Representing Text Chunks
Tjong Kim Sang & Veenstra 1999

Nikhil Dinesh
Edward Loper

Outline
• Introduction
• Representations
• Experiment 1 & Discussion
• Experiment 2 & Discussion
• Experiment 3 & Discussion
• Conclusions

Introduction
• Goal: Explore the effect of different output representations on performance
• Method:
  – Train chunk parsers using 6 different representations
  – Use 3 related learning algorithms
  – Compare performance
• Conclusion: No significant effect

Representations (Complete)

<table>
<thead>
<tr>
<th>Representation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOB1</td>
<td>The first word inside a baseNP immediately following another gets a B tag</td>
</tr>
<tr>
<td>IOB2</td>
<td>All baseNP-initial words receive a B tag</td>
</tr>
<tr>
<td>IOE1</td>
<td>The final word inside a baseNP immediately preceding another gets an E tag</td>
</tr>
<tr>
<td>IOE2</td>
<td>All baseNP-final words get an E tag</td>
</tr>
</tbody>
</table>
Representations (Partial)

[ ]
All baseNP-initial words get a [ tag and all other words get a '.' tag

]
All baseNP-final words get a ] tag and all other words get a '.' tag

IO
All words inside a baseNP receive an I tag and all other words get an O tag

Experiment 1: Methods

• Algorithm: memory based learner (IB1-IG)
  – Use information gain to define distance metric
  – Use the classification of the nearest neighbor
• Features:
  – Surrounding words and POS tags
  – Same basic features used by R&M 95
• All tagging is independent
  – No dependence on previous predictions
  – No cascaded decisions
  – For combination reps (e.g., "[+IO]"), each tag is assigned independently.

Combinations of Partial Reps

[+] BaseNP = [.....]

[+IO]
[ + I = B. (similar to IOB2)

IO+[ I + ] = E (similar to IOE2)

Over-fitting the data?

• Find the optimal context size for each output representation.
• How do we decide what context to use?
  "The optimal context size will be determined by comparing the results of different context sizes on the training data." (§2.3)
• Training on the test data!
  – For experiment 1: 25 possible contexts
  – For experiment 2: 256 possible contexts
  – For experiment 3: ~16,000 possible contexts
Experiment 1: Results

- No (statistically) significant differences
- "[+IO] and IO+" do slightly better.
- Interesting to think about why that might be
- But remember that these differences are unreliable at best.
- Conclusion: output representation doesn't matter (in this case).

<table>
<thead>
<tr>
<th>Rep</th>
<th>Context</th>
<th>F_{β=1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOB1</td>
<td>L=2/R=1</td>
<td>89.17</td>
</tr>
<tr>
<td>IOB2</td>
<td>L=2/R=1</td>
<td>88.76</td>
</tr>
<tr>
<td>IOE1</td>
<td>L=1/R=2</td>
<td>88.67</td>
</tr>
<tr>
<td>IOE2</td>
<td>L=2/R=2</td>
<td>89.01</td>
</tr>
<tr>
<td>[+ ]</td>
<td>2/1 + 0/2</td>
<td>89.32</td>
</tr>
<tr>
<td>[+IO]</td>
<td>2/0 + 1/1</td>
<td>89.43</td>
</tr>
<tr>
<td>IO+</td>
<td>1/1 + 0/2</td>
<td>89.42</td>
</tr>
</tbody>
</table>

Why might output representation matter?

- When do we expect output representation to matter?
- The output representation provides the algorithm with a way of dividing up the problem.
  - Similar to features?
- It's known that feature choice has a major effect.
- What about the choice of output representation?
  - Representations with the same information content?
  - Representations with different information content?

Generality of Conclusion

- How general is the conclusion that output representation doesn't matter?
  - What about other algorithms?
    - Transformational (e.g., R&M)
    - Maxent
    - SVM
  - What about other domains?
    - POS tagging
    - PP attachment
  - What about other representations?
    - (I, O, B, B')

Information Content

- All 6 representations have the same information content:
  - From the tagging for any representation, we can trivially derive taggings for all other reps.
- Is it just the information content that matters?
- Or is the information packaging also important?
Information Packaging

- Clearly, information packaging matters in some cases.
- Example: consider the representation "+AB":
  - A if inside a base NP and word index is odd
  - A if outside a base NP and word index is even
  - B if outside a base NP and word index is odd
  - B if inside a base NP and word index is even

- Same information content
- Very poor performance for most algorithms

Output Rep & Features

- Performance of an output representation depends on which features are chosen.
  - Example: for "+AB", adding an "even" feature would improve performance considerably.
- Output representation must be a "good fit" with the features.
- TKS&V picked best features for each output rep
  - Try comparing performance of different output reps on a common feature set?

Information Packaging (2)

- Output representation must be "natural" for the learning algorithm and for the data.
- Example: for SVMs, we apply a transformation to make the problem linearly separable.

Follow-up Experiments

- Many interesting follow-up experiments to examine the effect of output representation choice. E.g.:
  - Replicate [TKS&V99] with different algorithms
  - Replicate [TKS&V99] with different domains
  - Replicate [TKS&V99] with different encodings
    - Encodings with the same information content
    - Encodings with different information content
  - Compare performance of different output reps on a common feature set
- (anyone still looking for a final project? 😊)
Experiment 2

- “Cascaded” classifier
- Objective: To find the optimal no. of extra classification tags.

\[ P(c_n | < w_{n-1}, t_{n-1}, t_{n-2}, c_{n-2} > \ldots) \]

instead of

\[ P(c_n | < w_{n-1}, t_{n-1}, t_{n-2} > \ldots) \]

Why Use Cascades?

- The reason for doing it in a cascade is because they wanted to consider right context as well
- Classification tag contexts in the range 0 to 3
- 256 * E1 combinations for complete reps
- 256 * 256 * E1 for partial reps

Experiment 2: Context for IOB1

Experiment 2: Results

<table>
<thead>
<tr>
<th>Word/POS context</th>
<th>Chunk Tag</th>
<th>F score</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOB1 L=2 / R=1</td>
<td>1 / 2</td>
<td>90.12</td>
</tr>
<tr>
<td>IOB2 L=2 / R=1</td>
<td>1 / 0</td>
<td>89.30</td>
</tr>
<tr>
<td>IOE1 L=1 / R=2</td>
<td>1 / 2</td>
<td>89.55</td>
</tr>
<tr>
<td>IOE2 L=1 / R=2</td>
<td>0 / 1</td>
<td>89.73</td>
</tr>
<tr>
<td>[ + ] 2 / 1 + 0 / 2</td>
<td>0 / 0 + 0 / 0</td>
<td>89.32</td>
</tr>
<tr>
<td>[ + IO ] 2 / 0 + 1 / 1</td>
<td>0 / 0 + 1 / 1</td>
<td>89.78</td>
</tr>
<tr>
<td>IO + ] 1 / 1 + 0 / 2</td>
<td>1 / 1 + 0 / 0</td>
<td>89.86</td>
</tr>
</tbody>
</table>
Experiment 2: Results

• What is improvement w.r.t the F measure?
• An increase in recall or precision with not too much of a decrease in the other metric
• But the recall took a big hit for a smaller improvement in the precision (in the IB1-IG paper) and scoring by the F measure won’t consider this an advantage
• There seems to be a relationship between the metric used and the representation performance

Experiment 3

• Add classification of 3, 4 and 5 experiments of the first series in addition to the optimal one to the second cascade
• Objective: To use different context sizes

\[ \lambda_1 P(c_n | <\text{context 1}> ) + \lambda_2 P(c_n | <\text{context 2}> ) \]....

Experiment 3: Context for IOB1

Experiment 3: Contexts

• Combinations of 3, 4 or 5 experiments of the following lists
  • (0/0, 1/1, 2/2, 3/3, 4/4, 5/5) (Equal)
  • (0/1, 1/2, 2/3, 3/4) (Right heavy)
  • (1/0, 2/1, 3/2, 4/3) (Left heavy)
  • (16 + 5 + 5) * E2 combinations
  • 26 * 26 * E2 for partial reps
Experiment 3: Results

<table>
<thead>
<tr>
<th>Word/POS</th>
<th>Chunk</th>
<th>Combinations</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOB1</td>
<td>2/1</td>
<td>1/1</td>
<td>0/0 1/1 2/2 3/3</td>
</tr>
<tr>
<td>IOB2</td>
<td>2/1</td>
<td>1/0</td>
<td>2/1</td>
</tr>
<tr>
<td>IOE1</td>
<td>1/2</td>
<td>1/2</td>
<td>0/0 1/1 2/2 3/3</td>
</tr>
<tr>
<td>IOE2</td>
<td>1/2</td>
<td>0/1</td>
<td>1/2</td>
</tr>
<tr>
<td>[ + ]</td>
<td>2/1 + 0/2</td>
<td>0/0+0/0</td>
<td>- + -</td>
</tr>
<tr>
<td>[ + IO ]</td>
<td>2/0 + 1/1</td>
<td>0/0+1/1</td>
<td>- + 0/1 1/2 2/3 3/4</td>
</tr>
<tr>
<td>[IO + ]</td>
<td>1/1 + 0/2</td>
<td>1/1+0/0</td>
<td>0/1 1/2 2/3 3/4+ -</td>
</tr>
</tbody>
</table>

Experiment 4: Results

<table>
<thead>
<tr>
<th>W/T</th>
<th>C</th>
<th>Combinations</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOB1</td>
<td>3/3(3)</td>
<td>1/1</td>
<td>0/0(1) 1/1(1) 2/2(3) 3/3(3)</td>
</tr>
<tr>
<td>IOB2</td>
<td>3/3(3)</td>
<td>1/0</td>
<td>3/3(3)</td>
</tr>
<tr>
<td>IOE1</td>
<td>2/3(3)</td>
<td>1/2</td>
<td>0/0(1) 1/1(1) 2/2(3) 3/3(3)</td>
</tr>
<tr>
<td>IOE2</td>
<td>2/3(3)</td>
<td>0/1</td>
<td>2/3(3)</td>
</tr>
<tr>
<td>[ + ]</td>
<td>4/3(3) + 4/4(3)</td>
<td>0/0+0/0</td>
<td>- + -</td>
</tr>
<tr>
<td>[ + IO ]</td>
<td>4/3(3) + 3/3(3)</td>
<td>0/0+1/1</td>
<td>- + 0/1(1) 1/2(3) 2/3(3) 3/4(3)</td>
</tr>
<tr>
<td>IO + ]</td>
<td>3/3(3) + 2/3(3)</td>
<td>1/1+0/0</td>
<td>0/1(1) 1/2(3) 2/3(3) 3/4(3) + -</td>
</tr>
</tbody>
</table>

Experiment 4

- K nearest neighbors
- Experiment 1 was repeated with k=3
- Experiment 3 repeated with k=3 wherever it outperformed k=1

Conclusions & Questions

- Outline
  - What makes a good output representation?
  - Output representation & feature selection
  - Questions for Discussion
What Makes a Good Output Rep?

• A good output representation depends on:
  – The learning algorithm
  – The features
  – The data

• Intuition: an output representation is good if it divides data into groups with similar features.
  – “Similarity” depends on the learning algorithm

• Example: chunk parsing
  – For a given algorithm & feature set, which words tend to have similar feature values?
    • Words at the beginning of all base NPs?
    • Words at the beginning of base NPs preceded by base NPs?
    • Etc.

Questions

• Does output representation matter?
  – When does output representation matter?
• What makes a good output rep?
  – What factors do we need to consider?
• Is automatic output rep selection feasible?
• How do features relate to the output rep?
• What effect does the “size” (# of bits) of the output rep have?

Output Rep & Feature Selection

• Can we automatically choose a good output representation from a set of candidates?
  – c.f. feature selection

• Decision ordering:
  – Feature selection first?
  – Output representation selection first?
  – Consider them both at the same time?

Extra Slides
(if there’s time and/or interest)
Information Content

- What is the effect of using output representations that do not have the same information content?
- Example:
  - Task: classify texts as fiction or nonfiction
  - Use a fine grained output rep: scifi, thriller, reference, biography, etc.
- Advantage: output rep is more likely to divide the data into groups with similar features.
- Disadvantage: sparse data, more difficult to create the corpus.

Experiment 2 & HMMs

- We can think of experiment 2 (simple cascading) as an approximation to Viterbi decoding.
- In particular, experiment 1 gives us the most likely individual tags; but experiment 2 tries to give us the most likely tag sequences.
- Advantage of experiment 2: we can use predictions from both directions.
- Disadvantage of experiment 2: it’s less principled, and so it can still give unlikely tag sequences.

Experiment 3 & Backoff

- We can think of experiment 2 (simple cascading) as an approximation to backoff.
- Combines evidence from different context sizes.
- Less principled than backoff?