The shortcomings of a tagger

Kristin Hagen, Janne Bondi Johannessen og Anders Nøklestad
The Text Laboratory, The University of Oslo
{kristiha, jannebj, noklesta}@hedda.uio.no

Abstract
The tagger used for the Oslo Corpus of Tagged Norwegian Texts has very good statistical results. In spite of this, it makes mistakes. In this paper we take a closer look at some of them. Although some mistakes are of a kind that would disappear if we improved the tagger, many are impossible or very difficult to do anything about. They are due to errors in the corpus (spelling errors, foreign words, non-standard spellings), to elliptic sentences, such as headlines, and to structural ambiguity, which abounds to a surprising extent. Proofreading the corpus would have removed the first kind of problems, but the other two types cannot be resolved in any obvious way.

1. Introduction

The first version of the first ever comprehensive tagger for Norwegian is ready. Both the *nynorsk* and the *bokmål* (the two Norwegian language varieties) versions have been used to tag a large number of texts (= the Oslo Corpus of Tagged Norwegian Texts). The corpus has an advanced web-based user interface, which often gives nice results, but it also makes it easy to discover mistakes and shortcomings of the tagger. The present paper will focus on these.

The tagger is of a Constraint Grammar-type (Karlsson et al 1995). The linguistic constraints (rules) were developed by the Text Laboratory, while the software came from Lingsoft, Helsinki. A CG tagger takes as input a multitagged text, where each word form has as many tags as the lexicon allows, and gives as output a text where the tags are disambiguated by the given linguistic constraints according to the context for each word in question. The statistical results are good: The bokmål tagger has a recall of 99,2% and a precision of 96,8%. For nynorsk the results are slightly worse: 98,8 % recall and 95,6 % precision.

The tagger, then, makes some mistakes. One kind of shortcoming involves cases where some ambiguity remains (this influences the precision rate) - for a number of reasons, of which structural ambiguity is the most severe one: Sometimes extralinguistic knowledge would be required to disambiguate a certain ambiguity. Another shortcoming has to do with mistaken lexical analysis: We have problems when a text contains words that are unknown to the lexicon or that are analyzed wrongly by our compound analyzer, or if they even contain a wrong language (common in citations, loanwords etc.).

Before we go into these mistakes, however, let us give an example which shows that in spite of the errors, the overall impression is that the tagger actually does a good job. In the following example, we have asked the corpus to give us all occurrences of the word *stemme* ('vote') used as a verb; we therefore do not want any occurrences of the same wordform used as a noun meaning 'vote' or 'voice'. And indeed, the overall impression is that we get what we wanted:

Proceedings of NODALIDA 1999, pages 66-75

(1) Example of an arbitrary selection of hits from a search for the verb *stemme* 'vote', as opposed to the noun *stemme* 'vote' or 'voice'

Søkestreng: [word="stemme" & tagg=".* verb.*" & (src="AV.*" | src="SA.*" |
src="SK.*")] med 30 tegn på
venstre side og 40 tegn på høyre side.

AV/Ad96/01: re hele det norske folk ved å *stemme* mot, eller utfordre NATO-kolleger ved
AV/Ad96/01: utfordre NATO-kolleger ved å *stemme* for. I slike saker gjelder diplomatiet
AV/Ad96/01: re hele det norske folk ved å *stemme* mot, eller utfordre NATO-kolleger ved
AV/Ad96/01: utfordre NATO-kolleger ved å *stemme* for. I slike saker gjelder diplomatiet
AV/Ad96/01: møter opp, men avstår fra å *stemme*, vil i praksis støtte eventuelle fusjo
AV/Ad96/01: og Mosvold Farsund Invest vil *stemme* for fusjonen. - Gjensidige sier ja til
AV/Ad96/01: n med Danmark og Island fra å *stemme* da FNs nedrustningskomite i går
AV/Ad96/01: var to var i tvil, én ville *stemme* nei, men bademester Eivind Nilsen (ek
AV/Ad96/01: re, sa begge at de kom til å *stemme* for svensk medlemsskap i EU. Eivind
AV/Ad96/01: e, mens ingen tidligere fikk *stemme* for mer enn 12,5 prosent. I forbindels
AV/Ad96/01: egjering som et mottiltak vil *stemme* nei til alle reformer av EU-samarbeidet
AV/Ad96/01: ens 38 land avholdt seg fra å *stemme*. Resolusjonen er ikke bindende, men
AV/Ad96/01: neringen, men fikk det til å *stemme* rimelig bra mot TPG, sa Håvard. - Vi
AV/Ad96/01: var to var i tvil, én ville *stemme* nei, men bademester Eivind Nilsen (ek
AV/Ad96/01: re, sa begge at de kom til å *stemme* for svensk medlemsskap i EU. Eivind
AV/Ad96/01: spennende hvordan dette ville *stemme* med søkernes ønsker. Vi så for oss
AV/Ad96/01: sjonelle ferdighetene » til å *stemme* igjen. Han har 14 dager på seg før den
AV/Ad96/01: tradisjonelt hatt for vane å *stemme* sammen med Arbeiderpartiet og SV.
AV/Ad96/01: ndre tusen nye velger som vil *stemme* på meg! Sjakk i UKE-Adressa | En
AV/Ad96/01: dene hadde forsiktig begynt å *stemme* sine bakben (som de gnir mot en
AV/Ad96/01: den britiske regjering med å *stemme* nei i EUs ministerråd i alle spørsmål,
AV/Af94/01: ene søkeren at hun kom til å *stemme* på ham dersom hun deltok på møtet,
AV/Af94/01: peacemedlemmer fullmakt til å *stemme* på landets vegne. Selv var de ikke til
AV/Af94/01: pålegge sine representanter å *stemme* efter en vedtatt partilinje. Men konse
AV/Af94/01: nde. Hvordan skulle da Høyre *stemme* i Stortinget? Partilandsmøtet løste sa
AV/Af94/01: ervere seg dersom de ikke kan *stemme* for i Stortinget, sier hun. Efter man
AV/Af94/01: t partiet Rukh sier at de vil *stemme* imot avtalen når den kommer opp i parla
AV/Af94/01: år. På Lillehammer skal alt *stemme* i første forsøk, legger han til. - Ko
AV/Af94/01: en i fjor, sier nå at de vil *stemme* Høyre igjen. Høyrelederens opptreden s
AV/Af94/01: ripe inn i tidens politikk og *stemme* med Venstre i unionsstriden også når ha
AV/Af94/01: r frem til velgerne går for å *stemme*. Og kommer velgere som vil ha en

2. Structural ambiguity

Each time a word is left ambiguous between two categories, the corpus user will tend to think that the tagger is unsatisfactory. However, there is a lot of structural ambiguity in language. Most of it goes unnoticed, because our pragmatic and world knowledge guides us towards the right interpretation. But a tagger has only access to form, i.e. morphology and syntax, and will not be able to know which interpretation is the correct one when the formal features are the same.

Let us look at some examples.

(2)

Norwegian, BM: Jeg kjente meg glad og *lettet* da hun gikk
Two readings: a. I felt happy and *relieved* when she left
b. I felt happy and *took off in the air* when she left

Ambiguity: *lettet*: *"lette" 'take off in the air' verb pret*
"lette" 'relieved' adj masc fem ind sg

Our world knowledge tells us that the pronoun *jeg* refers to a person, and we also know that people do not have wings, and therefore normally will stay on the ground, unless something in the context tells us otherwise. We also know that the feeling of being *glad* ('happy') often goes together with the feeling of being *lettet* ('relieved'). As human beings, we therefore interpret the sentence in (2) in the only pragmatically correct way; the a-reading. But the tagger has no world knowledge, and must leave the sentence ambiguous, i.e. leave the word *lettet* with both tags.

(3)

Norwegian, NN: Vaskehjelper som vaskar *skular...*
Two readings: a. Cleaning women/men who wash *schools...*
b. Cleaning women/men who wash *stare...*

Ambiguity: *skular:* *"skule" 'schools' noun plural indef*
"skule" 'stare verb present tense

When seeing a sentence like (3), we know immediately that the relative clause would be a tautology if it only contained the verb without its object. We therefore understand the last word as the object of the verb rather than as a verb. But the tagger is in no position to decide which of the meanings would be meaningless, and has to leave the word *skular* with both tags.[1]

(4)

Norwegian, NN: Ho skulle sleppa fara på åker og eng, *berre* ho ville sjå til huset
Two readings: a. She would not have to travel in fields and meadows, *if only* she would look after the house
b. She would not have to travel in fields and meadows, she was the only one to look after the house

Ambiguity: *berre: "berre" 'if only' subjunction*
"berre" 'only' adverb

In (4), we understand that the most likely interpretation is that the second clause is a condition for the first clause. But the tagger finds the second reading, in which the second clause is a juxtaposed main clause, just as likely. Therefore, the word *berre* must be left ambiguous, keeping both the subjunction and the adverb tags.

(5)

Norwegian, NN: Ho kysste han gang på gang før ho og vart riven bort
Two readings: a. She kissed him time and time again before her and was taken away
b. She kissed him time and time again before she too was taken away

Ambiguity:	*før:*	*"før" 'before' preposition*
		"før" 'until' subjunction
	og:	*"og" 'and' conjunction*
		"og" 'too' adverb
	ho:	*"ho" 'she' pronoun nominative*
		"ho" 'her' pronoun accusative

In (5), both readings are equally likely without knowing more about the context. In the a-reading, there are two women involved, where one kissed the male before the other one, and was subsequently taken away. In the b-reading, there is only one woman, who kissed the man until she - in addition to somebody else - was taken away. There is no way the tagger would be able to choose betwen these readings, and three words have to be left ambiguous as a result.

(6)

Norwegian, BM: Smidsrød har *arbeidet som forsker* ved NTNFs Norsk Institutt for Tang- og tareforskning fra 1961
Three readings: a. Smidsrød has worked as researcher at NTNF...
b. ??Smidsrød has (his) work which does research...
c. ?? Smidsrød has (his) work as researcher...

Ambiguity:	*arbeidet:*	*"arbeide" 'work' verb past participle*
		"arbeid" 'work' noun sg def
	som:	*"som" 'as' preposition*
		"som" 'which' relative subjunction
	forsker:	*"forsker" 'researcher' noun sg ind*
		"forsker" 'research' verb pres

In (6), the meanings of the italicized words tell us that the word *arbeidet* should be interpreted as a verb. E.g., we know that *arbeidet* can never be an agentive noun, and therefore never be the subject of a verb *forsker*. We also know that the italicized words should not be interpreted as a noun phrase, as would have been the case in e.g. *Smidsrød har arbeidet som hobby*. Again, the tagger cannot choose, and will have to leave three words ambiguous.

3. Headlines have too little grammatical information

Headlines and titles generally are very rudimentary sentences that often lack a verb and function words. There is therefore very little information that can guide the tagger in the right direction when it comes to choosing between different readings:

(7)

Norwegian, BM: Rushfeldt *for dyr* for Viking?
Two readings: a. Rushfeldt *too expensive* for Viking
b. Rushfeldt *for animals* for Viking

Ambiguity:	*for:*	*"for" 'for' preposition*
		"for" 'too' adverb

dyr: *"dyr" 'animals' noun ind plural*
"dyr" 'expensive' adjective ind sg masc

As Norwegians, we know that Rushfeldt is a footballplayer, that Viking is a football club, and that football players often require a lot of money to change clubs. The a-reading is the only appropriate one. But the tagger does not know that the other reading is impossible, in which Rushfeldt would make a statement in favour of animals. Two tags are left ambiguous as a result.

(8)

Norwegian, BM: Luftpistol - Internasjonal *gren* og olympisk øvelse
Two readings: a. Air pistol - International branch and olympic event
b. *Air pistol - International cried and olympic event

Ambiguity: *gren:* *"grine" 'cry' verb preterite*
"gren" 'branch' noun ind singular masc

We know that, although an adjective can be the subject of a clause in Norwegian, in this particular sentence, which has to to with Olympic events, it is obvious that the word *gren* refers to air pistols a a branch, and is not a verb. But since this is a headline, there is no requirement for a finite verb, and indeed it does not have one, which might otherwise have helped disambiguate this word. And with no world knowledge, the tagger cannot choose between the two readings.

4. Wrong language or dialect causes problems

In an open text corpus there will always be examples of words and phrases that belong to other languages and dialects. We have not wanted to clean the corpus of this type of occurrences. Obviously, then, words from other languages will not be correctly analyzed by our monolingual tagger. This may in turn create problems for the tagging of the other words surrounding the unknown word - since disambiguation to a large extent depends on the local context of each word. Below are some examples of foreign elements:

(9) Dialect: Uknown word causes unresolved ambiguity in preceding word:

Norwegian, BM: Det *va* et godt forslag. Deinn første kjæresten m...
(*va* instead of *var*)
Two readings: •It was a good proposal
•The was a good proposal

Ambiguity: *det:* *"det" 'the' determiner demonstrative sg neuter*
"det" 'it' pronoun sg neuter

In (9), the italicized word *va* is a Trøndelag dialect word for the standard word *var* (preterite of the verb *være* 'be'). Since *va* is the infinitive form of a verb meaning to walk in water, the tagger finds no finite verb. It will then not know that the first word is a subject, and will not be able to understand that it is the pronoun reading, and not the determiner one, that should be chosen for *Det*. It is left ambiguous.

Other languages and dialects are actually quite common in texts generally - here are some more examples:

(10) German word:

Det er *«schönt»*, for han var mitt forbilde, sier Anders.
It is *wonderful* for he was my idol, Anders says.

(11) Trondheim dialect word:

Ni straffekast til Old Girls og bare tre til *ungpian* startet kampen.
Nine penalties for the Old Girls and only three to the young lasses started the match.

(12) Nynorsk Norwegian sentence in a Bokmål text:

Zimmer-utvalet har kome med framlegg til ny lov her til lands om eigedomsskatt.
The Zimmer committee has suggested a new law in this country about property tax

5. Words written in a way that is believed to be right, but isn't

Orthography is not easy, and indeed lots of people are unaware of how to write or even how to inflect certain words according to the norm. The tagger uses lexicons that follow the standard norm (Bokmålsordboka, Nynorskordboka, IBM's lexical database). Although we have made an effort to enlarge our lexical database to include the most common misconceptions (see Hagen, Johannessen and Kristoffersen 1997), it is not possible to foresee all possible mistakes, as can be seen below.

(13) A phrase believed to be a compound

Norwegian, NN: Det er ikkje nokon *kvensomhelst* som no står fram som ja-mann
Wrong analysis: • *It is not any *who-who-rather* who now stands forward as a yes-man
(should have been *anybody*)

Wrong analysis: compound adverb	*kvensomhelst:*	*"kven-som-helst" 'who-who-rather'*
Should have been:	*kven som helst*	*"kven som helst" 'anybody'*

The way Norwegian creates 'free choice items', like the English *any*, is by adding the (untranslatable) phrase *som helst* to the word in question. Since this is a set phrase, it may easily be conceived of as being compounded with the word it modifies. This has happened in this particular context. Since the word is not in the lexicon, the compound analyzer belonging to the tagger has, correctly, found the three words it consists of, but has treated it like all other compounds, giving the compound as a whole the tag of its last member. This of course gives the wrong result: The compound is given the tag adverb rather than pronoun, or even noun in this particular context. (Indeed, this particular word probably ought to have been added to the lexicon as a nominal compound.)

(14) Wrong inflection

Norwegian, NN: Ho er fødd i Kristiansand og *vaks* opp på Gjøvik
Wrong analysis: She was born in K. and *of the fish being on the feed* up at Gjøvik.
(should have been: *voks*)

Wrong analysis: *vaks: "vak" 'fish being on the feed near the surface' noun sg gen*
Should have been: *voks: "vokse" 'grow' verb preterite*

Strong verbs - verbs that inflect with an internal vowel change - sometimes have an inflectional norm that does not comply with what people actually believe to be the case. In (14) the verb is therefore analyzed by the tagger as a noun, which of course will prevent the correct analysis and disambiguation of the rest of the sentence.

These mistaken beliefs are actually very common. Actually, the Oslo Corpus contains 96 occurrences of the wordform *vaks* (supposed to be preterite of the verb meaning "grow"), compared to 145 correct ones. In other words, 40 per cent of the occurrences are written in a nonstandard way. The same is true in the results from a general Alta Vista web search: 173 pages contain the form "vaks " and 840 the form "voks ". Given that many of the latter ones also must have belonged to the homonymous noun meaning "wax", we can conclude that this mistaken belief with regard to spelling is very common.[2]

6. Spelling errors and mistakes generally cause problems

Every time there is a mis-spelt word or mistakes in punctuation, the tagger will have problems. A mis-spelt word will either not be analyzed or be analyzed wrongly, with the result that other words surrounding that word will also be difficult to analyze. For example, if a noun is wrongly identified as a verb, then the determiner of that noun will not be analyzed correctly, since a determiner needs a noun to be identified. If there is a mistake in punctuation, the tagger will not know where the clause ends. This has serious consequences. Since the tagger, for example, accepts only one finite verb for each clause, a missing full stop will make it impossible to identify two finite verbs in what is really two clauses.

(15) Lack of full stop:

Norwegian, BM: Du kan også svare på fax : 72501468 eller via e-mail : ole-einar.andersen@adresseavisen.no *Vi* må ha svaret innen kl. 12.00.

One wrong reading: You can also answer by fax:... or by e-mail:... Vi must have the answer by 12 Midday.

Wrong analysis: *Vi: "Vi" proper name*
Should have been: *Vi: "vi" 'we' pronoun pl*

In (15), the word *Vi* is of course a pronoun that is written with a capital letter because it is the first word of a sentence. Our knowledge of language makes it immediately possible to interpret it correctly, and to spot that there is a missing full stop in front of this word. However, the tagger has more limited knowledge, and instead analyzes this unknown word as a proper name, wrongly of course, with bad results for further identification of the words in the clause.

Below are a couple of more examples of printing and spelling errors that are problematic:

(16) Two words written together:

I Trondheim er *mellom30* og 50 stellebord av denne typen solgt.
In Trondheim, *between30* and 50 changing units of this kind have been sold.

(17) Wrong spelling:

Hvis betingelsene for *forskninng* er bedre i andre land enn her hjemme, vil forskningen etter hvert flyttes ut. (should have been: *forskning*)

If the conditions for research are better in other countries than here at home, research will be moved out after a while.

7. Wrong for other reasons

There are cases in which the tagger would have had better results had we improved it in certain ways. Below are a couple of such examples.

(18) A word unknown to the lexicon

Norwegian, BM: ... i Europa. *Per-Åke* Palmquist som alle de andre...
Two readings: • *...in Europe. *Per-go* Palmquist like all the others...
• ...in Europe. *Per-Åke* Palmquist like all the others...

Ambiguity: *Per-Åke: proper name*
Per-Åke: "per-åke" 'per-go' verb infinitive

If a word with a capital letter follows a full stop, it is possible to analyze it as a proper name if the word is not in the lexicon. But if the word is ambiguous between a proper name and a word in the lexicon, or is a possible compound, it is more difficult for the tagger to make the right choice. In (18), the name is interpreted as a compound, since the last part of it could be a verb. A list of names or a statistical module telling the tagger that the verb *åke* is very rare might have solved this problem, but as it stands, without these, the problem remains.

(19)

Norwegian, NN: Han kjende berre *noko voks* oppunder ermstaupet på han.
Only reading: a. He only felt *some wax* under his armpit.
Not analyzed: b. He only felt something grew under his armpit

Ambiguity: noko: "noko" 'something' pronoun neuter sg
"noko" 'some' determiner neuter sg
voks: "vokse" 'grow" verb preterite
"voks" 'wax' noun sg ind masc

In (19), the problem for the tagger is that it has to understand that the italicized words, in addition to being a noun phrase consisting of a determiner plus a noun, can also be analyzed as a pronoun followed by a relative clause without a relative subjunction. But this second reading is overall less likely, and to accept this kind of reading would probably give many more ambiguities in the rest of the tagging.

8. Conclusion

As long as each text is not cleaned before tagging, some problems are bound to remain unsolved. We have chosen this inclusive perspective for The Oslo Corpus because we believe that our users appreciate the possibility of being able to do searches in a large corpus. If we had chosen a restrictive attitude to the way the corpus texts should look before they were taggable, our corpus would have been considerably smaller, because we would have had to proofread it. The tagging mistakes which are due to wrong spelling and to wrong language and dialect are therefore impossible to prevent.

Some of the mistakes are due to people's mistaken beliefs. This kind of mistake, which is finite in number, can be accommodated by expanding the lexicon to include nonstandard spellings and inflections of words. We have already done this to some extent, and we have also done the opposite - reduced the lexicon by removing some extremely infrequent correct wordforms that are homonymous with some very frequent ones.

Structural ambiguity and ambiguity due to headlines are two problems that we do not see that we can solve. They require world knowledge of a kind that is hard to include in even very domain specific AI systems, and are impossible to include in tagging of completely open text corpora.

The fact that there turns out to be a surprising amount of structural ambiguity, however, is interesting with respect to the evaluation of taggers more generally. There are basically two types of taggers: those that leave ambiguity where it cannot be decided, like the Constraint Grammar type (Karlsson et al 1995) that we have used for the Oslo Corpus, and those that always make a choice, like statistical taggers (e.g. Kupiec 1992). It is possible that languages differ with respect to how much structural ambiguity they allow. We believe, after having worked with tagging of Norwegian, that a tagger which allows structural ambiguity to remain unsolved is preferable to one that does not.

Notes

1. One might ask whether the rest of this sentence would make the exerpt unambiguous, but this is not the case:

(i) Reinhaldsarbeidarar eller vaskehjelper som vaskar *skular*, kommunehus, bibliotek og andre kommunale hus bør få nøye opplæring i korleis dei skal utføre arbeidet sitt.

'Cleaning personell or washing people who wash [a. schools/b. stare], city halls, libraries or other buildings belonging to the council ought to be taught how to perform their work properly.'

With the a-interpretation, the noun *skular* is in a multiple coordination with other kinds of buildings, all being part of the object of *vaskar*. With the b-interpretation, the verb *skular* ends the clause, while the other kinds of buildings are the (pragmatically odd) subject of a second, asyndetically coordinated, main clause.

[2] We are grateful to Øystein Alexander Vangsnes for making us aware of these facts.

References

Bokmålsordboka. 1993. Landrø, M.I and B.Wangensteen (ed.). Universitetsforlaget, Oslo.

Hagen, K., J.B. Johannessen and K.E. Kristoffersen. 1997. Problemer ved bruk av andres lister til taggerformål. Paper presented at Møter om norsk språk 7, University of Trondheim.

Karlsson, F., A. Voutilainen, J. Heikkilä and A. Anttila. 1995. *Constraint Grammar.* A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin.

Kupiec, J. 1992. Robust part-of-speech tagging using a hidden Markov model, *Computer Speech and Language* 6, 225-242.

Nynorskordboka. 1998. Hovdenak, M., L. Killingbergtrø, A. Lauvhjell, S. Nordlie, M. Rommetveit and D. Worren (red.), Samlaget, Oslo.

The Oslo Corpus of Tagged Norwegian Texts:
http://www.tekstlab.uio.no/norsk/bokmaal/
http://www.tekstlab.uio.no/norsk/nynorsk/