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)
   )
)

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

<table>
<thead>
<tr>
<th>word</th>
<th>POS</th>
<th>chunk</th>
<th>O₁</th>
<th>O₂</th>
<th>O₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coach</td>
<td>NNP</td>
<td>B-NP</td>
<td>S</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>them</td>
<td>PRP</td>
<td>B-NP</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>in</td>
<td>IN</td>
<td>B-PP</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>handling</td>
<td>NN</td>
<td>O</td>
<td>S</td>
<td>X</td>
<td>(S*)</td>
</tr>
<tr>
<td>complaints</td>
<td>NNS</td>
<td>O</td>
<td>X</td>
<td>E</td>
<td>*S)</td>
</tr>
<tr>
<td>so</td>
<td>RB</td>
<td>B-SBAR</td>
<td>S</td>
<td>X</td>
<td>(S*)</td>
</tr>
<tr>
<td>that</td>
<td>IN</td>
<td>I-SBAR</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>they</td>
<td>PRP</td>
<td>B-NP</td>
<td>S</td>
<td>X</td>
<td>(S*)</td>
</tr>
<tr>
<td>can</td>
<td>MD</td>
<td>B-VP</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>resolve</td>
<td>VB</td>
<td>I-VP</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>problems</td>
<td>NNS</td>
<td>B-NP</td>
<td>X</td>
<td>X</td>
<td>*</td>
</tr>
<tr>
<td>immediately</td>
<td>RB</td>
<td>B-ADVP</td>
<td>X</td>
<td>E</td>
<td>*S)</td>
</tr>
</tbody>
</table>

CoNLL-2001

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.

$F_{\beta-1}$: \((\beta^2 + 1) \cdot \text{precision} \cdot \text{recall} \div \beta^2 \cdot \text{precision} + \text{recall}.

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

Evaluation software was available to all participants.

CoNLL-2001

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 Márquez (tree boosting)

The authors will present their systems themselves.

CoNLL-2001

Results bracket estimation

<table>
<thead>
<tr>
<th>test part 1</th>
<th>precision</th>
<th>recall</th>
<th>$F_{\beta-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carreras &amp; Márquez</td>
<td>93.96%</td>
<td>89.59%</td>
<td>91.72</td>
</tr>
<tr>
<td>Tjong Kim Sang</td>
<td>92.91%</td>
<td>85.08%</td>
<td>88.82 *</td>
</tr>
<tr>
<td>Molina &amp; Pla</td>
<td>89.54%</td>
<td>86.01%</td>
<td>87.74 *</td>
</tr>
<tr>
<td>Déjean</td>
<td>93.76%</td>
<td>81.90%</td>
<td>87.43</td>
</tr>
<tr>
<td>Patrick &amp; Goyal</td>
<td>89.79%</td>
<td>84.88%</td>
<td>87.27 *</td>
</tr>
<tr>
<td>baseline</td>
<td>98.44%</td>
<td>95.58%</td>
<td>93.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>test part 2</th>
<th>precision</th>
<th>recall</th>
<th>$F_{\beta-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carreras &amp; Márquez</td>
<td>90.04%</td>
<td>88.41%</td>
<td>89.22</td>
</tr>
<tr>
<td>Tjong Kim Sang</td>
<td>84.72%</td>
<td>79.96%</td>
<td>82.28</td>
</tr>
<tr>
<td>Patrick &amp; Goyal</td>
<td>80.11%</td>
<td>83.47%</td>
<td>81.76 *</td>
</tr>
<tr>
<td>Molina &amp; Pla</td>
<td>79.57%</td>
<td>77.68%</td>
<td>78.61 *</td>
</tr>
<tr>
<td>Déjean</td>
<td>99.28%</td>
<td>48.90%</td>
<td>65.47</td>
</tr>
<tr>
<td>baseline</td>
<td>98.44%</td>
<td>48.90%</td>
<td>65.34</td>
</tr>
</tbody>
</table>

* results differ from those mentioned in the proceedings

CoNLL-2001
Results full task

<table>
<thead>
<tr>
<th>test part 3</th>
<th>precision</th>
<th>recall</th>
<th>$F_{\beta=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carreras &amp; Márquez</td>
<td>84.82%</td>
<td>73.28%</td>
<td>78.63</td>
</tr>
<tr>
<td>Molina &amp; Pla</td>
<td>70.89%</td>
<td>65.57%</td>
<td>68.12 *</td>
</tr>
<tr>
<td>Tjong Kim Sang</td>
<td>76.91%</td>
<td>60.61%</td>
<td>67.79 *</td>
</tr>
<tr>
<td>Patrick &amp; Goyal</td>
<td>73.75%</td>
<td>60.00%</td>
<td>66.17 *</td>
</tr>
<tr>
<td>Déjean</td>
<td>72.56%</td>
<td>54.55%</td>
<td>62.77</td>
</tr>
<tr>
<td>Hammerton</td>
<td>55.81%</td>
<td>45.99%</td>
<td>50.42</td>
</tr>
<tr>
<td>baseline</td>
<td>98.44%</td>
<td>31.48%</td>
<td>47.71</td>
</tr>
</tbody>
</table>

* results differ from those mentioned in the proceedings

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

Comparison AdaBoost - TiMBL

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

<table>
<thead>
<tr>
<th>development part 1</th>
<th>precision</th>
<th>recall</th>
<th>$F_{\beta=1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carreras &amp; Márquez</td>
<td>95.77%</td>
<td>92.08%</td>
<td>93.89</td>
</tr>
<tr>
<td>C&amp;M with TKS ftrs</td>
<td>94.19%</td>
<td>88.62%</td>
<td>91.32</td>
</tr>
<tr>
<td>TKS with C&amp;M ftrs</td>
<td>93.16%</td>
<td>89.33%</td>
<td>91.20</td>
</tr>
<tr>
<td>Tjong Kim Sang</td>
<td>92.94%</td>
<td>86.87%</td>
<td>89.80 *</td>
</tr>
<tr>
<td>baseline</td>
<td>96.32%</td>
<td>38.08%</td>
<td>54.58</td>
</tr>
</tbody>
</table>

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

System combination

<table>
<thead>
<tr>
<th>development part 1</th>
<th>systems used</th>
<th>all</th>
<th>some</th>
</tr>
</thead>
<tbody>
<tr>
<td>majority voting</td>
<td>92.26</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>accuracy voting</td>
<td>92.26</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>precision voting</td>
<td>92.26</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>precision-recall voting</td>
<td>92.26</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>pairwise voting</td>
<td>92.45</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>stacked classifier</td>
<td>93.78</td>
<td>93.89</td>
<td></td>
</tr>
<tr>
<td>stacked classifier + POS</td>
<td>93.32</td>
<td>94.02</td>
<td></td>
</tr>
<tr>
<td>Carreras &amp; Márquez</td>
<td>93.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>average</td>
<td>90.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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 ) ). )

- Six systems have participated in the CoNLL-2001
  shared task: clause identification.

- The best results have been obtained by Xavier
  Carreras and Lluís Márquez 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.