Training Googles SyntaxNet to understand Norwegian Bokmål and Nynorsk

Bjarte Johansen
Department of Information Science and Media Studies
University of Bergen
bjarte.johansen@uib.no

Abstract

We use Google's open source neural network framework, SyntaxNet, to train a fully automatic part-of-speech language tagger for Norwegian Bokmål and Nynorsk. Using SyntaxNet, we are able to get comparable results with other tagger when tagging Bokmål and Nynorsk with part-of-speech. Both taggers are available and released as open source.

1 Introduction

We use Google's SyntaxNet, an open source neural network framework, to train a part-of-speech (PoS) tagger for Norwegian bokmål and nynorsk (Andor et al., 2016). PoS are categories of words that share the same grammatical properties. Examples of PoS are nouns, verbs, adjectives etc. PoS tagging can help computers reason about text by disambiguating the meaning of words in context. A word like "tie" is a noun in the sentence "His tie had an interesting pattern", but would be a verb in the sentence "He wasn't able to tie a sailor's knot."

We show that using an off the shelf, state of the art, learning platform allows us to create PoS taggers for both of the Norwegian written languages that are at, or exceed, what other researchers are able to do on the same task. Both PoS taggers are available online at `http://github.com/ljos/anna_lyse` and are released under an open source license. The nynorsk tagger is the only PoS tagger, as far as we know, available that produces unambiguous output ready for use in other machine learning tasks.

In this paper we

1. show how we ran the training and experiments for our taggers and present the results in section 4.
2. discuss the results and compare them to what other researchers have achieved in section 5. We also present some of the issues and problems with this way of doing PoS tagging.
3. lastly we discuss what we want to do with this research in the future in section 6.

This paper was presented at the NIK-2016 conference; see http://www.nik.no/.

2 Related work

Marco (2014) used the FreeLing open source text processing tool to create a PoS tagger and uses a hidden Markov model (HHM) to find the tags; they achieve an $F_{\beta=1}$ score of 97.3% for bokmål.

Hagen et al. (2000) created a tagger based on a morphological constraint grammar; the Oslo-Bergen Tagger (OBT). As they are interested in using the tagger as a tool for linguists to search for specific grammatical structures, OBT reports any ambiguities that it finds. They report a precision of 96.0% and a recall of 99.0%, which results in a $F_{\beta=1}$ score of 97.5%, for bokmål (Bick et al., 2015). For tagging nynorsk they reports a precision of 93.6% and a recall of 98.7, with a $F_{\beta=1}$ score of 96.2.

Johannessen et al. (2011) added a statistical disambiguator to the bokmål part of OBT based on a HMM approach. They achieve an $F_{\beta=1}$ score of 96.56%, but without any ambiguities in the output.

Using the Universal Dependency data set (Øvrelid and Hohle, 2016), Google was able to train SyntaxNet to tag bokmål with PoS at an $F_{\beta=1}$ score of 97.44% (Google, 2016b).

3 SyntaxNet

SyntaxNet is a "feed-forward neural network that operates on a task-specific transition system." It is not recurrent and uses beam search together with a conditional random field (CRF) objective to globally normalize the learning model. They also perform full backpropagation for all neural network parameters based on the CRF loss (Andor et al., 2016).

4 Evaluation

Recently Google released the source code to SyntaxNet, their neural network framework for syntax learning. They provide a detailed explanation for how to use SyntaxNet for learning new languages and examples on the web page for the source code for SyntaxNet (Google, 2016a). We decided to run a simple grid search using the Norwegian Dependency Treebank for both the bokmål and nynorsk version which follows the Norwegian Reference Grammar (Solberg et al., 2014).

The task we are trying to achieve is to correctly classify the PoS for every token in the data set.

The experiment followed a very simple setup. For each language we split the data set into a training set containing 50% of the sentences, a test set containing 25% and, a verification set containing the remaining 25%. As the data set contains sentences from different sources like newspaper text, government reports, parliament transcripts and, blogs, we first randomized the order of the sentences to get a fair distribution of each type of text in all of the data sets.

We then used a grid search to find the best performing parameters for training SyntaxNet. SyntaxNet have many different parameters that we can change, but we focused on the layer size, learning rate and, momentum as these are the ones that the people behind SyntaxNet recommends. Layer size control how many neurons are in each layer, learning rate is how fast each neuron learns by acting on the weight update of the backpropagation alorithm and, momentum helps the update gradient of backpropagation keep moving in the same direction. Even though we were able to run some of these training sessions in parallel it still took many days to run through all of the variations in the grid search.

Language	$F_{\beta=1}$
Bokmål	97.54%
Nynorsk	96.83%

Table 1: Results of training SyntaxNet.

The results from the grid search are available in table 1, but we were able to get an $F_{beta=1}$ score of 97.54% for bokmål and 96.83% for nynorsk.

5 Discussion

As one can see from table 1 we have been able to get good scores for both of our taggers. Comparing to the other taggers we presented in section 2 we can see that we are slightly better than all of the previous attempt at creating a PoS tagger for both bokmål and nynorsk.

There is one caveat, comparing to the OBT is somewhat problematic as it also report ambiguities and ours do not. We argue that to compare them we should look at the precision of OBT and not the F-score as that is the measure of when OBT is able to unambiguously tell if a tag is the correct one or not. Comparing against the F-score or recall would not work as it is measuring something else than we are measuring. The precision is the score that would be closest to what we are trying to measure.

We can also see that the tagger from Google gets almost the same score as we do; we believe that is because the Universal Dependency data set is the same data set we are using, just with a translated tag set.

There are also some problems with the current implementation of the taggers we have presented here. The first issue is that the tagger only accepts tokenized text, which means we cannot run the tagger on just plain text documents; they need to be preprocessed first.

The second is that since the method used involves neural nets, it is questionable what we can learn from analyzing the process itself to see what we can learn about Norwegian language from how the taggers are trained and operate.

Further, if one is interested in also capturing the ambiguity in the language and not always be presented with what the machine calculates to be the correct answer, the approach chosen by the OBT is better. If OBT detects a word that can be tagged in multiple different ways it will present all of them if it cannot chose. Our approach will always choose one definite answer, correct or not.

6 Future work

One of the problems we talked about in the discussion was that there doesn't exit a tokenizer for Norwegian. The OBT does contain one, but, as of now, it is not trivial to separate it from the tagger itself. Getting a standard and good tokenizer for both Norwegian languages would make it easier to develop new PoS tagger. It would also make it easier to experiment with different tokenization processes etc.

SyntaxNet doesn't just support PoS tagging, it is possible to also do dependency parsing. It would be quite trivial to adapt this project to also do this, but we did not attempt it as we didn't have the time to run another round of training.

There are many ways to experiment with SyntaxNet to see if it is possible to improve on the current taggers. We mostly followed the standard setup and tested different changes to the parameters that the creator behind SyntaxNet suggested. One could f.ex. change the features that the tagger looks at instead of just the standard set. We also believe that it could be possible that if we had larger data sets to learn from that the taggers could perform better. We see evidence of this in that the biggest difference between the tagger for bokmål and nynorsk is that the data set for bokmål is larger and it performs better.

Further, this project can help improve results in Named-Entity Recognition and other chunking tasks for both bokmål and nynorsk by providing a more accurate base to build from.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*, 2016.

Eckhard Bick, Kristin Hagen, and Anders Nøklestad. Optimizing the oslo-bergen tagger. In *Proceedings of the Workshop on "Constraint Grammar-methods, tools and applications" at NODALIDA 2015, May 11-13, 2015, Institute of the Lithuanian Language, Vilnius, Lithuania*, number 113, pages 11–17. Linköping University Electronic Press, 2015.

Google. Syntaxnet: Neural models of syntax. `github.com/tensorflow/models/tree/master/syntaxnet`, 2016a. Accessed: 2016-08-23.

Google. Parsey's cousins. `github.com/tensorflow/models/blob/master/syntaxnet/universal.md`, 2016b. Accessed: 2016-08-23.

Kristin Hagen, Janne Bondi Johannessen, and Anders Nøklestad. A constraint-based tagger for norwegian. In *Odense Working Papers in Language and Communication 19*, 17th Scandinavian Conference of Linguistics, pages 31–48, 2000.

Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad, and André Lynum. Obt+ stat: Evaluation of a combined cg and statistical tagger. *Constraint Grammar Applications*, pages 26–34, 2011.

Cristina Sánchez Marco. An open source part-of-speech tagger for norwegian: Building on existing language resources. In *LREC*, pages 4111–4117, 2014.

Lilja Øvrelid and Petter Hohle. Universal dependencies for norwegian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2016.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. The norwegian dependency treebank. In *LREC*, 2014.