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)

IOB1	The first word inside a baseNP immediately following another gets a B tag
IOB2	All baseNP-initial words receive a B tag
IOE1	The final word inside a baseNP immediately preceding another gets an E tag
IOE2	All baseNP-final words get an E tag

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

5

Combinations of Partial Reps

[+]	BaseNP = [.....]
[+IO	[+ I = B. (similar to IOB2)
IO+]	I +] = E (similar to IOE2)

6

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.

7

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

8

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).**

Rep	Context	$F_{\beta=1}$
IOB1	L=2/R=1	89.17
IOB2	L=2/R=1	88.76
IOE1	L=1/R=2	88.67
IOE2	L=2/R=2	89.01
[+]	2/1 + 0/2	89.32
[+IO	2/0 + 1/1	89.43
IO+]	1/1 + 0/2	89.42

9

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'}

10

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?

11

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?

12

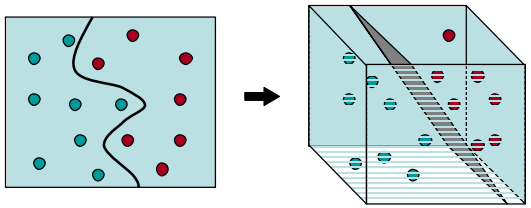
Information Packaging

- 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

13

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.*



14

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?

15

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? ☺)

16

Experiment 2

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

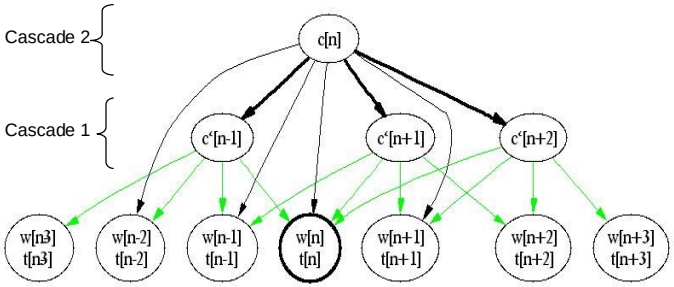
$$P(c_n \mid < w_n, t_n >, < w_{n-1}, t_{n-1}, c'_{n-1} >, < w_{n-2}, t_{n-2}, c'_{n-2} > \ldots)$$

instead of

$$P(c_n \mid < w_n, t_n >, < w_{n-1}, t_{n-1} >, < w_{n-2}, t_{n-2} > \ldots)$$

17

Experiment 2: Context for IOB1



18

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

19

Experiment 2: Results

	Word/POS context	Chunk Tag	F score
IOB1	L=2 / R=1	1 / 2	90.12
IOB2	L=2 / R=1	1 / 0	89.30
IOE1	L=1 / R=2	1 / 2	89.55
IOE2	L=1 / R=2	0 / 1	89.73
[+]	2 / 1 + 0 / 2	0 / 0 + 0 / 0	89.32
[+ IO	2 / 0 + 1 / 1	0 / 0 + 1 / 1	89.78
IO +]	1 / 1 + 0 / 2	1 / 1 + 0 / 0	89.86

20

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

21

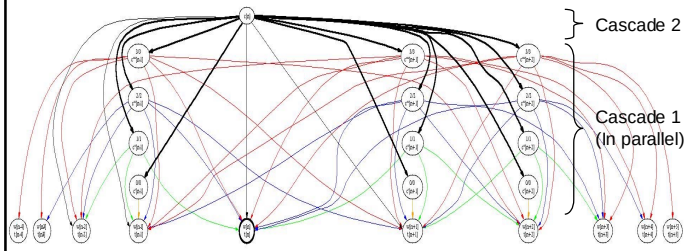
Experiment 3

- 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

$$\lambda_1 P(c_n \mid <context\,1>) + \lambda_2 P(c_n \mid <context\,2>)....$$

22

Experiment 3: Context for IOB1



23

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

24

Experiment 3: Results

	Word/POS	Chunk	Combinations	F
IOB1	2/1	1/1	0/0 1/1 2/2 3/3	90.53
IOB2	2/1	1/0	2/1	89.30
IOE1	1/2	1/2	0/0 1/1 2/2 3/3	90.03
IOE2	1/2	0/1	1/2	89.73
[+]	2/1 + 0/2	0/0+0/0	- + -	89.32
[+ IO	2/0 + 1/1	0/0+1/1	- + 0/1 1/2 2/3 3/4	89.91
IO +]	1/1 + 0/2	1/1+0/0	0/1 1/2 2/3 3/4+ -	90.03

25

Experiment 4

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

26

Experiment 4: Results

	W/T	C	Combinations	F
IOB1	3/3(3)	1/1	0/0(1) 1/1(1) 2/2(3) 3/3(3)	90.89 (0.63)
IOB2	3/3(3)	1/0	3/3(3)	89.72 (0.79)
IOE1	2/3(3)	1/2	0/0(1) 1/1(1) 2/2(3) 3/3(3)	90.12 (0.27)
IOE2	2/3(3)	0/1	2/3(3)	90.02 (0.48)
[+]	4/3(3) + 4/4(3)	0/0+0/0	- + -	90.08 (0.57)
[+ IO	4/3(3) + 3/3(3)	0/0+1/1	- + 0/1(1) 1/2(3) 2/3(3) 3/4(3)	90.35 (0.75)
IO +]	3/3(3) + 2/3(3)	1/1+0/0	0/1(1) 1/2(3) 2/3(3) 3/4(3) + -	90.23 (0.73)

27

Conclusions & Questions

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

28

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.

29

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?

30

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?

31

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.

33

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.

34

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?

35