Introduction to the CoNLL-2001 Shared Task: Clause Identification

Erik Tjong Kim Sang, University of Antwerp
Hervé Déjean, University of Tübingen

Motivation

We want to evaluate different learning algorithms on a natural language processing task.

Clause boundaries are useful information for a syntactic analysis of sentences.

The CoNLL-2001 shared task consists of identifying clauses in text.

Task description

(S Coach them in

    (S handling complaints S)

    (S so that

        (S they can resolve problems immediately S)

    S)

    .

S)

- We are interested in all clauses and do not restrict ourselves to base clauses.
- Type and function information have been disregarded.
- The shared task has been split in three parts to allow basic learning algorithms to participate as well.

Data

- We use sections 15-18 of the Wall Street Journal part of the Penn Treebank-2 as training data, section 20 as development data and section 21 as test data.
- Data files consisted of four columns: words, part-of-speech (POS) tags, chunk tags and clause tags.
- POS tags and chunk tags have been estimated in order to obtain realistic evaluation rates.
- Only phrases with labels starting with S have been included in as clauses (omitting RRC and FRAG).

Data example

word	POS	chunk	O_1	O_2	O_3
Coach	NNP	B-NP	S	X	(S*
them	PRP	B-NP	X	X	*
in	IN	B-PP	X	X	*
handling	NN	O	S	X	(S*
complaints	NNS	O	X	E	*S)
so	RB	B-SBAR	S	X	(S*
that	IN	I-SBAR	X	X	*
they	PRP	B-NP	S	X	(S*
can	MD	B-VP	X	X	*
resolve	VB	I-VP	X	X	*
problems	NNS	B-NP	X	X	*
immediately	RB	B-ADVP	X	E	*S)S)
.	.	O	X	E	*S)

Evaluation

We register the number of completely correct clauses and compute precision, recall and $F_{\beta=1}$ rates:

Precision: number of correct clauses divided by the number of clauses found by the algorithm.

Recall: number of correct clauses divided by the number of clauses in the corpus.

$F_{\beta=1}$: (β^2+1)*precision*recall divided by β^2*precision +recall.

Baseline performances have been obtained with an algorithm which puts every sentence in a single clause.

Evaluation software was available to all participants.

Participants

Six groups have participated in the CoNLL-2001 shared task. They have used connectionist techniques, memory-based methods, statistical techniques, symbolic methods and tree/graph boosting:

- Patrick and Goyal (graph boosting)
- Hammerton (connectionist techniques)
- Déjean (symbolic methods)
- Tjong Kim Sang (memory-based methods)
- Molina and Pla (statistical techniques)
- Carreras and Màrquez (tree boosting)

The authors will present their systems themselves.

Results bracket estimation

test part 1	precision	recall	$F_{\beta=1}$	
Carreras & Màrquez	93.96%	89.59%	91.72	
Tjong Kim Sang	92.91%	85.08%	88.82	*
Molina & Pla	89.54%	86.01%	87.74	*
Déjean	93.76%	81.90%	87.43	
Patrick & Goyal	89.79%	84.88%	87.27	*
baseline	98.44%	36.58%	53.34	

test part 2	precision	recall	$F_{\beta=1}$	
Carreras & Màrquez	90.04%	88.41%	89.22	
Tjong Kim Sang	84.72%	79.96%	82.28	
Patrick & Goyal	80.11%	83.47%	81.76	*
Molina & Pla	79.57%	77.68%	78.61	*
Déjean	99.28%	48.90%	65.47	
baseline	98.44%	48.90%	65.34	

* results differ from those mentioned in the proceedings

Results full task

test part 3	precision	recall	$F_{\beta=1}$	
Carreras & Màrquez	84.82%	73.28%	78.63	
Molina & Pla	70.89%	65.57%	68.12	*
Tjong Kim Sang	76.91%	60.61%	67.79	*
Patrick & Goyal	73.75%	60.00%	66.17	*
Déjean	72.56%	54.55%	62.77	
Hammerton	55.81%	45.99%	50.42	
baseline	98.44%	31.48%	47.71	

* results differ from those mentioned in the proceedings

- Four systems perform approximately equally well.
- Hammerton did not use all training data.
- Carreras & Màrquez perform a lot better than the rest (their error rate is 33% lower than second best).

Comparison AdaBoost - TiMBL

The Carreras and Màrquez approach uses more features than the other approaches. Does this account for the large performance differences with the other systems?

development part 1	precision	recall	$F_{\beta=1}$	
Carreras & Màrquez	95.77%	92.08%	93.89	
C&M with TKS ftrs	94.19%	88.62%	91.32	
TKS with C&M ftrs	93.16%	89.33%	91.20	
Tjong Kim Sang	92.94%	86.87%	89.80	*
baseline	96.32%	38.08%	54.58	

The performance differences between the Carreras and Màrquez approach and the other approaches are both related to the choice of features and the choice of system (Adaboost).

System combination

development part 1	systems used: all	systems used: some
majority voting	92.26	93.89
accuracy voting	92.26	93.89
precision voting	92.26	93.89
precision-recall voting	92.26	93.89
pairwise voting	92.45	93.89
stacked classifier	93.78	93.89
stacked classifier + POS	93.32	94.02
Carreras & Màrquez	93.89	
average	90.43	

- Background info: Van Halteren et al., Coling 1998.
- Apart from a small increase for a stacked classifier with extra information, system combination does not improve the best single result.
- The reason for this is that there is a large difference between the best individual system and the others.

Problematic sentences (1)

(Refcorp was created
(to help fund the thrift bailout) .)

(“ (Improving profitability of U.S. operations)
is an extremely high priority in the company . ”)

(Advancing and declining issues finished
(about even) .)

(“ But (it ’s not mediocre) ,
(it ’s a real problem) . ”)

(Trouble was ,
(nobody thought (they looked right)) .)

((He will also remain a director) ,
(US Facilities said) , but
(won’t serve on any board committees) .)

Problematic sentences (2)

(Then , it rebounded
(to finish down only 18.65 points) .)

(The stock recovered somewhat
to finish 1 1/4 lower at 26 1/4 .)

(The death of CIA Director William Casey and
resignation of Oliver North allowed
(anti-Noriega political forces to gain influence) .)

(Small-business suppliers want
(prisons to stop getting high priority) ,
(especially as
(prison production grows with
swelling inmate populations)) .)

Concluding remarks

- Six systems have participated in the CoNLL-2001 shared task: clause identification.
- The best results have been obtained by Xavier Carreras and Lluís Màrquez from Spain.
- Their excellent results have both been made possible by the choice of the learning algorithm (AdaBoost applied to decision trees) and their choice of features for describing the domain.