Homework

- Read chapter 7
- Repeat confusable experiments using Timbl. Try different options settings to see what works best.
- **WARNING:** Timbl’s idea of C4.5 format isn’t quite the same as C4.5’s
- Now try the same thing without using part-of-speech tags. You can use words, parts of words, capitalization, etc., but not tags
- How well does this work? What kinds of things does it still get wrong?

Memory-based Tagging

- MBT (Zavrel, Daelemans, van den Bosch) is a part-of-speech tagging system based on decision tree and instance based learners
- First step: construct a lexicon of words in training data, annotated with ‘ambitags’
- Next, construct classifiers for known and unknown words using a sliding window of features
- Classifier for known words uses two previously assigned tags, ambitag of current word, and ambitag of next word
- Classifier for unknown words uses previously assigned tag, ambitag of next word, and the first letter and last three letters of current word

---

Memory-based Tagging

- Case representation for known word IGTREE classifier:

<table>
<thead>
<tr>
<th>Pierre</th>
<th>d d f a t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinken</td>
<td>np np np</td>
</tr>
<tr>
<td>years</td>
<td>cd nns jj-np</td>
</tr>
<tr>
<td>old</td>
<td>cd nns jj-np</td>
</tr>
</tbody>
</table>

- Case representation for unknown word IB1 classifier:

<table>
<thead>
<tr>
<th>Pierre</th>
<th>P d a s s s t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vinken</td>
<td>V np np k e n</td>
</tr>
<tr>
<td>years</td>
<td>y cd jj-np a r s</td>
</tr>
<tr>
<td>old</td>
<td>o nns o l d</td>
</tr>
</tbody>
</table>

- Tagging proceeds left to right, with no backtracking or global optimization
- Training and application of model is relatively cheap, not too far from HMM methods
- Results are competitive with the state of the art: 96.4% on WSJ
- Used as tagger for Corpus Gesproken Nederlands project
- Source code freely available, easy to use
Memory-based Tagging

- LOB corpus tagging experiment:

<table>
<thead>
<tr>
<th>System</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigram</td>
<td>96.1</td>
</tr>
<tr>
<td>TBL</td>
<td>96.6</td>
</tr>
<tr>
<td>MBT</td>
<td>97.0</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>97.4</td>
</tr>
<tr>
<td>Ensemble</td>
<td>97.9</td>
</tr>
</tbody>
</table>

- Tagging homework assignment:

<table>
<thead>
<tr>
<th>System</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>90.5</td>
</tr>
<tr>
<td>bigram</td>
<td>94.1</td>
</tr>
<tr>
<td>MBT (bigram)</td>
<td>94.1</td>
</tr>
<tr>
<td>TnT</td>
<td>96.1</td>
</tr>
<tr>
<td>MBT</td>
<td>96.5</td>
</tr>
</tbody>
</table>

Instance-based learning

- IBL classifiers relate query instances to most similar instance in training data
- Applies a very simple learning rule (use majority class) to a local neighborhood of the feature space
- Can be used for both discrete and continuous features and target functions
- Variants like IB2 and TRIBL can improve efficiency without reducing accuracy too much
- Can run into overfitting problems, but in a way (it’s been argued) that is particularly well suited to NLP problems

Rule induction

- Early AI systems and not-so-early NLP systems used large collections of hand written rules
- Rule induction system try to produce sets of rules automatically from training data
- Classic rule-based systems: RIPPER (propositional logic), FOIL (first order)
- Inductive Logic Programming produces Horn clauses (Prolog programs)
- Grammar induction
- Decision trees, Transformation-based learning

Trees to rules

- Decision trees/lists/stumps have a natural interpretation as rule sets
- Each path from the root to a leaf is a rule of the form:

  \[
  \text{if } \text{outlook} = \text{rain} \land \text{wind} = \text{strong} \quad \text{then } \text{class} = \text{no}
  \]
- Explicitly rewriting trees as rules may give model greater generalization power and transparency
- C4.5 includes an algorithm for performing this transformation (c4.5rules)
Trees to rules

- Decision tree for $F=1$ or $J=1$:

![Decision tree diagram]

- Some rules from this tree include superfluous conditions:

```
if $F=1$ & $G=0$ & $J=1$ & $K=1$
then class=yes
```

- By deleting conditions, we can get to the rule we want:

```
if $J=1$ & $K=1$
then class=yes
```

Deleting conditions

- Should we delete condition $X$ from rule $R$?

<table>
<thead>
<tr>
<th>Satisfies $X$</th>
<th>Class $C$</th>
<th>$Y_1$</th>
<th>$E_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doesn't satisfy $X$</td>
<td>$Y_2$</td>
<td>$E_2$</td>
<td></td>
</tr>
</tbody>
</table>

- The original rule $R$ assigns class $C$ to $Y_1 + E_1$ examples that satisfy $X$ ($E_1$ of them incorrectly)

- The new rule would assign class $C$ to $Y_2 + E_2$ examples that don't satisfy $X$ ($E_2$ of them incorrectly)

- Only retain $X$ if the difference is statistically significant

- Or, we could use the pessimistic error rate estimate: remove $X$ if doing so doesn't increase the upper bound on the expected error

Given a decision tree, we can iterate over the paths deleting conditions (greedy search)

- This gives the result of **global** pruning and tree reconfiguration

- But, now each instance may have more than one rule that could apply to it, or none at all!

- We need a way to resolve conflicts between applicable rules

- Extrinsic ordering is always popular, but can be difficult to implement in a transparent manner
Resolving conflicts

- C4.5 organizes all rules into sets, one for each class
- Now within each set, rule order doesn’t matter since all rules have the right-hand side
- Collectively a rule set is a binary classifier which can be evaluated by its precision and recall
- Each rule set can be optimized using the Minimum Description Length Principle

Minimum Description Length

- Suppose $S$ and $R$ both have a copy of the training data, but only $S$ has the values of the target function
- $S$ could just transmit a list of classes to $R$
- Or, $S$ could send a classification theory (e.g., a decision tree), plus a list of exceptions
- Which is better, a simple theory with lots of exceptions, or a complex theory with no exceptions?
- The Minimum Description Length (MDL) Principle (Rissanen 1983) says the best theory minimizes the sum of the lengths of the theory and the exceptions:

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

Minimum Description Length

- Remember maximum a posteriori estimation:

$$h_{MAP} = \arg\max_{h \in H} P(D|h) P(h)$$
$$= \arg\max_{h \in H} (\log_2 P(D|h) + \log_2 P(h))$$
$$= \arg\min_{h \in H} (\log_2 P(D|h) - \log_2 P(h))$$

- And remember the Source Coding Theorem:

$$L_{C_X}(x) = -\log_2 P(x)$$

- So, we can rewrite the MAP estimate:

$$h_{MAP} = \arg\min_{h \in H} L_{C_H}(h) + L_{D|h}(D|h)$$

and $h_{MAP} = h_{MDL}$ when $C_1 = C_H$ and $C_2 = C_{D|h}$
Minimum Description Length

- The cost of a set of rules is the sum of the discounted cost of their LHS’s, but again the order doesn’t matter so we give ourselves credit.

- Say the rules cover \( r \) out of \( n \) cases, with \( f_p \) false positives and \( f_n \) false negatives. Then cost of the exceptions is bounded by:

\[
\log_2 \left( \frac{r}{f_p} \right) + \log_2 \left( \frac{n-r}{f_n} \right)
\]

- This tends to overestimate the length of the theory, so we discount it:

\[
L = \text{exception bits} + W \times \text{theory bits}
\]

Trees to rules

- Once the MDL subset of each rule set has been found, we need to order the subsets.

- False positives are worse than false negatives, because they prevent later rule sets from applying.

- So, we choose as the first rule set the one which makes the fewest false positive errors, recompute the error rates, and repeat.

- Next, the default class is the one which is most common among those training examples to which no rule applies.

- Finally, if there are any rules whose removal would reduce the error rate on the training data, get rid of them.

Minimum Description Length

- We want to find a subset of our rule set which minimizes \( L \).

- Description length \( L \) has lots of local minima, so hill climbing fails. Instead, Quinlan uses simulated annealing.

- Choose a rule at random.

- If removing the rule (or putting it back in) would reduce \( L \), do so.

- If it would increase \( L \) by \( d \) bits, do so with probability \( e^{-d/K} \).

- Repeat, gradually reducing the ‘temperature’ \( K \) until convergence.

Trees to rules

- C4.5’s method of producing rule sets from trees allows more effective pruning than a simple bottom-up strategy.

- The resulting rules are more compact and more easily understood than a decision tree.

- Rule systems can in theory be edited by hand (though I’m not sure under what circumstances that would be a good idea).
Transformation-based learning

- To keep things reasonable, HMM taggers can take only a small window of context into account
- This leads them to make predictable errors which can be corrected by a set of fix-up rules
- Post-processing can also be used to adapt a model to a different situation (e.g., written → spoken language)
- These rules can be written by hand, but can be expensive to write and especially to maintain
- Wouldn’t it be cool if we could do it automatically?

Brill (1992) proposed TBL as a solution to this problem (pp. 361–370)
- Start with a very simple baseline tagger
- Randomly generate a set of transformations which replace tags under certain conditions
- Choose the one which decreases the error rate the most, and add it to the system
- Iterate until you’ve had enough
- The collection of transformations can be converted to FST’s, composed, determinized, and minimized for very efficient storage and application (much faster than HMMs)

Transformation-based learning

- The LHS’s of candidate rules are generated from a set of templates:
  - word—1 (word+1) is tagged \( t \)
  - word—2 (word+2) is tagged \( t \)
  - one of word word—1…—2 (word+1…+2) is tagged \( t \)
  - one of word word—1…—3 (word+1…+3) is tagged \( t \)
  - word—1 is tagged \( t_1 \) and word+1 is tagged \( t_2 \)
  - word—1 (word+1) is tagged \( t_1 \) and word—2 (word+2) is tagged \( t_2 \)

- For each rule template \( T \), we get a set of rules of the form:
  
  \[
  \text{if } T \text{ then replace tag } X \text{ with tag } Y
  \]

- This yields an astronomical number of rules, but we don’t really have to construct them all. Clever indexing will let us only consider the tiny subset of rules which actually apply to something in the training data

When applied to the WSJ, the first rules it learns are:

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>When</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>VB</td>
<td>previous tag is TO</td>
</tr>
<tr>
<td>VBP</td>
<td>VB</td>
<td>one of the previous three tags is MD</td>
</tr>
<tr>
<td>NN</td>
<td>VB</td>
<td>one of the previous two tags is MD</td>
</tr>
<tr>
<td>VB</td>
<td>NN</td>
<td>one of the previous two tags is DT</td>
</tr>
<tr>
<td>VBD</td>
<td>VBN</td>
<td>one of the previous three tags is VBZ</td>
</tr>
<tr>
<td>VBN</td>
<td>VBD</td>
<td>previous tag is PRP</td>
</tr>
<tr>
<td>VBN</td>
<td>VBD</td>
<td>previous tag is NNP</td>
</tr>
<tr>
<td>VBD</td>
<td>VBN</td>
<td>previous tag is VBD</td>
</tr>
<tr>
<td>VBP</td>
<td>VB</td>
<td>previous tag is TO</td>
</tr>
<tr>
<td>POS</td>
<td>VBZ</td>
<td>previous tag is PRP</td>
</tr>
<tr>
<td>VB</td>
<td>VBP</td>
<td>previous tag is NNS</td>
</tr>
</tbody>
</table>

Transformation-based learning
Transformation-based learning

- The same strategy can be used to learn lexical relationships:

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>When</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN</td>
<td>RB</td>
<td>word two positions to the right is <em>as</em></td>
</tr>
<tr>
<td>VBP</td>
<td>VB</td>
<td>one of the two previous words is <em>n’t</em></td>
</tr>
</tbody>
</table>

- Some results on *closed vocabulary* experiments:

<table>
<thead>
<tr>
<th>System</th>
<th>Training size</th>
<th>Model size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>64,000</td>
<td>6,170</td>
<td>96.3</td>
</tr>
<tr>
<td>HMM</td>
<td>1,000,000</td>
<td>10,000</td>
<td>96.7</td>
</tr>
<tr>
<td>Lexical TBL</td>
<td>64,000</td>
<td>215</td>
<td>96.7</td>
</tr>
<tr>
<td>Lexical TBL</td>
<td>600,000</td>
<td>447</td>
<td>97.2</td>
</tr>
<tr>
<td>Non-lexical TBL</td>
<td>600,000</td>
<td>378</td>
<td>97.0</td>
</tr>
</tbody>
</table>

Unknown words

- Some unknown word rules:

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>When</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>NNS</td>
<td>word ends in <em>-s</em></td>
</tr>
<tr>
<td>NN</td>
<td>CD</td>
<td>word contains character .</td>
</tr>
<tr>
<td>NN</td>
<td>JJ</td>
<td>word contains character -</td>
</tr>
<tr>
<td>NN</td>
<td>VBN</td>
<td>word has suffix <em>-ed</em></td>
</tr>
<tr>
<td>NN</td>
<td>VBG</td>
<td>word has suffix <em>-ing</em></td>
</tr>
<tr>
<td>??</td>
<td>RB</td>
<td>word has suffix <em>-ly</em></td>
</tr>
<tr>
<td>??</td>
<td>JJ</td>
<td>adding suffix <em>-ly</em> results in a known word</td>
</tr>
<tr>
<td>NN</td>
<td>CD</td>
<td>word $ can appear to the left</td>
</tr>
</tbody>
</table>

Unsupervised learning

- TBL as described so far is a supervised learning method

- TBL can also be used *unsupervised* (well, semi-supervised) to learn adaptation rules

- Base classifier tags each word with its ‘ambitag’

- Proceed as before, but with a new scoring procedure which takes advantage of unambiguous words
Unsupervised learning

- Suppose we have the rule \textit{Replace }X\textit{ with }Y\textit{ in context }C\textit{, where }Y \in X\textit{. Now consider each tag }Z \in X \textit{ where }Z \neq Y\textit{ and find:}

\[
R = \arg\max_Z \frac{\text{freq}(Y)}{\text{freq}(Z)} \times \text{incontext}(Z; C)
\]

The score for the rule is:

\[
\text{incontext}(Y, C) - \frac{\text{freq}(Y)}{\text{freq}(R)} \times \text{incontext}(R, C)
\]

- This is supposed to prefer transformations that pick out tags which occur more frequently in this context, correcting for relative frequencies of tags

Transformation-based learning

- TBL automatically induces a set of error-correcting rules for a simple statistical classifier
- Training is slow, but the final model is very fast
- TBL seems to be immune to overtraining – later rules don’t help much, but also don’t seem to hurt
- This may be because it is directly minimizing the error rate, rather than a proxy (likelihood, entropy, etc.)
- There are lots of extensions imaginable (e.g., more sophisticated base taggers, multiple base taggers, genetic algorithms for ‘evolving’ rule templates, . . .)

Unsupervised learning

- For instance \textit{can} could be a noun, a verb, or a modal. In the context \textit{The can will be crushed}, we know it must be a noun because when an unambiguous word occurs after \textit{the}, it’s almost always a noun
- Some unknown word rules:

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
<th>When</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN_VB_VP</td>
<td>NN</td>
<td>previous tag is NNS</td>
</tr>
<tr>
<td>NN_VB</td>
<td>VB</td>
<td>previous tag is MD</td>
</tr>
<tr>
<td>JJ_NNP</td>
<td>JJ</td>
<td>following tag is NNS</td>
</tr>
</tbody>
</table>

- In closed vocabulary experiments, WSJ accuracy improved from 90.7% to 95.1% and Brown accuracy improved from 89.9% to 95.6%
- In similar experiments, Baum-Welch estimation yields 83–92% accuracy