Grouping Words

Linguistic Objects in this Course

- Trees (with strings at the nodes)
  - Syntax, semantics
  - Algorithms: Generation, parsing, inside-outside, build semantics

- Sequences (of strings)
  - n-grams, tag sequences
  - morpheme sequences, phoneme sequences
  - Algorithms: Finite-state, best-paths, forward-backward

- “Atoms” (unanalyzed strings)
  - Words, morphemes
  - Represent by contexts - other words they occur with
  - Algorithms: Grouping similar words, splitting words into senses

A Concordance for “party” from www.webcorp.org.uk

What Good are Word Senses?

- thing. She was talking at a party thrown at Daphne’s restaurant in
  - have turned it into the hot dinner-party topic. The comedy is the
  - selection for the World Cup party, which will be announced on May 1
  - in the 1983 general election for a party which, when it could not bear to
  - to attack the Scottish National Party, who look set to seize Perth and
  - that had been passed to a second party who made a financial decision
  - the by-pass there will be a street party. “Then,” he says, “we are going
  - number-crunchers within the Labour party, there now seems little doubt
  - political tradition and the same party. They are both relatively Anglophilic
  - he told Tony Blair’s modernised party they must not retreat into “warm
  - "Oh no, I’m just here for the party,” they said. “I think it’s terrible
  - A future obliges each party to the contract to fulfil it by
  - be signed by or on behalf of each party to the contract.” Mr David N
What Good are Word Senses?

- John threw a “rain forest” party last December. His living room was full of plants and his box was playing Brazilian music ...

What Good are Word Senses?

- Replace word w with sense s
  - Splits w into senses: distinguishes this token of w from tokens with sense t
  - Groups w with other words: groups this token of w with tokens of x that also have sense s

What Good are Word Senses?

- number-crunchers within the Labour party, there now seems little doubt
- political tradition and the same party. They are both relatively Anglophilic
- he told Tony Blair’s modernised party they must not retreat into “warm thing. She was talking at a party thrown at Daphne’s restaurant in
- have turned it into the hot dinner-party topic. The comedy is the
- selection for the World Cup party, which will be announced on May 1
- the by-pass there will be a street party. “Then,” he says, ”we are going
- “Oh no, I’m just here for the party,” they said. “I think it’s terrible
- an appearance at the annual awards bash, but feels in no fit state to
- -known families at a fundraising bash on Thursday night for Learning
- Who was paying for the bash? The only clue was the name Asprey,
- Mail, always hosted the annual bash for the Scottish Labour front-
- popular. Their method is to bash sense into criminals with a short,
- just cut off people’s heads and bash their brains out over the floor,

What Good are Word Senses?

- adjacent words (or their senses)
- Grammatically related words (subject, object, …)
- Other nearby words
- Topic of document
- Sense of other tokens of the word in the same document

Cues to Word Sense
Word Classes by Tagging

- Every tag is a kind of class
- Tagger assigns a class to each word token

```
<table>
<thead>
<tr>
<th>prob from tag bigram model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start PN Verb Det Noun Prep Noun Prep</td>
</tr>
<tr>
<td>0.4 0.6 0.001</td>
</tr>
</tbody>
</table>
```

Bill directed a cortege of autos through

```
<table>
<thead>
<tr>
<th>prob from unigram replacement</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN   Verb    Det     Noun  Prep Noun   Prep</td>
</tr>
<tr>
<td>0.4 0.6 0.001</td>
</tr>
</tbody>
</table>
```

Words as Vectors

- Represent each word type \( w \) by a point in \( k \)-dimensional space
- e.g., \( k \) is size of vocabulary
- the 17th coordinate of \( w \) represents strength of \( w \)'s association with vocabulary word 17

```
(0, 0, 3, 1, 0, 7, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
```

Words as Vectors

- Represent each word type \( w \) by a point in \( k \)-dimensional space
- e.g., \( k \) is size of vocabulary
- the 17th coordinate of \( w \) represents strength of \( w \)'s association with vocabulary word 17

```
(0, 0, 3, 1, 0, 7, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
```

Plot all word types in \( k \)-dimensional space

Look for clusters of close-together types

Learning Word Classes

- Every tag is a kind of class
- Tagger assigns a class to each word token
  - \{party/N, bash/N\} vs. \{party/V, bash/V\}
  - What good are these groupings?
  - Good for predicting next word or its class!

- Role of forward-backward algorithm?
  - It adjusts classes etc. in order to predict sequence of words better (with lower perplexity)

```
From corpus:
Jim Jeffords abandoned the Republican party. There were lots of abbots and nuns dancing at that party. The party above the art gallery was, above all, a laboratory for synthesizing zygotes and beer.
```

Words as Vectors

- Represent each word type \( w \) by a point in \( k \)-dimensional space
  - e.g., \( k \) is size of vocabulary
  - the 17th coordinate of \( w \) represents strength of \( w \)'s association with vocabulary word 17

```
(0, 0, 3, 1, 0, 7, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
```

```
aband
```

```
abandon
```

```
abduct
```

```
above
```

```
abdu
```

```
zygote
```

```
zyurgy
```

```
plot
```

```
through
```
Learning Classes by Clustering

- Plot all word types in k-dimensional space
- Look for clusters of close-together types

Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: dist(A,B) = min dist(a,b) for a∈A, b∈B
  - Complete-link: dist(A,B) = max dist(a,b) for a∈A, b∈B

Bottom-Up Clustering – Single-Link

- Each word type is a single-point cluster
- merge
- Again, merge closest pair of clusters:
  - Single-link: clusters are close if any of their points are
    dist(A,B) = min dist(a,b) for a∈A, b∈B

Bottom-Up Clustering – Complete-Link

- Again, merge closest pair of clusters:
  - Complete-link: clusters are close only if all of their points are
    dist(A,B) = max dist(a,b) for a∈A, b∈B

Bottom-Up Clustering – Complete-Link

- Slow to find closest pair – need quadratically many distances

Example from Manning & Schütze
Bottom-Up Clustering

- Start with one cluster per point
- Repeatedly merge 2 closest clusters
  - Single-link: \( \text{dist}(A, B) = \min \text{dist}(a, b) \) for \( a \in A, b \in B \)
  - Complete-link: \( \text{dist}(A, B) = \max \text{dist}(a, b) \) for \( a \in A, b \in B \)
    - too slow to update cluster distances after each merge; but 3 alternatives!
  - Average-link: \( \text{dist}(A, B) = \text{mean} \text{dist}(a, b) \) for \( a \in A, b \in B \)
  - Centroid-link: \( \text{dist}(A, B) = \text{dist}(\text{mean}(A), \text{mean}(B)) \)
- Stop when clusters are “big enough”
  - e.g., provide adequate support for backoff (on a development corpus)
- Some flexibility in defining \( \text{dist}(a, b) \)
  - Might not be Euclidean distance; e.g., use vector angle

EM Clustering (for k clusters)

- EM algorithm
  - Viterbi version – called “k-means clustering”
  - Full EM version – called “Gaussian mixtures”
- Expectation step: Use current parameters (and observations) to reconstruct hidden structure
- Maximization step: Use that hidden structure (and observations) to reestimate parameters
- Parameters: \( k \) points representing cluster centers
- Hidden structure: for each data point (word type), which center generated it?

EM Clustering (for k clusters)

- [see spreadsheet animation]