Finnish, German, Swedish Part-of-Speech Tagging with Bidirectional Long-Short Term Memory Models and Auxiliary Loss

Yu An
yu.an@helsinki.fi

1 Introduction

1.1 Overview of part of speech

Part of speech, also known as POS, word classes or syntactic categories, describes the attribute of a word and its relationship with the neighboring words. For example, knowing a word is a verb tells us about its meaning, which describes some kind of action or state being; it suggests the probable attributes of its immediate neighbors, which is being preceded by a noun and followed by a noun. In English, only verbs and verb phrases are allowed to work as the predicate of a sentence (Ping, 2005), which shows its informativeness in syntactic parsing. Part of speech tags are useful features in NLP tasks such as named entities recognition, information extraction, coreference resolution and even speech recognition (Jurafsky and Martin, 2019), for these tasks involve the relationship between words, which is the very knowledge that POS is supposed to provide.

1.2 Available methods

The POS of a word can change over contexts. As a result, the POS tagging problem is in essence to seek the tag that best fits the particular instance of the word. Solutions to it belongs to the classification realm from the perspective of machine learning methodology. More specifically, the very definition of part of speech that it is concerned with knowledge about upper and lower stream rather than only about the instance word itself puts the pointer toward sequence labeling algorithms. Scherrer (2019)'s summary compared the common approaches to sequence labeling by describing them from three aspects: whether the model is generative or discriminative, what kind of input features to use and the scope of prediction. The summary is illustrated in Table 1.

The Algorithms column in Table 1 exhibits the approaches from the most straight-forward ones to the more refined ones. Naive Bayes with unigram input feature cannot be simpler: it makes decision by examining the posterior probability of words given a possible POS and chooses the one that gives the highest probability, which puts it into the type of generative models. Based on the same theorem, HMM models takes context information into account by computing the probability distribution over sequences of labels rather than a single label. Maximum entropy Markov model further overcomes the disadvantage of generative models, opening the seats to a larger set of possible input features and making it possible to look at previous, next words or even word suffixes. The more advanced model, CRF, manages to look at the whole sequence at each time step and hence implements bidirectionalty on top of the MEMM. It directly leads into the subject of Plank et al. (2016)'s work, the bi-LSTM models, which also equips with bidirectionality. Additionally, according to Jurafsky and Martin (2019), Taggers using the above approaches has achieved satisfactory performances in pre-deep learning era, at least on WSJ training corpus, a dataset in English. Plank et al. (2016) also showed in the work the performance of the baseline bigram HMM tagger with suffix trie (Brants, 2000), which reached around 95% accuracy across the 22 languages. The remarkable performance of baselines shows POS tagging is rather easy as a task, which can be imagined since it is known that every language has a few unambiguous words and some ubiquitous tokens like punctuation are generally predictable.

Deep learning sequence models tackled the limitation of Markov assumption, which is the constraint for the context view a tagger can take at each time step. Recurrent neural networks allows to handle variables of veritable lengths. This improvement also applies to Bi-LSTM, as a variant

Algorithms	Model type		Input features			Scope of prediction			
	Generative model	Discriminative model	Current word	Context feature (previous / next word)	Previous label	Unigram model with greedy inference	Sequence model with greedy inference	Sequence model with exact unidirectional inference	Sequence model with exact bidirectional inference
Naive Bayes	1		1		1	1			
Greedy HMM	1		1		1		1		
Viterbi HMM	1		1		1			1	
Maximum entropy Markov model (MEMM)		1	1	1	1			1	
Conditional random field (CRF)		1	1	1	1				1
Feed-forward neural networks			1	1		1			
Recurrent neural networks (RNN)			1	1			1		
Recurrent neural networks with CRF layer			1	1	1				1

Table 1: The overview of sequence labeling algorithms (Scherrer, 2019)

of RNN.

2 Models

The innovation (Plank et al., 2016) introduced is to use a multi-task bi-LSTM for POS tagging. If we follow the stream of the input, the bottom layer the data are passed to should be the embedding. To test the effect of the granularity of the representations, two parallel embedding layers are set to encode information from different levels of the input token. Then, depending on the the granularity of the embedded sequence, the vectors proceed to different branches to further encode the input. Plank et al. (2016) used the term "sequence bi-LSTM" to refer to the "encoder" layer. The input to the layer includes only the embedding vectors. Sequences are read from both directions, and the output is would be a concatenation of the last hidden state of the forward and backward passes. Such architecture only applies to sub-token level embeddings, that is, only byte embeddings and character embeddings will be processed by the sequence bi-LSTM, if they are chosen. Word embeddings are directly concatenated to the vectors representing the sub-token information yielded from the sequence bi-LSTM, and do not need to be passed to another layer. Then the composite representation along with an additional input that indicates the word's position in the sentence are put to another bi-LSTM layer, which connects to a linear layer to map the hidden state space to the tag space. In the multi-task version of the model, the same output from the higher level bi-LSTM goes through another linear layer that is parallel to the tag space mapper, which predicts the log frequency of the word token in the training set. Hence the total loss that is back-propagated at each time step would be the sum of the losses from predicting the POS tag and the log frequency.

3 Datasets

Following the advice in project instruction, I tested the model performance for Finnish, German and Swedish languages on data from the latest release of Universal Dependences project (Nivre et al., 2019) by March 2020, UD v2.4. The Finnish dataset is *UD_Finnish-TDT*, containing 12217 sentences in training set, 1364 sentences in dev set, 1555 sentences in test set. Swedish *UD_Swedish-Talbanken* data set has 4303 sentences for training, 504 sentences for validation and 1219 sentences for testing. For German, UD German-HDT data set is quite large compared the the two languages before, with 68801 sentences in the training set file *de_hdt-ud-train-a.conllu*, 17028 sentences in the test set and 17293 in the development set. For the sake of trianing time, I used the first 12000 sentences in *de_hdt-ud-train-a.conllu* for training the German tagger and kept all the sentences in test and development set. German and Finnish are morphologically rich languages, while Swedish is relatively poor in morphological changes.

4 Experiments

All bi-LSTM models were implemented in PyTorch. I followed the exact setting as in the original study except the Gaussian noise for training labels, i.e. SGD optimizer with cross-entropy loss, no mini-batches, 20 epochs, 0.1 learning rate, 128 dimensions for word embeddings, 100 for character and byte embeddings, 100 for all hidden states. For I mistaken the noise setting, i.e. Gaussian noise with σ=0.2 for the initial state of hidden states, the

	Without auxiliary loss and using:			With auxiliary loss and using:		
	$\vec{w}$	$\vec{c}$+$\vec{b}$	$\vec{w}$+$\vec{c}$	$\vec{w}$	$\vec{c}$+$\vec{b}$	$\vec{w}$+$\vec{c}$
fi	81.94 (vs 94.85)	95.42 (vs 89.15)	93.19 (vs 87.95)	84.07	95.07 (vs 95.85)	93.97
de	86.62 (vs 92.94)	95.50 (vs 90.11)	94.28 (vs 90.33)	86.85	95.85 (vs 93.38)	94.72
sv	73.01 (vs 96.36)	92.18 (vs 95.50)	89.28 (vs 93.32)	75.20	92.93 (vs 96.69)	90.30

Table 2: Tagging accuracies on UD 2.4 test sets. $\vec{w}$: words, $\vec{c}$: characters, $\vec{b}$: bytes.

code for initializing hidden state was set to a random sample of Gaussian. For the performance of taggers showed were decent, I believe it has only trivial effect on the model. No pre-trained embeddings were loaded and no regularisation were used.

In the end, 3×3×2=18 models were presented in my experiment. For each of the three languages, I implemented two versions of models, that is, one with the auxiliary loss function and another without. And for both versions of the models, I tried 3 kinds of representations: word embeddings alone, a concatenation of word and character embeddings and a concatenation of byte and character embeddings.

Besides, I also tried the most-frequent-tag baseline for each language to see the difficulty of the task. It naively reports the most frequent POS tag as the answer to every instance.

5 Results

First, the performance of the most-frequent-tag baseline is rather substandard, achieving only 27.4%, 23.1% and 21.2% on Finnish, Swedish and German respectively. This fits the impression that despite POS tagging being a relatively easy task, it requires effort to reach a 50+% accuracy. My reproduced results are given in Table 2. The accuracies in parentheses are the performances of the models in the original work which used the same loss function and the same representation for training in the original work, as a comparison.

6 Conclusions and Discussion

Most of the results from single-task models are not consistent with those in Plank et al. (2016) in both absolute numbers and preference for representations. It can be my misunderstanding for the presented *bi-LSTM* model in the paper, or it can be the effect of label noise in the original paper working as regularization, since the difference between my training and development accuracies showed that 20 epochs has led to overfitting. For all the three languages, my results show that the combination of byte and character embeddings outperforms the other representations, achieving around 95% of accuracies. Character-level information definitely benefits when it is fused with words. Training with only word embeddings falls short on both versions of models, indicating that the modification of the loss function does not bring much strength to the model. Swedish suffers most if the representation only include word embeddings.

Although this experiment uses different datasets than the presented paper (UD 1.2 in Plank et al. (2016), UD 2.4 here), given the implication of Figure 3 in that on the effect of dataset sizes that training data that exceeds 1000 sentences does not attribute much to the accuracy, the difference between the original results and mine is more likely caused by the regularisation effect of label noise.

As a conclusion, this experiment showed the curicial role the choice of representation plays in bi-LSTM taggers. This fits the general impression that character-level information helps in informing POS, a characteristic between morphological and lexical level.

References

Thorsten Brants. 2000. Tnt: A statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, page 224–231, USA. Association for Computational Linguistics.

D Jurafsky and JH Martin. 2019. Speech and language processing (3rd (draft) ed.).

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou,

Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Kyung-Tae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2019. Universal dependencies 2.4. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

LI Ping. 2005. Contrasting the predicates between english and chinese in terms of word classes. *Journal of Huaihua University*, (3):30.

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. Association for Computational Linguistics.

Yves Scherrer. 2019. Lecture notes in computational syntax.