

N-gram統計量からの係り受け情報の復元

2011/09/22

NLP若手の会第6回シンポジウム

プリファードインフラストラクチャ

海野 裕也, 岡野原 大輔

{unno, hillbig}@preferred.jp

概要

- Nグラムの統計量のみから係り受け構造を復元する実験を行った
- 簡単な指標のみを使っても予想以上にうまく復元できた
- 実験データを見ながら問題点の整理を行った
- 本日は動機や諸問題に関する議論を行いたい

動機：文の構造を検索に応用できないか

クエリ：大規模並列処理



．．． 大規模超並列実時間処理を．．．



．．． 大規模コーパス中の並列句処理基準．．．

- 単語境界情報だけだと、修飾句の挿入に対処できない
- 修飾部を飛ばしてクエリの出現を探したい

係り受け構造を復元できないだろうか？

- 係り受け解析の教師あり・半教師あり学習は成功している
- 一方で**教師データの作成はコストが高い**
- 教師データがないと学習できない
 - 多言語、分野適応、単語単位・文字単位の係り受け

係り受け構造を低コストで復元できないか？

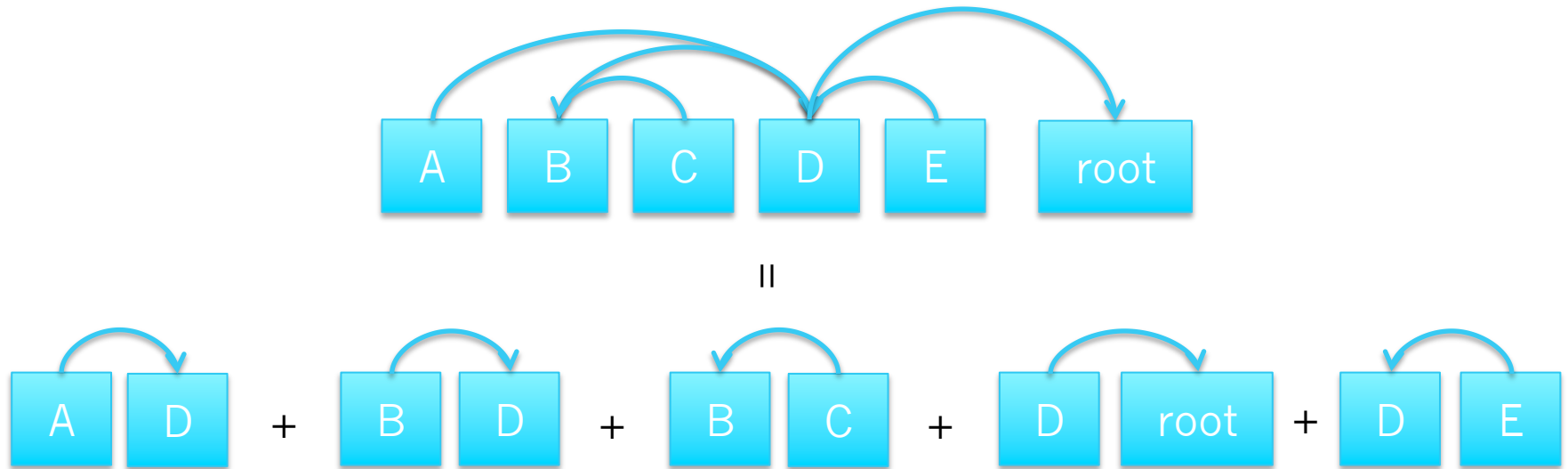
アイデア：大半の係り受け関係は隣接しているのなら、隣接頻度の情報から係り関係を復元できるはず

- 隣接している単語が係りやすい
 - 半分以上の係り受けは隣接することが知られている
 - 隣接頻度が係りやすさの指標になりそう
- 修飾部は省略されやすい
 - 例：「並列・分散・処理」
 - 修飾部の「分散」が省略された、「並列・処理」という表現もたくさん出現する
- **Nグラムの統計量のみで**係りやすさを表現できるのではないか？

今回は長い複合名詞を対象にした

- 複合名詞は堅い文書に特に多く、一部を探したいニーズもある
- 複合名詞の係り関係に関する正解データがない
- 医療用語
 - 大腿骨頸部内側骨折
 - 閉塞性動脈硬化症
- 政治用語
 - 各府省情報化統括責任者補佐官等連絡会議
 - 国家安全保障問題担当大統領補佐官
- その他
 - 記録的短時間大雨情報
 - 戦略的創造研究推進事業

Eisnerアルゴリズム [Eisner96]



- 係り受け木 T に対するスコア $S(T)$ を、局所的なスコアの和で表す
- $S(T) = \sum_{(m, h) \in T} s(m, h)$
 - (m, h) は T 中のすべての修飾・非修飾ペア
- $S(T)$ を最大にする T_{opt} は時間計算量 $O(n^3)$ で求まる

Google N-gramデータから推定した自己相互情報量 (PMI) でスコア関数を設計する

$$s(m, h) = \log \frac{P(m, h)}{P(m)P(h)} = \log \frac{\#(mh)}{\#(m)\#(h)} + \text{const}$$

- Google日本語Nグラムデータの頻度を利用する
 - $\#(mh)$ は m, h のバイグラム頻度
 - $\#(m)$ は m のユニグラム頻度
- Eisnerの式で足し合わせる $s(m, h)$ の個数は T によらず一定なので、上記の const は無視して良い

結果：予想以上にうまくいくが、変なところも



具体例的に間違った例を整理する

1. 不自然な修飾部ができてしまうケース

- 明らかに切れない、**接尾表現**から修飾部が始まる
- ～性、～県、～的、～系、～検討、～補償・・・

2. 修飾部に係ってしまうケース

- **接頭表現**などの明らかな修飾部に係ってしまう
- 大～、超～、準～、特別～、簡易～・・・

3. 構造的に不自然なケース

- ペアでは**妥当な候補が複数**あるときに、不自然な構造になる
- 例：「超・大・規模・分散・並列・処理」
 - 「超・大」「超・分散」「超・並列」はどれも自然

1. 不自然な修飾部ができてしまうケース



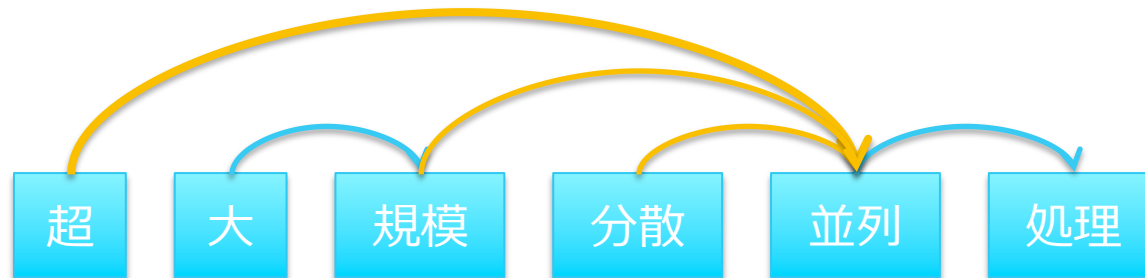
- 「性」が単独で「硬化」を修飾している
- 「閉塞・性」というフレーズが検出できていない
- 不自然に始まる修飾部を抑制するスコアがない
- 接尾表現は頻度が高いため、PMIが低くなりがち
 - 閉塞：580K 性：117M 閉塞・性：72K
 - 閉塞：580K 症：13.4M 閉塞・症：20.5K

2. 修飾部に係ってしまうケース



- 「大」を「情報」が修飾している
- 「情報・航海」という表現がそもそも出現しない
 - 情報：542M 大：114M 情報・大：68K
 - 情報：542M 航海：1.66M **情報・航海：77**
- 「大」は普遍的に修飾部になりやすい、という情報が必要

3. 構造的に不自然なケース



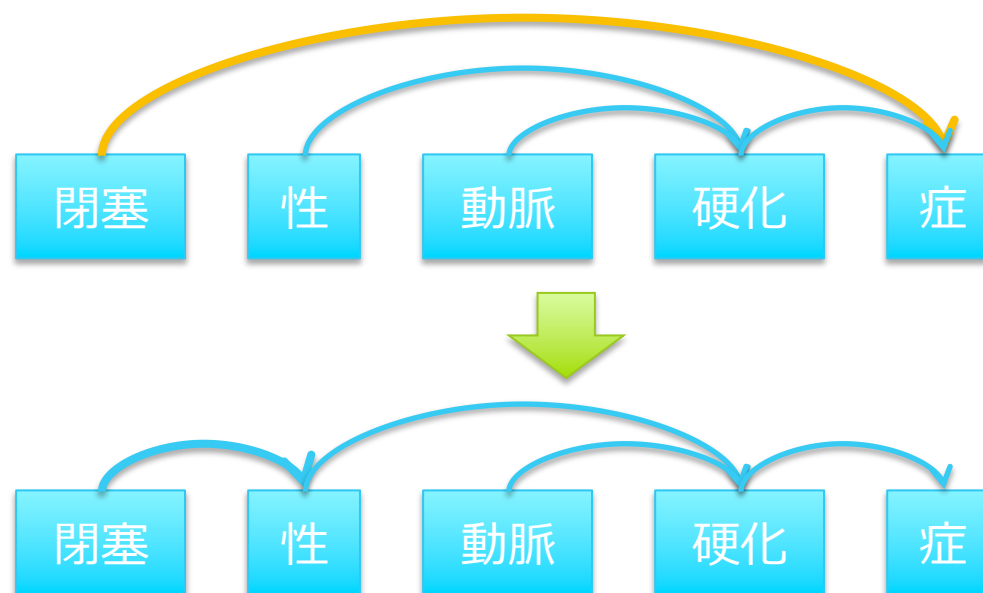
- 「超・大」「超・並列」など、妥当な候補が文中に複数あるとだいたいうまくいかない
- ローカルなスコアのみに依存しているのが問題
- 複雑な構造に対するペナルティが必要

改善案1：高頻度の接尾辞の影響を排除する

- PMIの代わりに、単位バイグラム言語モデルの接続スコアを使う

$$s(m, h) = \log P(h|m) = \log \frac{\#(mh)}{\#(m)}$$

- いくつかは改善された



議論1：構文情報は単語連接以上の情報をもたらすのか

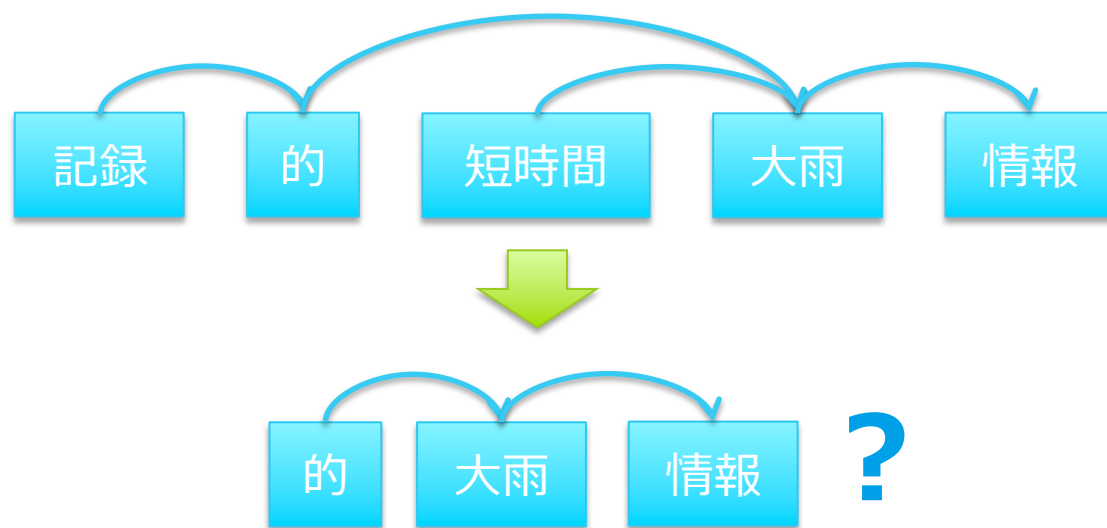
クエリ：大規模並列処理

・・・大規模超並列実時間処理を・・・

編集距離：2

- 編集距離のようなスコアを導入できれば大部分はうまく動くようにも思える
- 木構造が決定的に重要なケースはあるか？ 検索以外のアプリケーションで重要なシーンはあるか？

議論2：複合語における適切な係り受け構造とは？



- 係り受け木のパスを拾っても意味が通らないこともある
- そもそも構造が適切ではない？ 抽出の仕方が適切でない？
- 構文情報をどのように活用すべきか

関連研究

確率的単語分割 [工藤05][岡野原+06]

	確	率	的	単	語	分	割	コ	ー	パ	ス	
(1)	1	0	1	1	0	1	0	1	0	0	0	1
(2)	0.95	0.05	0.95	0.95	0.05	0.95	0.05	0.95	0.05	0.05	0.05	0.95
(3)	0.99	0.01	0.99	0.89	0.18	0.85	0.19	0.95	0.0	0.0	0.0	0.99

(1) 形態素解析結果 (2) 従来のSSC ($\alpha=0.95$) (3) 確率的単語分割

- 単語分割誤りに対してロバストにするために、単語分割を確率的に出力する
- 単語境界か否かの確率の積で単語の有無をスコア化する
- 検索結果がロバストになる

関連研究

大規模PMI情報の係り受け解析への応用 [Zhou+11]

- 単語間のPMIを統計的係り受け関係の特徴量に入れる
- PMIの推定にはGoogleでのヒットカウントを利用する
- これだけで、+1 ~ 2ポイントの精度向上

参考文献

- [Eisner96] J. M. Eisner. **Three New Probabilistic Models for Dependency Parsing: An Exploration.** COLING '96.
- [工藤05] 工藤拓. **形態素周辺確率を用いた分かち書きの一般化とその応用.** 言語処理学会全国大会 '05.
- [岡野原+06] 岡野原大輔, 工藤拓, 森信介. **形態素周辺確率を用いた確率的単語分割コーパスの構築とその応用.** NLP若手の会シンポジウム '06.
- [Zhou+11] G. Zhou, J. Zhao, K. Liu, L. Cai. **Exploiting Web-Derived Selectional Preference to Improve Statistical Dependency Parsing.** ACL '11.