Boosting Trees for Clause Splitting

Xavier Carreras & Lluís Màrquez

TALP Research Center
Technical University of Catalonia

6th July 2001

CoNLL 2001 Shared Task

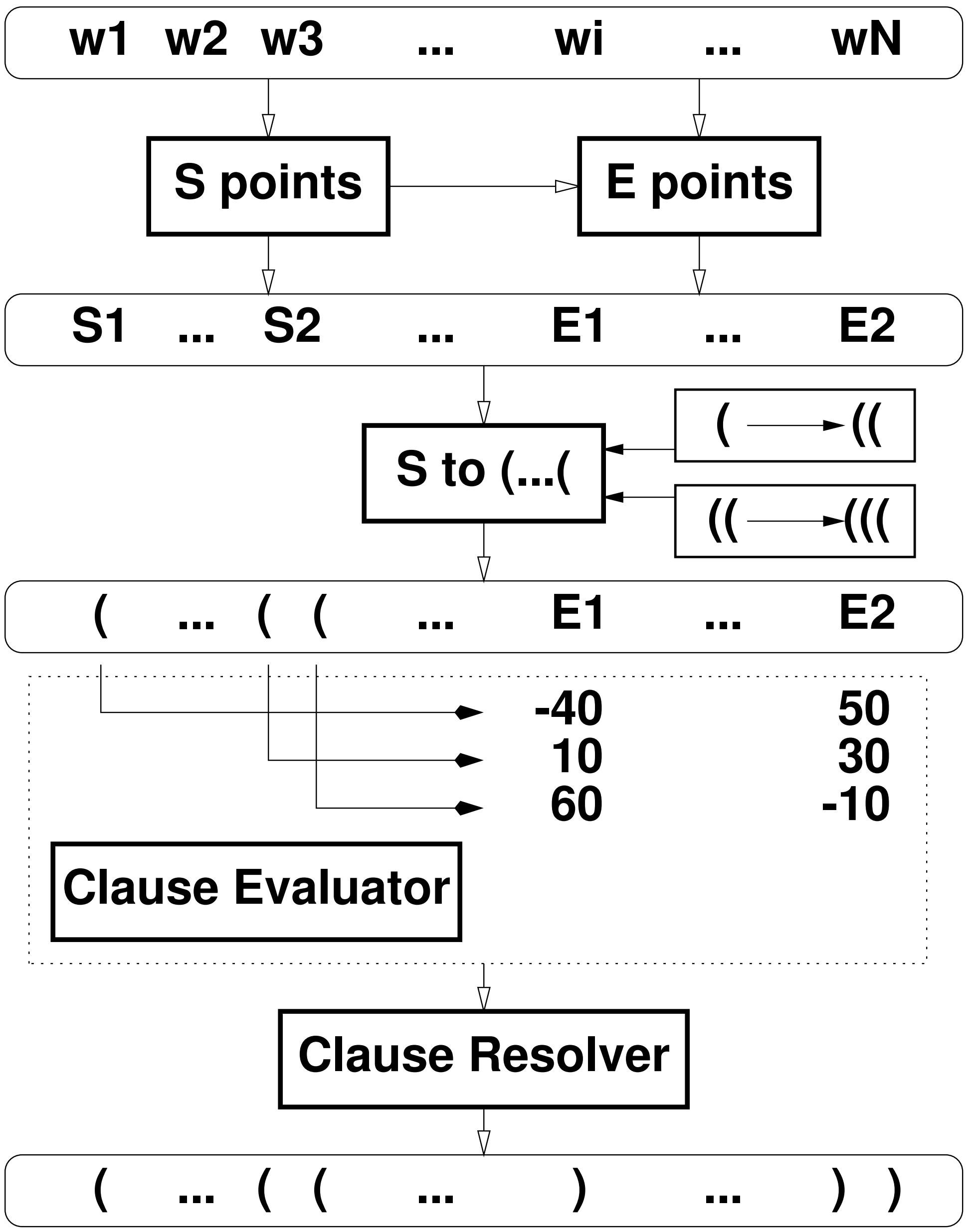
Approach

Decomposition of the whole problem into a combination of "simple" binary decisions.

- Learning algorithm: AdaBoost.MH
  - R. Schapire & Y. Singer, 1999.
  - General algorithm for combining classifiers.
  - Success on several NLP problems:
    - Text Categorization (Schapire & Singer, 00)
    - Word Sense Disambiguation (Escudero et al., 00)
    - PP-attachment (Abney et al., 99)
    - POS-tagging (Abney et al., 99)

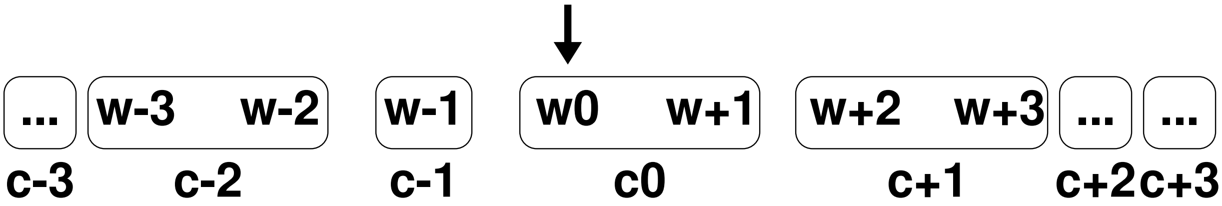
  - Weak Rules: decision trees of arbitrary fixed depth (1-5).
  - Binary classifier.
  - Binary features.
  - Predictions are given with a *confidence value*.

Clause Splitting Scheme



Features

- Window:
  - Word Window: word forms and POS.
  - Chunk Window: chunk labels.



- Whole Sentence:
  - Sentence patterns, representing the relevant structure of the sentence. Ex:

    `The deregulation of railroads` **`and`** `trucking companies` **`that began`** `in 1980` **`enabled`** `shippers to` **`bargain`** `for transportation .`

    Pattern: `and that VP VP VP .`
  - Sentence features:
    - Verbs
    - Relative pronouns
    - Punctuation marks
    - S points and E points

Sizes and Results

	#ex	#ex +	#feat	#wr	d
S points	138,069	22,950	43,356	2,000	3
E points	138,069	17,150	42,853	2,000	3
($\rightarrow$ ((	23,075	1,519	36,240	1,500	3
(($\rightarrow$ (((	1,600	81	5,361	100	3
Cl.Eval.	39,209	16,294	14,040	2,000	3

development	precision	recall	$F_{\beta=1}$
part 1	95.77%	92.08%	93.89%
part 2	91.27%	89.00%	90.12%
part 3	87.18%	82.48%	**84.77%**

test	precision	recall	$F_{\beta=1}$
part 1	93.96%	89.59%	91.72%
part 2	90.04%	88.41%	89.22%
part 3	84.82%	73.28%	**78.63%**

Comments and Conclusions

- Good results due to:
  - Rich feature set.
  - AdaBoost
- But ...
  - A greedy approach is followed.
  - Noise in the input of the final decisions.
- Maybe a recursive decomposition would perform better.