Tagging a Norwegian Dialect Corpus

Andre Kåsen	Kristin Hagen	Anders Nøklestad	Joel Priestley
Department of Informatics	The Text Laboratory	The Text Laboratory	The Text Laboratory
University of Oslo	University of Oslo	University of Oslo	University of Oslo

andrekaa@ifi.uio.no {kristin.hagen, anders.noklestad, joel.priestley}@iln.uio.no

Abstract

This paper describes an evaluation of five data-driven Part-of-Speech (PoS) taggers for spoken Norwegian. The taggers all rely on different machine learning mechanisms: decision trees, hidden Markov models (HMMs), conditional random fields (CRFs), long-short term memory networks (LSTMs), and convolutional neural networks (CNNs). We go into some of the challenges posed by the task of tagging spoken, as opposed to written, language, and in particular a wide range of dialects as is found in the recordings of the LIA (Language Infrastructure made Accessible) project. The results show that the taggers based on either conditional random fields or neural networks perform much better than the rest, with the LSTM tagger getting the highest score.

1 Introduction

The most commonly used PoS tagger for Norwegian is the the Oslo-Bergen tagger (OBT); a Constraint Grammar tagger for Bokmål and Nynorsk (Johannessen et al., 2012), the two written standards that exist for written Norwegian. For spoken language transcribed into Bokmål, the statistical NoTa tagger was developed and trained on Bokmål transcriptions from Oslo and the surrounding area (Nøklestad and Søfteland, 2007). A recent infrastructure project, LIA (Language Infrastructure made Accessible) has produced a large number of dialect transcriptions in Nynorsk, the other written standard. This creates a need for a new tagger that works on this written standard and that can also handle a diverse data set containing a wide range of dialects.

In this paper we will first describe the LIA dialect transcriptions and then the manually annotated training material for Nynorsk as well as some challenges in annotating spoken language. Afterwards we will describe a number of experiments with five different open source taggers.

2 Dialect transcriptions

The audio files were recorded between 1950 and 1990 in order to explore and survey the many different dialects in Norway. Most of the informants are older people and native speakers of their dialect. Typically, the recordings are interviews about old trades such as agriculture, fishing, logging and life at the summer farm. Other topics are weaving, knitting, baking or dialects. Sometimes the research questions also concern person or place names. The recordings are semi-formal to informal and often take place in an informant's home.

The original LIA transcriptions are semi-phonetically transcribed (Hagen et al., 2015). Example (1) below shows the semi-phonetic and normalized transcription. To make the transcriptions searchable and suitable for automatic tagging, they are semi-automatically transliterated to Nynorsk by the Oslo Transliterator, which is trained on more than 200 Norwegian dialects.

(1)	*hann*	*e*	*flinngke*	*te*	*driva*	*garen*
	han	er	flink	å	drive	garden

‘He is good at running the farm.’

Øvrelid et al. (2018) note that the segmentation heuristic in this material is such that segments do not necessarily correspond to sentences, but rather to (conversational) meaningful units.

3 The Training Corpus, Dialects and Spoken language PoS

The starting point was the annotation scheme of the Norwegian Dependency Treebank (NDT) described by Solberg et al. (2014). This is an extension of the OBT scheme (which is based on

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 350–355
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

(Faarlund et al., 1997)) with additions necessary for NDT. Table 1 shows the PoS tag set of the training corpus.

PoS tag	Description
adj	Adjective
adv	Adverb
det	Determiner
inf-merke	Infinitive marker
interj	Interjection
konj	Conjunction
nol	Hesitation
pause	Pause
prep	Preposition
pron	Pronoun
sbu	Subordinate conjunction
subst	Noun
ufullst	False start
verb	Verb

Table 1: The PoS tag set of the training corpus.

In addition to the traditional PoS classes, there is one for hesitations *nol*, one for pauses *pause* and one for false starts *ufullst*. Unlike the classification in British National Corpus where all these unclassified words seem to be classified as UNC (Burnard (2007))[1] this solutions gives us the possibility to experiment with the different types of pauses, hesitations etc., see the result chapter and the description of the different categories further below.

The manually corrected training corpus contains 163,687 tokens from 37 transcriptions and 29 dialects as listed in table 2, whereas the geographical distribution of the data is shown in figure 1.

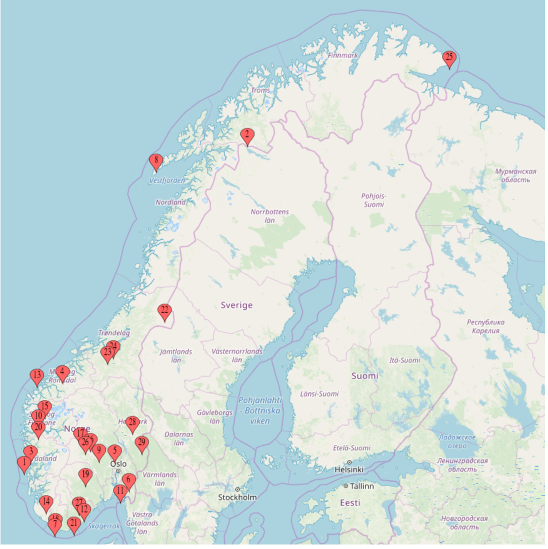

Figure 1: The map shows the locations of the 29 dialects in the training corpus.

Next we discuss some challenges encountered in spoken language when moving from an annotation scheme primarily developed for written language. Or as Miller and Weinert (1998) put it: “The terms ‘spoken language’ and ‘written language’ do not refer merely to different mediums but relate to partially different systems of morphology, syntax, vocabulary, and the organization of texts.”

Transcription was conducted in accordance with transcription guidelines (Hagen et al., 2015) that stipulate a strict, verbatim representation of speech, regardless of fluency or perceived correctness. Some frequent categories of phenomenon in speech have to be considered in this respect:

Disfluency (as described in Shriberg (1996)) is a category that goes beyond PoS tags, but has some relevance at the word level. For example, incomplete or interrupted words, i.e. false starts of different kinds, have to be tagged, and while such words are transcribed as far as possible, interruption and incompleteness are marked with a hyphen - (see example 2). False starts are marked with the tag *ufullst* ‘incomplete’. Pauses, which we transcribe with the ‘#’ symbol, are stops or interruptions in the speech flow of the speaker. We have marked them with the tag *pause*. Filled pauses or hesitations are standardized as *ee* and tagged *nol*.

[1] See in particular chap. 6 *Wordclass Tagging in BNC XML*

Dialect area	**# segments**	**# tokens**
Austevoll	1193	11191
Bardu	560	4205
Bergen	993	10416
Bolsøy	645	6669
Brandbu	404	6112
Eidsberg	679	5880
Farsund	351	3707
Flakstad	1201	11080
Flå	149	2808
Førde	332	3175
Fredrikstad	554	7676
Froland	378	6660
Giske	874	10821
Gjesdal	415	4101
Gloppen	526	5724
Gol	158	2414
Hemsedal	244	4436
Herad	214	2186
Hjartdal	354	4032
Høyanger	330	4357
Kristiansand	259	3713
Lierne	365	3867
Skaun	482	4661
Trondheim	216	3392
Vardø	481	6055
Ål	542	8685
Åmli	212	3128
Åmot	423	5123
Åsnes	466	7413
Total	14000	163687

Table 2: The manually corrected training corpus contains tokens from 37 transcriptions and 29 dialects.

(2) *så det var ganske m- # ee*
so it was very ufullst pause nol
mange der
many there
'There were a lot of people there.'

Another challenge is frequent and form-identical words. For example, sentential connectives or conjunctions are a well delimited group of words in written Norwegian. In spoken Norwegian, however, the usage patterns of certain words have yet to be examined, and the difference between certain conjunctions and pragmatic markers/particles is somewhat unclear.[2]. For instance, *så* seems to take on multiple functions:

(3) *så Kari løp fort*
so Kari ran fast.
'so Kari ran fast.'

The next two examples illustrate another challenge. Adverbs, interjections and particles are far more common in spoken language than in written text. The pragmatic particle *lell* probably has a function like the adverb *heller* ('just as well'), or some sort of particle. Then in example (5) we see a somewhat similar use pattern, but with a token that is form-identical with the conjunction *eller* ('or'). In both cases, we have chosen to tag the words as adverbs.

(4) *men huttetu eg greidde nå ikkje å*
but my I could PART not TO
sjå på det lell
see on it PART
'oh my I couldn't look at it.'

(5) *er det langt for deg å reise til #*
is it long for you TO travel to pause
til jobben da eller?
to work then or?
'do you have a long travel to work?'

Håberg (2010) describes and analyzes what is known as the preproprial article, which is form-identical with the third person pronoun:

(6) *så dæ skræiv hu F1 en særåppgave*
so then wrote she F1 a paper
omm dæ
about you
'Then F1 wrote a paper about you.'

The analysis given by Håberg (2010) states that the function of the preproprial article is more akin to that of a determiner, and therefore constitutes an ambiguity between the tags *det* and *pron*. In both of the cases above a heuristic that only considered form was employed, i.e. the preproprial article is tagged *pron*. Note also that the preproprial article is close to non-existent in written language.

Other problems that can be considered are variable word order in embedded structures (Rognes, 2011) or form-identical subjunctions and prepositions (Huus, 2018). To draw an intermediate conclusion, we can say that an investigation like that of Hohle (2016) is called for with regard to spoken language.

[2] Several case studies can be found in the special issue on pragmatic particles of The Norwegian Linguistic Journal http://ojs.novus.no/index.php/NLT/issue/view/196/showToc

4 Taggers

In order to find the most suitable tagger, an array of different taggers from different paradigms were tested. In the following, we give a short description of the systems in use in the present paper, along with references to them.

TreeTagger[3] In order to keep some continuity with the aforementioned NoTa tagger, new models were induced for the TreeTagger. TreeTagger is based on the decision tree paradigm (Schmid, 1999), and was shown by (Nøklestad and Søfteland, 2007) to be the best performing system for the NoTa data set.

TnT[4] is a second order HMM tagger (Brants, 2000). It has been used on multiple occasions (see Hohle et al. (2017), Velldal et al. (2017)) to tag Norwegian. It is therefore natural to include it among the systems in the present paper.

MarMoT[5] is a generic CRF tagger (Müller et al., 2013), and is widely used as a baseline tagger. It can with relative ease be extended to include morphological tags as well which is a natural next step for the present work.

Bilstm-aux[6] is a bidirectional LSTM tagger with auxiliary loss that has been shown to work well for Norwegian (Plank et al., 2016). Plank et al. (2016) report a tagging accuracy of 98.06% for the Norwegian part of the Universal Dependency Treebanks v1.2 (Nivre et al., 2015). The Norwegian UD part is the NDT mentioned earlier, converted to the UD standard (see (Øvrelid and Hohle, 2016; Øvrelid et al., 2018)).

Sclem2017-tagger[7] is a general purpose tagger utilizing a CNN with a character composition component and a context encoder (Yu et al., 2017). Yu et. al (2017) report a accuracy of 97.65% for Norwegian UD.

5 Results

In the current work, we only tested the performance of the taggers on the entire corpus, not on individual dialects, for several reasons:

First, there is considerable variation in the amount of material we have for the different dialects, preventing a balanced comparison between dialects. Furthermore, for many of the dialects the size of the material is too small to yield a reliable evaluation. Finally, the transcription into standard orthography by necessity removes parts of what distinguishes the dialects, in particular with respect to morphological features, and the amount of normalization is highest for those dialects that differ the most from standard written Nynorsk, again preventing a fair comparison of dialects.

All systems were evaluated intrinsically using 10-fold cross validation and reported with accuracy. Care has been taken to ensure that each fold has the relative equal distribution of dialects as the whole data set to prevent skewed folds. After splitting the whole data set (80-10-10) evenly w.r.t. dialects and distributing the 80% portion into 10 folds each with a hold out portion, the data was randomized. Table 3 shows the calculated accuracy for all the systems with the respective standard deviation for the ten folds. As is evident, the top performing taggers have relatively similar scores, but according to McNemars test, Bilstm-aux performs significantly better than the next best tagger, MarMoT ($p < 0.05$), and it also shows a somewhat smaller standard deviation. For the best system we also add a table for each PoS tags precision and recall (Table 4).

System	**Accuracy (std.)**
TreeTagger	95.16 (0.0020)
TnT	93.18 (0.18)
MarMoT	97.25 (0.14)
Sclem2017	97.16 (0.15)
Bilstm-aux	**97.33** (0.11)

Table 3: The PoS accuracy and standard deviation for the 10-fold cross validation for each system.

Both Sclem2017 and Bilstm-aux are evaluated with their integrated test function, whereas MarMot and TreeTagger are evaluated with an ad hoc python script. What sets these systems apart is the fact that the neural networks are given a development set at training time for early stopping purposes, while MarMot demands brown-like clusters induced with Marlin [8][9]

[3] https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[4] http://www.coli.uni-saarland.de/ thorsten/tnt/

[5] https://github.com/muelletm/cistern/tree/master/marmot

[6] https://github.com/bplank/bilstm-aux

[7] https://github.com/EggplantElf/sclem2017-tagger

[8] https://github.com/muelletm/cistern/tree/master/marlin

[9] Marlin was trained with the Nynorsk part of the Habit corpus and the Norwegian Newspaper Corpus.

(Martin et al., 1998; Müller and Schütze, 2015). This is most likely one of the reasons it performs so well compared to the neural taggers, and call for an investigation of neural taggers with pre-training as well, i.e. neither of the neural taggers was trained with pre-trained word embeddings.

PoS tag	Presicion	Recall
adj	89.45	90.87
adv	96.64	94.90
det	94.03	92.95
inf-merke	97.07	98.17
interj	99.32	99.08
konj	96.42	97.95
nol	100	100
pause	100	100
prep	97.77	98.11
pron	98.53	98.65
sbu	92.05	91.97
subst	95.85	96.86
ufullst	97.81	99.06
verb	98.20	98.20

Table 4: The precision and recall (averaged across all 10 folds) for the best performing system: Bilstm-aux (Plank et al., 2016)

5.1 Removal of pauses, hesitations and pauses+hesitations

In the style of Nøklestad and Søfteland (2007), evaluations where different speech specific tokens were removed were also carried out. Nøklestad and Søfteland (2007) report that this in fact lowered the performance of the systems they tested. The results that were obtained from the two best performing systems in the present paper are found in Table 5.

System	Accuracy (std.)
MarMoT$_{hesitations}$	97.19 (0.001)
Bilstm-aux$_{hesitations}$	97.27 (0.1)
MarMoT$_{pauses}$	97.08 (0.001)
Bilstm-aux$_{pauses}$	97.17 (0.14)
MarMoT$_{hesitations+pauses}$	97.03 (0.001)
Bilstm-aux$_{hesitations+pauses}$	97.07 (0.18)

Table 5: The PoS accuracy and standard deviation for the 10-fold cross validation with speech specific tokens removed. The subscripts indicate what kinds of tokens are removed in each case.

The accuracy deteriorates as speech specific tokens are removed, and for both systems removal of pauses have a greater impact on the accuracy than hesitations. This supports the findings by (Strangert et al., 1993) that pauses tend to occur at important positions in an utterance, including syntactic boundaries, and hence may provide important clues about the syntactic structure.

6 Conclusions and Further Work

The present paper has reported on new results for PoS tagging of Norwegian dialect data. It has also shown that, among the tagger technologies tested, the ones based on CRFs or neural networks show the best performance on this task.

A subset of the training material in this paper constitutes the LIA Treebank of Spoken Norwegian Dialects and it would be interesting to investigate whether removal of other phrasal disfluencies than the ones already tested would have an impact on the final accuracy score (see Dobrovoljc and Martinc (2018) and references therein). It would also be worth the effort to see whether neural taggers respond better if the input is semi-phonetic rather than normalized. Finally, if we are able to produce a considerable amount of material for a set of dialects, transcribed in a way that is more faithful to the peculiarities of each dialect, it would be interesting to test and compare the performance of the taggers on individual dialects.

References

Thorsten Brants. 2000. Tnt: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*, pages 224–231.

Lou Burnard. 2007. Reference guide for the british national corpus. Technical report.

Kaja Dobrovoljc and Matej Martinc. 2018. Er ... well, it matters, right? on the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 37–46.

Jan Terje Faarlund, Svein Lie, and Kjell Ivar Vannebo. 1997. *Norsk referansegrammatikk*. Universitetsforlaget.

Live Håberg. 2010. Den preproprielle artikkelen i norsk: ei undersøking av namneartiklar i kvæfjord, gausdal og voss. Master's thesis.

Kristin Hagen, Live Håberg, Eirik Olsen, and Åshild Søfteland. 2015. Transkripsjonsrettleiing for lia. Technical report.

Petter Hohle. 2016. Optimizing a pos tag set for norwegian dependency parsing. Master's thesis.

Petter Hohle, Lilja Øvrelid, and Erik Velldal. 2017. Optimizing a pos tagset for norwegian dependency parsing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 142–151.

Andrea Myklebust Huus. 2018. Distribusjonen av te som infinitivsmerke i norsk: En korpusbasert undersøkelse av utbredelsen av te som infinitivsmerke i norsk. Master's thesis.

Janne Bondi Johannessen, Kristin Hagen, André Lynum, and Anders Nøklestad. 2012. Obt+ stat. a combined rule-based and statistical tagger. In *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*.

Sven Martin, Jörg Liermann, and Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. *Speech communication*.

James Edward Miller and Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford University Press.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order crfs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332.

Thomas Müller and Hinrich Schütze. 2015. Robust morphological tagging with word representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, et al. 2015. Universal dependencies 1.2.

Anders Nøklestad and Åshild Søfteland. 2007. Tagging a norwegian speech corpus. In *Proceedings of the 16th Nordic Conference of Computational Linguistics*, pages 245–248.

Lilja Øvrelid and Petter Hohle. 2016. Universal dependencies for norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1579–1585.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The lia treebank of spoken norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 4482–4488.

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*, pages 412–418.

Stig Rognes. 2011. V2, v3, v4 (and maybe even more): The syntax of questions in the rogaland dialects of norway. Master's thesis.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of International Conference on Spoken Language Processing*, volume 96, pages 11–14.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The norwegian dependency treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 789–795.

E. Strangert, E. Ejerhed, and D. Huber. 1993. Clause structure and prosodic segmentation. In *FONETIK-93 Papers from the 7th Swedish Phonetics Conference*.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint ud parsing of norwegian bokmål and nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10.

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.