Memory-Based Clause Identification

Erik F. Tjong Kim Sang
CNTS - Language Technology Group
University of Antwerp
Belgium
erikt@uia.ua.ac.be

Approach

We have used the memory-based learning algorithm IB1-IG, a nearest-neighbor classifier.

Tokens have been represented by a set of features from a window of surrounding words, part-of-speech tags and chunk tags.

expected VBN I-VP as IN B-PP the DT B-NP S
as IN B-PP the DT B-NP impact NN I-NP S
the DT B-NP impact NN I-NP of IN B-PP X

All training data is stored and test data is classified by taking the class of the training data item that is closest to them in the feature space.

Feature tuning

	train1	0	1	2	3
1	w	61.77	84.40	83.74	81.08
2	p	30.44	80.40	80.47	76.85
3	c	13.67	76.76	79.05	78.71
4	wp	62.24	87.19	84.45	81.22
5	wc	67.95	87.31	85.74	82.97
6	pc	49.29	86.65	84.92	81.72
7	wpc	68.66	87.92	85.93	83.28
8	1+2+3	38.32	85.24	86.92	85.38
9	4+5+6	68.04	88.83	87.44	84.98
10	7+8+9	68.03	88.75	87.72	85.45
11	w-	54.05	83.70	83.48	81.25
12	c-	14.26	77.70	79.30	78.50
13	wc-	58.47	86.53	85.74	82.77

Results for a 10cv experiment on the training data.

Bracket combination rules

1. Assume that exactly one clause starts at each clause start position.
2. Assume that exactly one clause ends at each clause end position but
3. ignore all clause end positions when currently no clause is open, and
4. ignore all clause ends at non-sentence-final positions which attempt to close a clause started at the first word of the sentence.
5. If clauses are opened but not closed at the end of the sentence then close them at the penultimate word of the sentence.

Results

development	precision	recall	$F_{\beta=1}$	
part 1	92.94%	86.87%	89.80	*
part 2	83.80%	80.44%	82.09	
part 3	76.54%	67.20%	71.57	*

test	precision	recall	$F_{\beta=1}$	
part 1	92.91%	85.08%	88.82	*
part 2	84.72%	79.96%	82.28	
part 3	76.91%	60.61%	67.79	*

* results differ from those mentioned in the proceedings