechanism. The limitation of scarcity of annotated corpus of both of the domains can be overcome to some extent by using available pre-trained word embeddings or raw corpus to get better embeddings for this model as part of future work. In addition to this, additional linguistic information can be fused into the model to leverage the advantages of additional accessible annotations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas, November. Association for Computational Linguistics.

Chung-Cheng Chiu and Colin Raffel. 2017. Monotonic chunkwise attention. *arXiv preprint arXiv:1712.05382*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.

Cicero Dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *International Conference on Machine Learning*, pages 1818–1826.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2386–2395, Florence, Italy, July. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany, August. Association for Computational Linguistics.

Sara Meftah and Nasredine Semmar. 2018. A neural network model for part-of-speech tagging of social media texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pruthwik Mishra, Vandan Mujadia, and Dipti Misra Sharma. 2017. POS tagging for resource poor languages through feature projection. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 50–55, Kolkata, India, December. NLP Association of India.

Rudra Murthy, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2), December.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany, August. Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A twitter corpus for hindi-english code mixed pos tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 12–17.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks.

Juntao Yu, Mohab Elkaref, and Bernd Bohnet. 2015. Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies*, pages 1–10, Bilbao, Spain, July. Association for Computational Linguistics.

Othman Zennaki, Nasredine Semmar, and Laurent Besacier. 2019. A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages. *Natural Language Engineering*, 25(1):43–67.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-supervised domain adaptation for joint segmentation and POS-tagging. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 588–597, Gothenburg, Sweden, April. Association for Computational Linguistics.