FSNLP/
Ch.14 Clustering

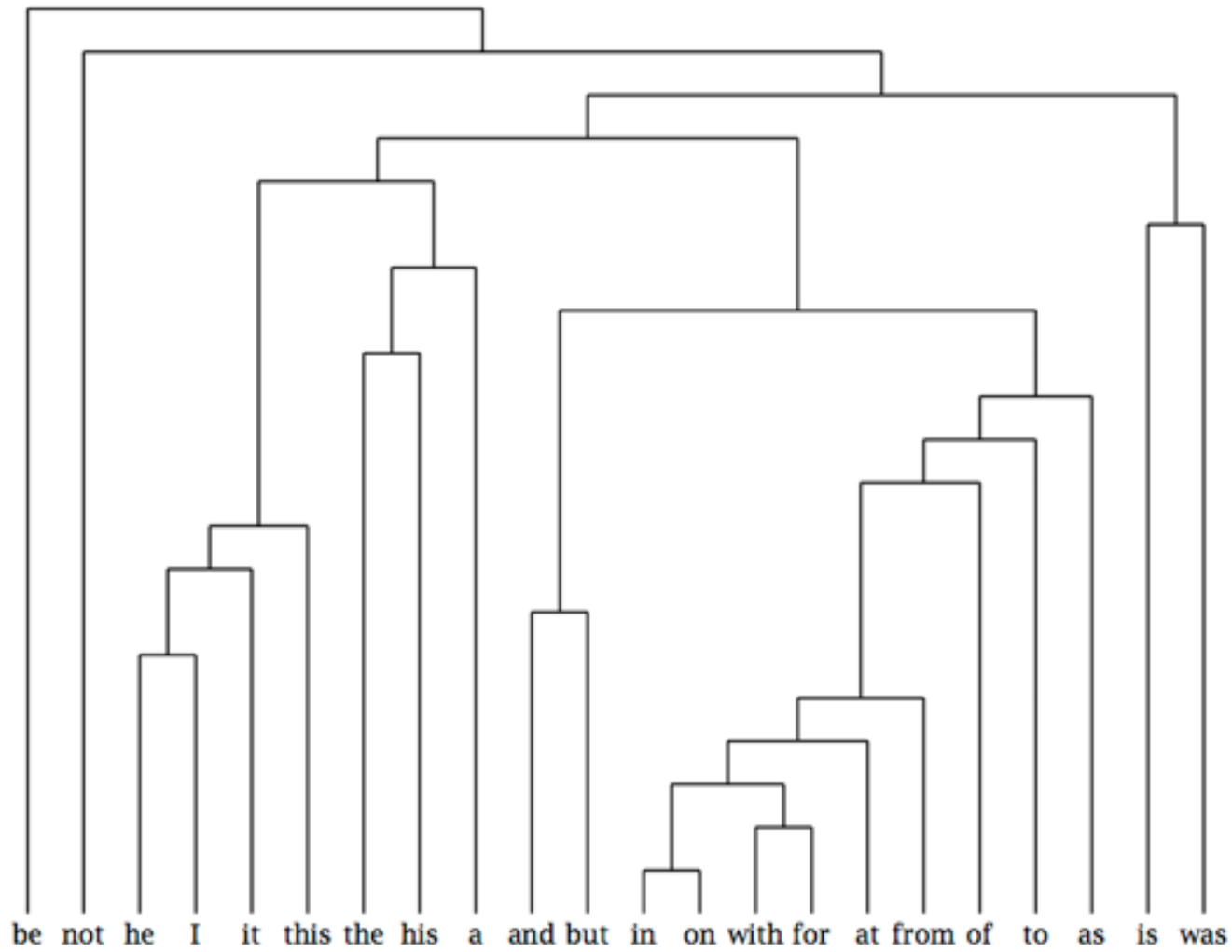
相澤研 B4 須田山強真

クラスタリング

- 類似度に基づく
  グループ分け

- 類似度
  - 文脈で近傍に登場する
    単語の分布

- 教師なし学習
  - 分類(教師あり学習)との違い



目的

- Explanatory data analysisの手法として
  - 現象の特性を知る
- 一般化
  - 類似した単語のクラスタから用法を得る
  - 出現 “On {Monday, Tuesday}”, Fridayは曜日の一つ
    - On Fridayは正しそう

クラスタリング

- 種類
  - 階層の有無
  - soft/hard 割り当て
- 距離と類似度の混同に注意

階層的クラスタ

- 類似度関数は単調とする:

  合併によって類似度は増加しない

- 構築
  - top-down:グループ内類似度を

    最大化するように分割を繰り返す

  - bottom-up: 一要素のみのクラスタからはじめて

    類似した2つのクラスタを合併

類似度関数の例

- 階層的な場合を仮定
- 局所/大域的な良さ
- single link: クラスタ内の要素をそれぞれとり、その間で最も高い類似度を用いる(MSTと同様に解く)
- complete link: 同様の設定で最も低い類似度を用いる
- 名前はグラフ理論的解釈に基づく

グループ平均

- {single, complete}-linkの間で妥協した類似度として
- 時間計算量 $O(n^2)$
  - 要素間での類似度を前計算: $O(n^2)$
  - 特徴ベクトルの和を持っておくことで合併はO(1)
  - 合併は O(n)

言語モデルへの応用

- bi-gramのもとでcross-entropyを最小化する
- 1次マルコフ性のクラスタへの一般化をともに用いる
  - c2の単語の出現はその直前の語が属するc1にのみ依存 (c1, c2はクラスタ)

言語モデルへの応用

$$P(w_1 \dots w_n) \approx P(w_1) \prod_{j=2}^{n} P(w_j|w_{j-1})$$

(1次マルコフ性)

$$H(L) = -\sum_{(w_1,w_2)\in \boldsymbol{Bi}} \frac{C(w_1,w_2)}{|\boldsymbol{Bi}|} \log P(w_2|w_1) \text{ where } L = w_1 \dots w_n$$

(cross-entropy最小化)

$$P(w_2|w_1) \approx P(w_2|c_2)P(c_2|c_1) \text{ with } w_1 \in c_1$$

(クラスタへの一般化)

十分長いテキストのもとで近似を用いると

$$H(L) = H(w) - I(c_1; c_2)$$

言語モデルへの応用

- 前の式から階層上で隣接するクラスタ間の相互情報量を最大化するものを選ぶ

flat clustering

- K-means (hard assignment)
- 適当な分布のもとでEM algorithm(soft assignment)

K-means

k: ハイパーパラメータ

入力: データセット, 2点の距離 d(x,y)

出力: k個の中心

1. k個の中心を初期化

2. 基準が満たされるまで繰り返す

    1. データ点に近い中心を割り当てる

    2. 中心に対して割り当てられた点から位置を再計算

EM algorithm

入力: パラメータを持つ分布, データセット

出力: 分布のパラメータ

1. パラメータを初期化
2. 以下を収束するまで繰り返す
    1. パラメータを固定して条件つき期待値を計算(E-step)
    2. 期待値のもとでパラメータに関して尤度を最大化(M-step)

- {E, M}-stepによって尤度は単調に増加する
- 少なくとも局所解を得られる

EM algorithmの例

- Baum-Welch algorithm
- Inside-outside algorithm
- Unsupervised WSD
- K-means
  - 分散を小さく固定したガウス混合分布として
  - 互いに近い中心があるケースの扱いは異なる