Multilingual Sequence Labeling With One Model

Alan Akbik
Zalando Research
Mühlenstraße 25
10243 Berlin

Tanja Bergmann
Zalando Research
Mühlenstraße 25
10243 Berlin

Roland Vollgraf
Zalando Research
Mühlenstraße 25
10243 Berlin

{firstname.lastname}@zalando.de

1 Introduction

A large family of NLP tasks such as named entity recognition (NER) and part-of-speech (PoS) tagging may be formulated as sequence labeling problems; text is treated as a sequence of words to be labeled with linguistic tags. However, industrial applications of sequence labeling are often faced with *multilingual text*, i.e. data in more than one natural language.

The standard approach to address multilingual data is to train separate sequence labeling models for each language, and to employ automatic language identification to select the appropriate model for a given text document. However, this approach effectively multiplies the effort of training a model, pre-training and selecting appropriate word embeddings, and requires an array of models to be managed in production use.

One model for many languages. In this extended abstract, we present a novel approach in which we train *a single model* to perform part-of-speech (PoS) tagging for 12 distinct languages from 4 different language groups. Our approach leverages neural language modeling with recurrent neural networks (RNNs) to model text at the character level. We show that it is possible to train a single language model (LM) over multilingual text data, causing it to implicitly capture language as a feature in its internal representation. This enables us to extract meaningful *word embeddings* from pre-trained, multilingual LMs against which we train a single downstream sequence labeling model over aggregate multilingual training data. Our experiments indicate that this model internally performs language identification and is thus immediately applicable to multilingual data.

2 Method

Our proposed approach is based on recent advances in neural language modeling (LM) that have allowed language to be modeled as distributions over sequences of characters instead of words. Recent work has shown that by learning to predict the next character on the basis of previous characters, such models learn internal representations that capture syntactic and semantic properties: even though trained without an explicit notion of word and sentence boundaries, they have been shown to generate grammatically correct text, including words, subclauses, quotes and sentences. Importantly, we presented in recent work a method to extract word-level representations from pre-trained character LMs that we showed to be highly meaningful when used as word embeddings in downstream tasks (Akbik et al., 2018).

Multilingual character-level language modeling. In this work, we apply the same approach, but train a language model over text in multiple languages. Since we train a purely character-level model, there is no computational impact of broadening to other languages. The vocabulary of characters (letters of the alphabet, numbers and special characters) remains nearly identical to training such a task over English-only text[1]. This produces a pre-trained multilingual language model which - next to syntax and semantics - also captures language as a latent

[1] This stands in stark contrast to training word-level models in which the vocabulary (the space of all unique words) would explode which each language added.

	HectaLM		Yasunaga et al. (2018)	Plank et al. (2016)	Berend (2017)	Nguyen et al.(2017)	Yu et al. (2017)
	fast	*default*					
English	94.37	**96.02**	95.82	95.16	93.47	94.7	94.76
German	93.48	**94.58**	94.35	93.38	90.73	92.7	92.77
French	96.83	**97.59**	96.63	96.11	94.96	96.0	96.27
Italian	96.95	**98.29**	98.08	97.95	96.28	97.5	97.74
Dutch	93.56	**96.47**	93.09	93.30	85.10	91.4	93.11
Polish	95.45	**98.09**	97.57	97.62	93.95	96.3	96.92
Avg	95.10	**96.84**	95.92	95.55	92.42	94.76	95.26

Table 1: Universal part-of-speech tagging on the 6 languages in HectaLM.

feature. This allows us to employ the approach from (Akbik et al., 2018) to extract word embeddings for text in any language the model was trained on, effectively producing a single model we use as *embedding layer* in a multilingual downstream task.

3 Experiments

3.1 Setup

For our experiments, we use the open-source Flair framework, which implements the standard BiLSTM architecture for sequence labeling and provides classes to experiment with language model embeddings.

Tasks. We utilize the data from the Universal Dependencies project in our experiments to train two tasks: (1) A *6-language universal PoS tagging task* in which we train over aggregate UD data for English, German, French, Italian, Dutch and Polish. (2) A *12-language universal PoS tagging task* in which we train over aggregate UD for the previous 6 languages plus Spanish, Swedish, Danish, Norwegian, Finnish and Czech.

Language models. We train character-level language models (which we refer to as HectaLM) over a corpus that spans six human languages, namely English, German, French, Italian, Dutch and Polish. The corpus contains approximately 10 billion tokens from various sources aggregated by the Opus project (Wikipedia, parliament speeches, movie subtitles, medical texts, news commentary, books). Our vocabulary contains 275 distinct characters which covers 99.99% of characters in the corpus, the rest are replaced by a special UNK token. We train two models: A "default" model with 2048 hidden states and one layer and a smaller model (dubbed "fast") with 1024 hidden states and one layer, projected to 512 dimensions before prediction. We train both models for one week on a v100 GPU.

Baselines. We compare against previous best-reported numbers on multilingual PoS tagging (Yasunaga et al., 2018; Yu et al., 2017; Nguyen et al., 2017; Berend, 2017; Plank et al., 2016).

3.2 Results

State-of-the-art on in-LM languages (Table 1). We first evaluate downstream task performance on the six languages that were used to train the HectaLM. As Table 1 shows, our single-model approach outperforms all previously published numbers across all languages. In addition to outperforming previous results, our approach has the practical advantages of requiring only a singular model, no language-specific word embeddings and no language identification.

Good results even on languages not in the LM (Table 2). Remarkably, as Table 2 indicates, our approach can even be applied to languages the language model has not been trained on and still produce usable word embeddings for many related European languages. Only on Finnish, which linguistically is further apart, the model struggles to produce meaningful embeddings.

	HectaLM		Yasunaga	Plank et	Berend	Nguyen et	Yu et al.
	fast	*default*	et al. (2018)	al. (2016)	(2017)	al.(2017)	(2017)
English	93.47	95.61	**95.82**	95.16	93.47	94.7	94.76
German	93.20	94.24	**94.35**	93.38	90.73	92.7	92.77
French	96.44	**97.61**	96.63	96.11	94.96	96.0	96.27
Italian	96.37	**98.18**	98.08	97.95	96.28	97.5	97.74
Dutch	92.82	**96.37**	93.09	93.30	85.10	91.4	93.11
Polish	94.57	**98.01**	97.57	97.62	93.95	96.3	96.92
Spanish	88.10	95.53	**96.44**	95.74	94.69	95.9	95.65
Swedish	93.07	94.71	**96.70**	96.69	94.62	–	96.28
Danish	88.27	94.08	**96.74**	96.35	93.32	95.8	96.06
Norwegian	82.95	96.39	**98.08**	98.03	95.67	97.4	97.65
Finnish	94.23	90.89	95.40	**95.85**	89.19	94.6	95.79
Czech	92.73	97.51	**98.81**	98.24	95.83	–	98.79
Avg	92.19	95.76	**96.48**	96.20	93.15	95.23	95.98

Table 2: Ablation experiment: Universal part-of-speech tagging on 12 languages. Even though only the upper 6 languages are in HectaLM, the model performs remarkably well even on the remaining 6 languages.

4 Conclusion and Outlook

We find these initial experiments encouraging for single-model approaches to multilingual text data. Our current work focuses on training larger multilingual language models and expanding experimentation to more downstream NLP tasks. By increasing the hidden states and layers of the respective models, we hope to close the quality gap between our proposed approach and current state-of-the-art single-language models for part-of-speech tagging. We integrate our approach into the Flair framework for use and reproduction by the research community.

References

[Akbik et al.2018] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

[Berend2017] Gáabor Berend. 2017. Sparse coding of neural word embeddings for multilingual sequence labeling. *Transactions of the Association for Computational Linguistics*, 5:247–261.

[Nguyen et al.2017] Dat Quoc Nguyen, Mark Dras, and Mark Johnson. 2017. A novel neural network model for joint pos tagging and graph-based dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies: August 3-4, 2017 Vancouver, Canada*, pages 134–142. The Association for Computational Linguistics.

[Plank et al.2016] Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 412.

[Yasunaga et al.2018] Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *NAACL 2018, North American Chapter of the Association for Computational Linguistics*, pages 976–986.

[Yu et al.2017] Xiang Yu, Agnieszka Falenska, and Ngoc Thang Vu. 2017. A general-purpose tagger with convolutional neural networks. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 124–129.