POS Taggers: A comparison

Ricardo Navares

Abstract

Two different *POS* taggers will be compared on an already tagged text from the *Corpus Brown*. The text will be presented in tables indicating the different tags per word produced by each tagger along with a brief comment about potential conflicts.

Introduction

A *Part-of-Speech* (POS) tagger is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., also referred as tags. These tags are used to categorize words in order to apply statistical analysis for pattern recognition.

An statitistical *POS* tagger is a rule-based or stochastic algorithm which uses probability or information theory to assign tags. A couple of algorithms proposed by Stanford department of computational linguistics [1] were chosen to tag a sample text from the Brown Corpus[2].

Text selection

It was intended to select a text which contains several conflicts such as word with several potential tags. For instance, the word *fears* can be tagged as a noun in plural or the third person singular verb as it appears in the *ca03* text from the Brown Corpus:

[1] `http://www-nlp.stanford.edu/links/statnlp.html`
[2] `http://www.nltk.org/nltk_data/`

August 29, 2017

```
Several defendants in the Summerdale police burglary trial made statements indicating
their guilt at the time of their arrest, Judge James B. Parsons was told in Criminal
court yesterday.

The disclosure by Charles Bellows, chief defense counsel, startled observers and was
viewed as the prelude to a quarrel between the six attorneys representing the eight
former policemen now on trial.

Bellows made the disclosure when he asked Judge Parsons to grant his client,
Alan Clements, 30, a separate trial. Bellows made the request while the all-woman jury
was out of the courtroom.

Fears prejudicial aspects ‘‘The statements may be highly prejudicial to my client’’,
Bellows told the court. ‘‘Some of the defendants strongly indicated they knew they
were receiving stolen property. It is impossible to get a fair trial when some of the
defendants made statements involving themselves and others’’.
```

the original tagged text is as follows:

```
Several/ap defendants/nns in/in the/at Summerdale/np police/nn
burglary/nn trial/nn made/vbd statements/nns indicating/vbg
their/pp$ guilt/nn at/in the/at time/nn of/in their/pp$ arrest/nn
,/, Judge/nn-tl James/np B./np Parsons/np was/bedz told/vbn in/in
Criminal/jj-tl court/nn yesterday/nr ./.

The/at disclosure/nn by/in Charles/np Bellows/np ,/, chief/jjs
defense/nn counsel/nn ,/, startled/vbd observers/nns and/cc was/bedz
viewed/vbn as/cs the/at prelude/nn to/in a/at quarrel/nn between/in
the/at six/cd attorneys/nns representing/vbg the/at eight/cd former/ap
policemen/nns now/rb on/in trial/nn ./.

Bellows/np made/vbd the/at disclosure/nn when/wrb he/pps asked/vbd
Judge/nn-tl Parsons/np to/to grant/vb his/pp$ client/nn ,/,
Alan/np Clements/np ,/, 30/cd ,/, a/at separate/jj trial/nn ./.
Bellows/np made/vbd the/at request/nn while/cs the/at all-woman/jj
jury/nn was/bedz out/in of/in the/at courtroom/nn ./.

Fears/vbz-hl prejudicial/jj-hl aspects/nns-hl
‘‘/‘‘ The/at statements/nns may/md be/be highly/ql prejudicial/jj
to/in my/pp$ client/nn ’’/’’ ,/, Bellows/np told/vbd the/at court/nn ./.
‘‘/‘‘ Some/dti of/in the/at defendants/nns strongly/rb indicated/vbd
they/ppss knew/vbd they/ppss were/bed receiving/vbg stolen/vbn property/nn ./.
It/pps is/bez impossible/jj to/to get/vb a/at fair/jj trial/nn when/wrb
some/dti of/in the/at defendants/nns made/vbd statements/nns involving/vbg
themselves/ppls and/cc others/nns ’’/’’ ./.
```

POS taggers

A rule-based and an stochastic tagger were selected based on their availability to be invoked by the wrapper developed in *Perl* which it is attached along with this document. The wrapper reads the already tagged text and

converts it into a plain text to subsequently tag it again via the proposed algorithm. Thus, it can compared the original tagging with the resulting from both algorithms. The wrapper can be tested on any other text belonging to the same Corpus

Tree Tagger.

Developed by Helmut Schmid [1], it is based on binary decision trees which uses *Penn* tags, consequently the algorith was trained on the *Penn TreeBank* Corpus.

As opposed to the N-grams, the algoriths estimates the transition probabilities through binary trees [2]. The probability of a trigram is defined by the branch, so given a trigram $\{A, B, C\}$, being A the child and C the parent node and using bayesian probability, is given by,

$$P(A) = P(C) \cdot P(B|C) \cdot P(A|B, C) \tag{1}$$

The tree is recursively built over a set of trigrams obtained in the training text, this set is recursively split in two subsets based on the probability distribution of the tag and the exploration of the trigram parent nodes to obtain information. Thus the probability of a tag is estimated based on its entropy.

Once the tree is built, it is revisited to check the information gain in order to prune the tree.

Lingua-N-Tagger.

Developed by Aaron Coburn y Maciej Ceglowski [3] is based on Hidden Markov Models on bigrams. A Markov Model is defined by a set of states and their transition probabilities. This model is suitable when there are observed (words) and hidden elements (tags) [4].

The algorithm assigns the sequence of tags which are more likely based on the output. Since in the Markov chain, an state depends only on the previous state, the probability is propagated and the scope is to maximize this probability. It also uses *Penn* tags and was trained on the *Penn TreeBank* Corpus.

Libraries needed for the wrapper

```
#!/usr/bin/perl
use strict;
use warnings;
use Path::Class;
use Lingua::EN::Tagger;
use autodie;
use Text::Table;
use 5.012;
```

Results

Results can be found in Tables 1, 2, 3 and 4. Five columns were defined per word where the first one is the word in the text, the second the original tag of the words, the next two the tag produced by each algorithm and the fifth a reference to a comment if applies. An empty fifth column cell means there was no inconsistency.

Each comment has a crossreference to its correspondent table. Thus *Comment 2.3* corresponds to the third comment of table 2. Each conflict is highlighted in red.

Comments Table 1

Comment 1.1 No conflict. Although is is tagged as a determiner (AP) in the Brown Corpus, both taggers define it as an adjetive (JJ). Without any further context it can be interpret as a determiner (*some*) or adjetive (*respective*) with reference to some other part of the original text.

Comment 1.2 No conflict according to [5]. Brown Corpus has a different tag for articles (AT) and determiner (AP) while Penn Treebank does not. The diffent tag between TreeTag (DT) and HMM (DET) is due to code implementation [3].

Comment 1.3 No conflict. TreeTagger uses this tag (VVD) to define verbs in past.

Comment 1.4 No conflict. See *Comment 1.3* for verbs in gerund.

Comment 1.5 No conflict. As HMM is implemented in Perl, the possesive pronoun is defined with PRPS instead of PP$ to avoid the use of a reserved symbol $ which defines variables.

Word	Brown	TreeTag	HMM	Com.	Word	Brown	TreeTag	HMM	Com.
Several	ap	JJ	JJ	(1.1)	Judge	nn-tl	NP	NNP	(1.6)
defendants	nns	NNS	NNS		James	np	NP	NNP	
in	in	IN	IN		B.	np	NP	NNP	
the	at	DT	DET	(1.2)	Parsons	np	NP	NNP	
Summerdale	np	NP	NNP		was	bedz	VBD	VBD	(1.7)
police	nn	NN	NN		told	vbn	VVN	VBN	
burglary	nn	NN	NN		in	in	IN	IN	
trial	nn	NN	NN		Criminal	jj-tl	JJ	NNP	(1.8)
made	vbd	VVD	VBD	(1.3)	court	nn	NN	NN	
statements	nns	NNS	NNS		yesterday	nr	NN	NN	(1.9)
indicating	vbg	VVG	VBG	(1.4)	.	.	SENT	PP	
their	pp$	PP$	PRPS	(1.5)	The	at	DT	DET	(1.2)
guilt	nn	NN	NN		disclosure	nn	NN	NN	
at	in	IN	IN		by	in	IN	IN	
the	at	DT	DET	(1.2)	Charles	np	NP	NNP	
time	nn	NN	NN		Bellows	np	NP	NNP	
of	in	IN	IN		,	,	,	PPC	
their	pp$	PP$	PRPS	(1.5)	chief	jjs	JJ	JJ	(1.10)
arrest	nn	NN	NN		defense	nn	NN	NN	
,	,	,	PPC		counsel	nn	NN	NN	

Table 1: Results

Comment 1.6 No conflict. The tag used in the Brown Corpus specifies that the name belongs to a title or status, Peen TreeBank tags it as proper noun.

Comment 1.7 No conflict. Corpus Brown differentiates between the verb *to be* and the rest while Penn TreeBank does not. In any case it is tagged as a verb in past.

Comment 1.8 Conflict. *Criminal* is not a title or status as in Comment 1.6 but a definer so it has to be tagged as an adjetive (JJ). There is a conflict with HMM which tags the wrd as a proper noun (NNP).

Comment 1.9 No conflict. Tree bank does not differentiate between adverbial nouns and others [5].

Comment 1.10 No conflict. Superlative. Beatrice Santorini [5] defines this case as fuzzy parts of text. Only can beconsidered superlatives those words ending in *-st* such as *worst*, *most* and *least*, ignoring those like *top* or *chief* as in this case.

Word	Brown	TreeTag	HMM	Com.	Word	Brown	TreeTag	HMM	Com.
,	,	,	PPC		policemen	nns	NNS	NN	
startled	vbd	VVD	NN	(2.1)	now	rb	RB	RB	
observers	nns	NNS	NNS		on	in	IN	IN	
and	cc	CC	CC		trial	nn	NN	NN	
was	bedz	VBD	VBD		.	.	SENT	PP	
viewed	vbn	VVN	VBN		Bellows	np	NP	NNP	
as	cs	IN	IN	(2.2)	made	vbd	VVD	VBD	
the	at	DT	DET		the	at	DT	DET	
prelude	nn	NN	NN		disclosure	nn	NN	NN	
to	in	TO	TO	(2.3)	when	wrb	WRB	WRB	
a	at	DT	DET		he	pps	PP	PRP	(2.5)
quarrel	nn	NN	VB	(2.4)	asked	vbd	VVD	VBD	
between	in	IN	IN		Judge	nn-tl	NP	NNP	
the	at	DT	DET		Parsons	np	NP	NNP	
six	cd	CD	CD		to	to	TO	TO	
attorneys	nns	NNS	NNS		grant	vb	VV	VB	
representing	vbg	VVG	VBG		his	pp$	PP$	PRPS	
the	at	DT	DET		client	nn	NN	NN	
eight	cd	CD	CD		,	,	,	PPC	
former	ap	JJ	JJ		Alan	np	NP	NNP	

Table 2: Results

Comments Table 2

To avoid duplication of comments, those cases already covered in the previous part will not be commented again.

Comment 2.1 Conflict. Lingua-EN-Tagger (HMM) tags the word as noun when it is a verb. TreeTagger tags it correctly.

Comment 2.2 No conflict. Brown tagging is more specific and allows to differentiate between conjunctions and prepositions while Penn TreeBank use the same tag for both cases (IN).

Comment 2.3 No conflict. According to Beatrice Santorini [5] there is no difference between *to* as an infinitive particle and *to* as a preposition.

Comment 2.4 Conflict. HMM tags *quarrel* as a verb while in this context it is a name. This is a case with two possible tags which depend on the context.

Comment 2.5 No conflict. Brown tagging specifies it as a singular pronoun (PPS) while TreeBanks is not that specific (PP). HMM uses PRP on purpose due to developer decision [3].

Word	Brown	TreeTag	HMM	Com.	Word	Brown	TreeTag	HMM	Com.
Clements	np	NP	NNP		courtroom	nn	NN	NN	
,	,	,	PPC		.	.	SENT	PP	
30	cd	CD	CD		Fears	vbz-hl	NNS	NNS	
,	,	,	PPC		prejudicial	jj-hl	JJ	JJ	
a	at	DT	DET		aspects	nns-hl	NNS	NNS	
separate	jj	JJ	JJ		“	“	“	PPL	
trial	nn	NN	NN		The	at	DT	DET	
.	.	SENT	PP		statements	nns	NNS	NNS	
Bellows	np	NP	NNP		may	md	MD	MD	
made	vbd	VVD	VBD		be	be	VB	VB	
the	at	DT	DET		highly	ql	RB	RB	(3.3)
request	nn	NN	NN		prejudicial	jj	JJ	JJ	
while	cs	IN	IN		to	in	TO	TO	
the	at	DT	DET		my	pp$	PP$	PRPS	
all-woman	jj	JJ	NN	(3.1)	client	nn	NN	NN	
jury	nn	NN	NN		”	”	”	PPR	
was	bedz	VBD	VBD		,	,	,	PPC	
out	in	RP	IN	(3.2)	Bellows	np	NP	NNP	
of	in	IN	IN		told	vbd	VVD	VBD	
the	at	DT	DET		the	at	DT	DET	

Table 3: Results

Comments Table 3

Comment 3.1 Conflict. Adjective compuended by adverb plus noun while HMM tags it as noun.

Comment 3.2 Conflict. Prepositions and particles appear close to verbs and are considered fuzzy parts of text and, consequently, difficult to distinguish. The TreeTagger cannot recognise this word as a preposition in this case.

Comment 3.3 No conflict. Qualifier. Again, Brown tagging provides more significance and defines the word as qualifier (QL) while TreeBank is more general using adverbs (RB).

Comments Table 4

Comment 4.1 No conflict. Tagged as determiner.

Comment 4.2 Conflict. Consider past participle as an adjetive, being correctly tagged as a verb by HMM (VBN) but not as an adjetive by TreeTagger (JJ).

Word	Brown	TreeTag	HMM	Com.	Word	Brown	TreeTag	HMM	Com.
court	nn	NN	NN		impossible	jj	JJ	JJ	
.	.	SENT	PP		to	to	TO	TO	
“	“	“	PPL		get	vb	VV	VB	
Some	dti	DT	DET	(4.1)	a	at	DT	DET	
of	in	IN	IN		fair	jj	JJ	JJ	
the	at	DT	DET		trial	nn	NN	NN	
defendants	nns	NNS	NNS		when	wrb	WRB	WRB	
strongly	rb	RB	RB		some	dti	DT	DET	
indicated	vbd	VVD	VBD		of	in	IN	IN	
they	ppss	PP	PRP		the	at	DT	DET	
knew	vbd	VVD	VBD		defendants	nns	NNS	NNS	
they	ppss	PP	PRP		made	vbd	VVD	VBD	
were	bed	VBD	VBD		statements	nns	NNS	NNS	
receiving	vbg	VVG	VBG		involving	vbg	VVG	VBG	
stolen	vbn	JJ	VBN	(4.2)	themselves	ppls	PP	PRP	
property	nn	NN	NN		and	cc	CC	CC	
.	.	SENT	PP		others	nns	NNS	NNS	
It	pps	PP	PRP		”	”	”	PPR	
is	bez	VBZ	VBZ		.	.	SENT	PP	

Table 4: Results

Conclusions

Table 5 shows the conflicts of each algorithm compared to the original reference tag from Brown Corpus. TreeTagger achieves a 98.73% of right guesses, slightly better thatn the Hidden Markov Model which obtains a 97.46% on this test text including punctuation marks.

Since TreeTagger uses trigrams, it seems that it better captures the context of each word than the bigram-based HMM. This conclusion is supported by the fact that TreeTagger does not generate as frequent conflicts when dealing with word with two possible tags such as adjetives-names (JJ-NN) or verb-names (VBD-NN).

On the other hand, we have seen that Corpus Brown provides more statistical significance due to a richer and more specific set of tags. For instance articles and determiners are differentiated (AT-DT) providing larger analysis capabilities.

Word	Brown	TreeTag	HMM	Table	Comment
Criminal	jj-tl	JJ	NNP	1	(1.8)
startled	vbd	VVD	NN	2	(2.1)
quarrel	nn	NN	VB	2	(2.4)
all-woman	jj	JJ	NN	3	(3.1)
out	in	RP	IN	3	(3.2)
stolen	vbn	JJ	VBN	4	(4.2)
Stats					
Total Conflicts		2	4		
Percentage correct (total 158 words)		98.73%	97.46%		

Table 5: Summary

References

[1] Helmut Schmid. Institute for Computational Linguistics. University of Stuttgart. *TreeTagger - a language independent part-of-speech tagger*. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`.

[2] Helmut Schmid. Int Conf on New Methods in Language Processing. 44–49 *Probabilistic Part-of-Speech Tagging Using Decision Trees*. 1994.

[3] Aaron Coburn, Maciej Ceglowski. CPAN. *Lingua-EN-Tagger*. `http://search.cpan.org/~acoburn/Lingua-EN-Tagger/`.

[4] Michael Collins. University of Columbia. Lecture Notes *Tagging with Hidden Markov Models*. 2011.

[5] Beatrice Santorini. University of Pennsylvania *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. 1990.