Exploring the Effectiveness of Topic Modelling as an Auxiliary Task for POS Tagging

Angelo Basile
University of Malta / Msida, Malta (MT)
angelo.basile.17@um.edu.mt

Abstract

Most recent work in NLP is showing that multi-task learning can give an advantage in many scenarios. In this paper we investigate the use of topic modelling as an auxiliary task for pos tagging using a state-of-the-art bidirectional LSTM tagger. We simulate a low-resource scenario and we argue that since topic modelling can be done in an unsupervised way, in case it works as an auxiliary task it can help indeed when data are scarce. We experiment on Italian with different network architectures and topic models. The results are not yet satisfactory.

1 Introduction

Part-of-Speech (POS) tagging is the task of assigning to each word in a text a tag that describes its corresponding grammatical category. This task is fundamental in almost every natural language processing pipeline: parsing, named entity recognition, machine translation, authorship attribution are just a few examples of NLP tasks that can greatly benefit from a layer of POS annotation.

The problem is traditionally modelled as a sequence labelling — or structured prediction — task: in the past, have been successfully used Hidden Markov Models (Cutting et al., 1992), Decision Trees (Schmid, 1994b) and neural networks since the early 90's (Schmid, 1994a).

The performance of taggers varies greatly depending mostly on the domain: the performances per token for state-of-the-art taggers are around 97% in accuracy but performances per sentences are much lower (i.e. around 55% accuracy): these numbers are part of a interesting discussion in (Manning, 2011). Most taggers are trained on newspapers and score high results when evaluated in the same domain, but other domains (e.g. Twitter and other social media) are either more difficult to handle or simply there is a need for more work. Plank (2016) discuss this problem and show that indeed tagging performances rapidly decrease when train and testing domains differ.

In this work we try to build upon recent work of Plank et al. (2016), who show that neural networks can be used also with relatively small datasets. We simulate a low-resource language scenario working with Italian and using up to three thousands annotated sentences: we want to test if it is possible to improve the performance of a state-of-the-art tagger by introducing topic information about the sentence that was gather in a completely unsupervised manner. By combining the findings of Plank et al. on the effectiveness of neural networks with the idea that topic modelling can be done in an unsupervised fashion, we argue that if this system works, then it can be used for real low-resourced languages, assuming that at least a medium sized non-annotated corpus is available for training a topic model.

For including topic information in the tagger, we model the problem as multi-task learning (MTL) problem, where the main task is POS tagging and topic classification is the auxiliary task: the results on MTL in natural language processing reported by Søgaard and Bingel (2017) show that it is not always the case that MTL helps, but it works in many combinations of tasks and subtasks.

Our hypothesis to be proved here is the following: the topic of a text can help the disambiguation of ambiguous words and therefore it can help in tagging.

We assign to each training instance a topic using unsupervised methods (i.e. Latent Semantic Analysis (Deerwester et al., 1990) and Latent Dirichlet Allocation (Blei et al., 2003)) and we add an extra

layer to our POS tagging network that predicts the topic.

Contributions In this paper we investigate if topic can help as an auxiliary task in POS tagging. We propose this approach as a potential solution for low-resourced languages. The code and the trained model for Italian are available online at https://github.com/anbasile/mtl-tagging.

2 Data

The corpus used to conduct the experiments come from the CoNNL 2007 shared task (Nivre et al., 2007) on Parsing and it is a subset of the Italian ISST Treebank (Fanciulli et al., 2003). It consists of 3359 sentences from newspapers (namely *La Repubblica* and *Il Corriere della Sera*) and periodicals that were selected with the aim of balancing topics. Each sentence is annotated in the CoNNLL format; the tagset consists of 12 different tags. We choose to work on this dataset because we wanted our results to be comparable with those coming from Li et al. (2012) which exploit a freely available resource (i.e. *Wiktionary*) for improving the performance in the same simulate low-resource scenario.

3 Experiments

To implement our model, we use the *dyNet* toolkit (Neubig et al., 2017): it makes it easy to build a multi-headed model thanks to its underlying graph-based system. For the topic model, we use the *Gensim* framework (Řehůřek and Sojka, 2010). All the code is written using Python version 3.5 and the required libraries are listed in the project's repository for making the experiments easily reproducible.

3.1 Topic Modelling

A word is characterized by the company it keeps (Firth, 1968). This is the fundamental idea behind vector space models and distributional semantics: words that have the same meaning tend to occur in the same contexts. This contexts or clusters of meaning can then be condensed in *topics* and this is what techniques like Latent Semantics Analysis (LSA) do.

In this work we try to exploit topic to help the tagger disambiguate words and hence improve its performance.

One of the issues with topic modelling — at least in the way it is usually done — is that the results are not immediate to interpret: each documents gets assigned a topic label which has no meaning and each topic can be seen as just a list of the salient words that define it. For our experiments this is not a problem, since we use the topic labels just to eventually improve the tagger.

We build two different topic models, one using LSA and one using LDA, setting the number of topics to 50 for both models: we choose this number considering that the number of section that usually forms a newspaper (most of the sentences in the training corpus come from these) and also considering the fact that a small number of periodicals is included in the corpus. We manually inspected the results and after hypothesizing an interpretation of the topics we tried different preprocessing to improve the modelling: for our final models, in both cases, we decided to lemmatize the text, filter out stop words and punctuation and apply a tf-idf transformation.

The output of this stage consists of a list of ids which is paired with each training sentence and these ids are predicted by our model as an auxiliary task to tagging.

3.2 Tagging

Our tagger is heavily based on the Bilty tagger from Plank et al. (2016). The structure of the network consists of an embedding layer — we use no pre-trained embeddings — and two LSTM layers that read the text in both direction; furthermore, in some experiments we use a Multi-Layer Perceptron (MLP) at the end that predicts the tags; a second MLP, connected directly to the LSTM layers, predicts the topic. Figure 1 shows the structure of the network.

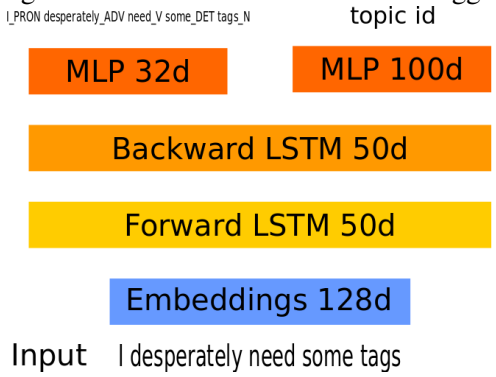
We use Stochastic Gradient Descent to optimize the weights with the default learning rate of 0.1; we perform various experiments where we train the network for 20 epochs and for 100 epochs.

4 Evaluation and Results

Metric Since we are interested only in assessing whether the topic plays a role or not in a multi-task learning scenario with POS tagging, we decided that accuracy per token alone is a good metric.

Baseline As a baseline, we have the TreeTagger from Schmid (1994b), which is a probabilistic tagger based on decision trees: this system scored

Figure 1: The structure of our neural tagger.



75% in accuracy.

The test set consists of 249 documents and is a split of the same corpus that was used for training.

The results are highlighted in Table 1. The taggers performs overall almost as state-of-the-art, even more considering the size of the dataset. We notice that training epochs does not have a significant effect on the results: at 20 epochs the system can already perform well.

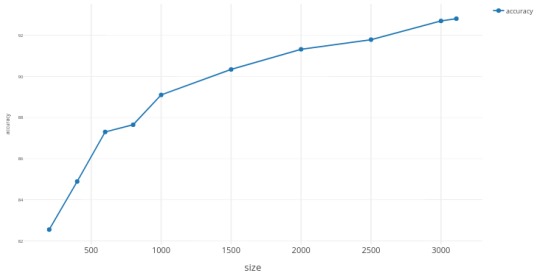
topic	MTL	epochs	accuracy
yes	no	20	92.46
yes	no	100	92.71
no	no	100	92.75
yes	yes	100	92.81
no	yes	100	**92.99**
yes	yes	20	92.58
no	yes	20	92.48

Table 1: Results of our neural POS tagger. MTL stands for Multi-Layer Perceptron: we experimented with having or not such a layer as an output layer for predicting the tags. For predicting topic we always use it.

Going to the heart of our work — assessing whether topic modelling helps or not in POS tagging — we can see that the model that is not predicting the topic is consistently better than its counterpart which is. We experimented with different numbers of training epochs because we hypothesized that an auxiliary task can require extra training in order to be properly learned, but we have to reject this hypothesis. Furthermore, the results are consistent across system architectures — with and without an extra MLP layer for tag prediction — which means that somehow the auxiliary task is always confusing the model.

We built a learning curve in order to understand how much data does the network need to achieve state-of-the-art performances: Figure 2 shows the results.

Figure 2: The learning curve of the network, going from 200 sentences to 3110.



Discussion At this point we could conclude that topic modelling does not help in POS tagging and not only that, but it also confuses the system. Such a conclusion would however be premature.

Before discussing the main results, we want to highlight that we confirm a recent finding in the literature: neural networks can perform well with limited training instances and indeed our tagger trained with only 200 sentences outperforms the baseline by a significant margin (i.e. 82% in accuracy vs. 75%).

Now, considering the main research question of this paper we have three important points to discuss.

First, topic modelling with very short texts is a hard task: what is the topic of a sentence? We tried our best with this dataset, because we wanted anyway to simulate a low-resource scenario, but we need to repeat the experiments with either gold labels (e.g. having the section of the newspaper from which the sentences are from) or with longer texts.

Second, we might have been adding redundant information — which could translate to noise — to the system: word embeddings and LSA models can be seen to some extent as two ways of achieving very similar results. Are perhaps embeddings already capturing topic?

Third, we know from Søgaard and Goldberg (2016) that some auxiliary tasks are better predicted at some particular levels and although it seems that lower levels are better suited for predicting tasks related to a lower linguistic abstraction, it is not clear whether to predict topic. What would happen if we would predict it at another level?

5 Related Work

In the Introduction we mention some related work that we used for our project, but plenty of other research has been done for improving the performances of systems for low-resource languages. Agic et al. (2015) show that good results can be obtain in extreme condition, by using only the Bible which is translated in many under-represented languages; discrete results can be obtain with no supervision at all (Das and Petrov, 2011); Li et al. (2012) show that freely available language resources can be used in creative ways to improve the performances of simple models when data are scarce; tagging performance can be also be improved by adding non-textual features to the model, such as demographic information of the author of the text to be tagged (Hovy and Søgaard, 2015). On adding additional features (such as a topic) in a neural network model, Le and Zuidema (2014) propose an interesting architecture.

6 Conclusions

We built a multi-task learning system for POS tagging hoping that predicting the topic of a text could give an advantage that would be useful in low-resourced settings, since topic can be obtained in an unsupervised way. Basically, we failed. Before reporting these results as negative results, we believe however that a follow-up study with gold topic labels and more data would be needed.

A part from that, we confirm a recent finding in the literature. neural networks can be used with a relatively small dataset.

Acknowledgments

This paper originates from the class of Advanced Issues in Language Technology, held in Malta in 2018: I am very grateful to prof. Barbara Plank for all her patience. I also want to thank my friends and colleagues Kenny W. Lino and Jovana Urosevic for having hosted me in their home while taking the course.

References

Zeljko Agic, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the bible: Learning pos taggers for truly low-resource languages. In *ACL*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Douglas R. Cutting, Julian Kupiec, Jan O. Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *ANLP*.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL*.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS* 41:391–407.

Filippo Fanciulli, M. Massetani, Ricciarda Raffaelli, Maria Teresa Pazienza, Dan Saracino, and Fabio Massimo Zanzotto. 2003. The italian syntactic-semantic treebank: Architecture, annotation, tools and evaluation.

John Rupert Firth. 1968. *Selected papers of JR Firth, 1952-59*. Indiana University Press.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *ACL*.

Phong Le and Willem H. Zuidema. 2014. The inside-outside recursive neural network model for dependency parsing. In *EMNLP*.

Shen Li, João Graça, and Ben Taskar. 2012. Wikily supervised part-of-speech tagging. In *EMNLP-CoNLL*.

Christopher D. Manning. 2011. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *CICLing*.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan T. McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The conll 2007 shared task on dependency parsing. In *EMNLP-CoNLL*.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp. *CoRR* abs/1608.07836.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *CoRR* abs/1604.05529.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New*

Challenges for NLP Frameworks. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

Helmut Schmid. 1994a. Part-of-speech tagging with neural networks. In *COLING*.

Helmut Schmid. 1994b. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Anders Søgaard and Joachim Bingel. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *EACL*.

Anders Søgaard and Yoav Goldberg. 2016. Deep multi-task learning with low level tasks supervised at lower layers. In *ACL*.