A Support Vector Machine Approach to Dutch Part of Speech Tagging

Mannes Poel, Luite Stegeman & Rieks op den Akker

Dept. Computer Science
Human Media Interaction Group
University of Twente

September 2007

Outline

CGN: Corpus Spoken Dutch

- Large corpus: $\approx$ 9 million transcribed words.
- 15 different categories:

Type	Size in words
Face to face conversations	2626172
Interview with teacher Dutch	565433
Phone dialogue (recorded with mini disc)	853371
Business conversations	136461
Political debates, discussions and meetings	360328
Sport comments	208399
Masses and ceremonies	18075
Lectures and discourses	140901

- Morpho-syntactically annotated.

Part of Speech Tags

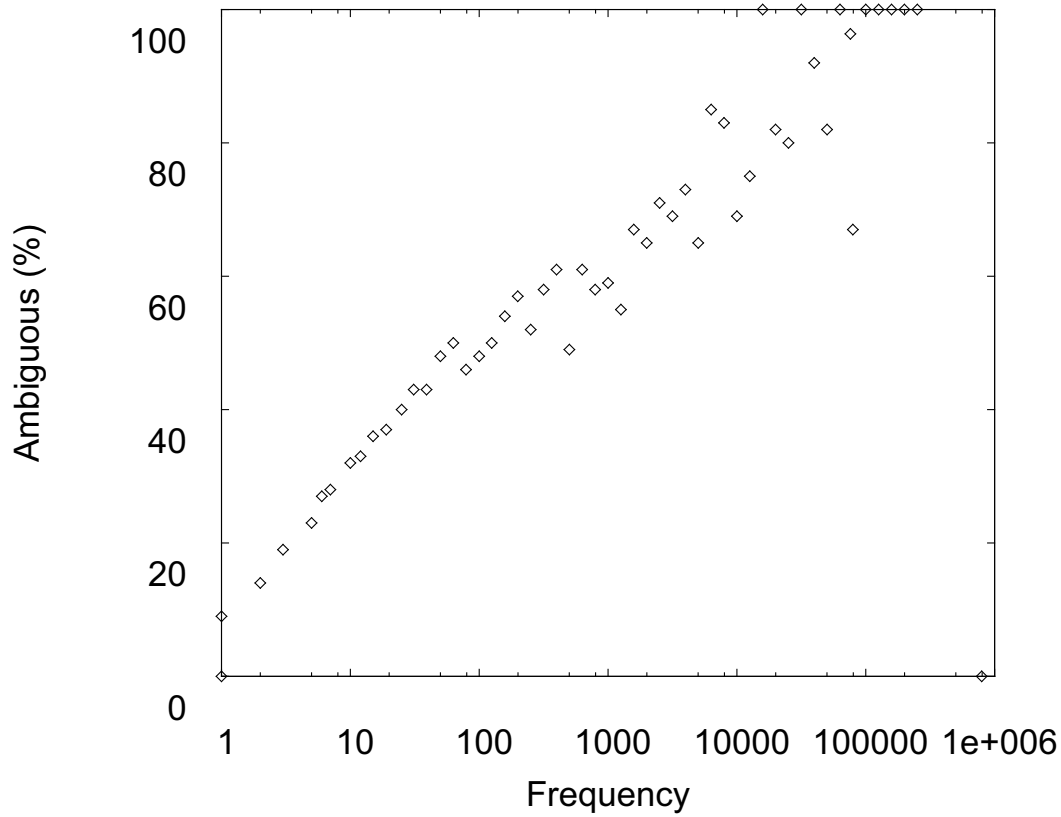
- Full part of speech tags for the CGN consists of 316 different tags. Very fine tuned many tags occur only a few times.
- Main classes: 12 tags such as NOUN, VERB, LET (punctuation mark), SPEC (special). Much smaller then commonly used.
- We considered a tag set consisting of 72 tags.

Tag set

Tag numbers	Part-of-Speech Tag	Tags in the CGN corpus
1 . . . 8	Noun	N1, N2, . . ., N8
9 . . . 21	Verb	WW1, WW2, . . ., WW13
22	Article	LID
23 . . . 49	Pronoun	VNW1, VNW2, . . ., VNW27
50, 51	Conjunction	VG1, VG2
52	Adverb	BW
53	Interjections	TSW
54 . . . 65	Adjective	ADJ1, ADJ2, . . ., ADJ12
66 . . . 68	Preposition	VZ1, VZ2, VZ3
69, 70	Numeral	TW1, TW2
71	Punctuation	LET
72	Special	SPEC

Part-of-Speech Tagging

Part-of-Speech Tagging (PoST) is the process of determining the right tag for (ambiguous) words



Goal & Challenge

The goal is to design a SVM for tagging of the ambiguous words. The tag (class) depends on the context.
Different approaches:

- Giménez and Márquez constructed accurate SVM PoS taggers for English and Spanish. In their approach a linear kernel was used.
- Nakagawa, Kudo and Matusmoto constructed a polynomial kernel SVM PoS tagger for English.

However both of the above approaches are applied to written text only and are applied to a corpus of a much smaller size than the CGN corpus.
The main challenge is to construct a SVM PoS tagger based on the large CGN corpus.

Design of the SVM tagger

- Single pass left to right tagger, using a sliding window of 7 words $w_1 w_2 w_3 w_4 w_5 w_6 w_7$ where w_4 is the word to be tagged.
- Input coding

Type	Used for	Coding
PoST	w1, w2, w3	"1-out-of-N"
relative tag frequencies	w4, w5, w6, w7	vector of relative tag frequencies
word	w1, . . ., w7	"1-out-of-N"
capitalization	w1, . . ., w7	0 for no capitals, 1 for the first letter, 2 for more then one letter
length	w1, . . ., w7	single number
number	w1, . . ., w7	0 if word does not contain number, 1 if it contains at least one number, 2 if the first character is a number, 3 if all characters are numbers
suffix	w4	"1-out-of-N"
PoST bigrams	w1, w2, w3	"1-out-of-N"
PoST trigrams	w1, w2, w3	"1-out-of-N"
Reduced PoST bigrams	w1, w2, w3	"1-out-of-N"
Reduced PoST trigrams	w1, w2, w3	"1-out-of-N"
word bigrams	w1, w2, w3, w5, w6, w7	"1-out-of-N"

Decomposing the SVM

Reason for decomposition: huge amount of data.

- w_4 common word (frequency > 50 in the training set): SVM for each word, resulting in 795 SVM's. Relative tag frequency w_4 is constant and can be discarded from the input coding.
- w_4 uncommon word: SVM for each reduced w_3 tag. 12 reduced tags, hence 12 different SVM's
- w_4 unknown word: SVM for each reduced w_3 tag. 12 reduced tags, hence 12 different SVM's. Relative frequency of w_4 is set to zero, hence can be discarded.

Initial Committee of SVM's

if *w*4 = common word **then**
  select *SVM-common*(*w*4)
else-if *w*4 = uncommon word **then**
  select *SVM-uncommon*(reduced_tag(*w*3))
else *w*4 is unknown word
  select *SVM-unknown*(reduced_tag(*w*3))

Training and test data

- From the CGN 11 sets were constructed: First sentence of the CGN corpus was put in *set0*, second sentence in *set1*, . . ., eleventh sentence in *set10*, twelfth sentence in *set0* etc.
- *set1* up to and including *set10* are used for training and validation.
- *set0* is used for the ultimate performance test.

Initial evaluation on validation set

Kernel type	common	uncommon	unknown	overall
rbf	97.65	87.42	54.14	97.81
2nd order	97.67	87.14	53.47	97.82
3rd order	97.66	87.64	53.47	97.82
linear	97.65	87.25	52.90	97.81

Average unknown word performance: $\approx 53\%$.
Compound analysis of the unknown word can be used to improve performance rate on unknown words.

Compound analysis

In order to improve the unknown word performance we use so-called compound analysis. Many words in Dutch are compounds, i.e. consisting of two or more words glued together. For instance:

- schoenveter (shoestring)
- fietsband (tire of a bicycle)
- fietsventieldopje

Method used: decompose unknown words into compounds and use the second compound as an indication for the PoST.

Compound analysis [2]

Strict: Both compounds must be in the lexicon. Coverage 38.15%, performance on compounds 83.04% and overall unknown word performance 65.72%

Relaxed: Second part must be in the lexicon. Coverage 64.99%, performance on compounds 72.33% and overall unknown word performance 68.32%

Coverage and performance on validation set.

The Final Committee of SVM's

if *w*4 = common word **then**
  select *SVM-common*(*w*4)
else-if *w*4 = uncommom word **then**
  select *SVM-uncommon*(reduced_tag(*w*3))
else-if compound analysis *w*4 succeeds
  %*w*4 is unknown word
  select *SVM-uncommon*(reduced_tag(*w*3))
else % *w*4 is unknown word
  select *SVM-unknown*(reduced_tag(*w*3))

Overall performance

common	uncommon	unknown	overall (all words)
97.28	**88.40**	**70.00**	**97.52**

Tagging performance (in %) on the test set of the final committee of taggers. The overall performance also includes the non-ambiguous words.

- Memory based PoST of Canisius and van den Bosch: 95.96%
- Neural Network based approach: 97.35% (97.88% on known words and 41.67% on unknown words)

More detailed performance analysis

Best scoring category: Phone dialogue

common	uncommon	unknown	overall (all words)
97.94	**88.06**	**65.97**	**98.15**

Worst scoring category: Masses & Ceremonies

common	uncommon	unknown	overall (all words)
96.28	**75.00**	**62.07**	**96.03**

Conclusions

- Design of a committee of SVM's to tackle PoST for large corpora:

if *w*4 = common word **then**
  select *SVM-common*(*w*4)
else-if *w*4 = uncommom word **then**
  select *SVM-uncommon*(reduced_tag(*w*3))
else-if compound analysis *w*4 succeeds
  %*w*4 is unknown word
  select *SVM-uncommon*(reduced_tag(*w*3))
else % *w*4 is unknown word
  select *SVM-unknown*(reduced_tag(*w*3))

Performance:

common	uncommon	unknown	overall (all words)
97.28	**88.40**	**70.00**	**97.52**

- Compound analysis improves unknown word performance: from 53% to 70%
- Reasonable tagging speed: 1000 words/sec

Future work: Combine SVM and NN based approach