3T-02

# 未解析コーパスからの依存関係単語対の収集

麻野間 直樹　中岩 浩巳
NTT コミュニケーション科学基礎研究所
{asanoma,nakaiwa}@cslab.kecl.ntt.co.jp

## 1 はじめに

テキストコーパスから共起単語を収集することは，自然言語処理ではしばしば行われる．その目的としては，定型的表現などの言語知識の獲得[Smadja 93]や，共起単語の頻度情報をもとに語義曖昧性解消[Leacock 98][Shütze 92]などの他のタスクへの利用がある．

この際に，「動詞と目的語」といったような直接依存関係にある２つ組の単語対（以下，依存単語対）を共起単語として収集したほうが，共起情報の有効性を期待できる．

テキストコーパスから依存単語対を得るには，未解析コーパスに構文解析系（パーザ）を適用するか，Penn Treebank[Marcus 93]のような人手で解析情報を付与したコーパスを用いる．前者の方法では，この解析にはある程度の時間がかかり，得られる解析結果には解析誤りを含むことが多い．また後者の方法では，単語間の依存関係まで記述されておらず，解析済みコーパスを大量に取得することも難しい．

本稿では，入手の容易な未解析のコーパスから，依存単語対を効率よく収集する方法を検討する．

## 2 共起取得法

従来，未解析のコーパスから共起単語を取得する方法（共起取得法）においては，以下のような共起単語の選択基準を用いていた．

一文内共起：一文内に同時に出現する単語を共起単語とする

n単語ウィンドウ：単語間距離がn内に同時に出現する単語を共起単語とする（n＝1 の場合は“単語 bigram”）

m文字ウィンドウ：m 文字内に同時に出現する単語を共起単語とする

また我々は，品詞情報付きのコーパスを用いることを前提とした共起単語の収集方法を，以下のように提案している．

品詞別最近接：品詞毎にそれぞれ最も近くに共起する単語を共起単語とする［麻野間 2000］

従来手法では，依存関係のない単語対を多くとったり（一文内共起，n単語ウィンドウ），依存単語対を多くとりこぼしたり（単語 bigram）する欠点がある．また“品詞別最近接”は，遠距離の依存単語対を収集できる反面，収集条件の緩さから依存関係にない単語対を依然多く収集する傾向にあった．

## 3 共起品詞パタン

提案手法は，比較的効率のよかった“単語 bigram”の共起取得法をベースとする．これに加え，共起単語それぞれに対する品詞のパタン（共起品詞パタン）をあらかじめ設定して，離散する共起単語を収集すること行う．これにより，“品詞別最近接”によって取得してしまう不適切な単語対の排除が期待できる．さらに“n単語ウィンドウ”と同様の位置的条件を採用して，共起単語間の距離を限定する．

以上より，本稿の“共起品詞パタンを用いる方法”は，次の２つの共起を取得する．

1. 連鎖共起：単語 bigram を収集
2. 離散共起：位置的な単語距離 $d$ 内で共起品詞パタンに合致する単語対を収集

実際に発生する離散共起の品詞の組とその前後関係を，英語コーパスから予備調査したところ，次のような関係が比較的高頻度であることがわかった．

- 名詞とそこから文頭方向で最も近い｛前置詞，冠詞，動詞｝からなる単語対
- ｛前置詞，副詞｝とそこから文頭方向で最も近い動詞からなる単語対

これらの品詞の条件を共起品詞パタンとして採用する．

また，この共起品詞パタンを用いる方法の単語距離制限 $d$ の値は，予備調査において最も性能がよかった $d$=3 をとる．

図 1 の文例から，離散共起は（switched, focus）（switched, quickly）の２つを取得できる．

---

Extracting dependent pairs of words from non-parsed corpora.
Naoki Asanoma and Hiromi Nakaiwa.
NTT Communication Science Laboratories.

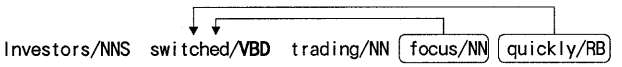


図 1：共起品詞パタンによる共起単語取得

## 4 実験

### 4.1 正解データの作成

依存単語対の正解データは，Penn Treebank 2 (PTB2) の 1989 年 Wall Street Journal コーパスからランダムに選んだ 30 の英文（総単語数：533）に，人手で単語間の依存関係情報（総依存関係数：490）を付与して得る．単語間の依存関係を付与する際には，既に句単位の要素間の解析がなされている PTB2 内のコーパスを用いて，その構造情報に矛盾しないように依存単語対を決定していく．

### 4.2 共起単語収集

実験では，上記で選んだ 30 の英文に対して，前記の共起取得法それぞれを使って共起単語を抽出する．

実験で行う共起取得法は，“一文内共起”，“5 単語ウィンドウ”，“単語 bigram”，“品詞別最近接”，共起品詞パタンを用いる方法，および参考データとしてパーザを用いる方法の 6 つである．パーザは，単語間の依存関係の抽出が可能な Link Grammar Parser[1] を用いた．

なお品詞情報を使う共起取得法については，PTB2 から対応する品詞タグ付のコーパスを用いる．“品詞別最近接”については，収集対象とする品詞を，名詞，動詞，形容詞，副詞に限定する．

### 4.3 評価方法

共起取得法の評価は，収集効率と高速性によって行う．収集効率の評価基準は，再現率 $R$，精度 $P$，および収集効率の総合評価として $R$ と $P$ の要約値 $F$ を用いる．

$$R = \frac{\text{抽出された正解の依存単語対の数}}{\text{正解の依存単語対の数}}$$

$$P = \frac{\text{抽出された正解の依存単語対の数}}{\text{抽出された依存単語対の数}}$$

$$F = \frac{2RP}{R+P}$$

また高速性の評価として，PTB2 の WSJ から選んだ 1000 文（総単語数：15383）から共起取得に要する時間を計る．

[1] http://www.link.cs.cmu.edu/link/

### 4.4 結果

収集効率と高速性の結果を表 1 に示す．所要 $t$ のそれぞれの値は“単語 bigram”の絶対所要時間を 1 とした時の比の値を示している．

表 1：収集実験の結果

| 共起取得法 | 再現率$R$ | 精度$P$ | 要約値$F$ | 所要$t$ |
|---|---|---|---|---|
| 一文内共起 | 1.00 | 0.10 | 0.18 | 6.2 |
| 5 単語ウィンドウ | 0.96 | 0.22 | 0.35 | 2.9 |
| 単語 bigram | 0.64 | 0.63 | 0.63 | 1.0 |
| 品詞別最近接 | 0.35 | 0.67 | 0.45 | 3.5 |
| 共起品詞パタン | 0.77 | 0.59 | 0.66 | 2.0 |
| パーザ | 0.97 | 0.97 | 0.97 | 1200 |

結果として，共起品詞パタンを用いる方法は，“単語 bigram”と比較して総合評価値 $F$ を上回った．また，参考データのパーザを用いる方法は桁違いの解析時間を要するのに対し，共起品詞パタンを用いる方法は比較的高速であった．

この実験結果より，“単語 bigram”に置き換わる方法として，共起品詞パタンを用いる方法の可能性を示すことができた．

## 5 おわりに

本稿は未解析のテキストコーパスから依存単語対を収集するため，収集すべき共起単語の品詞のパタンに基づいた共起取得法を提案し，実験の結果，高速かつ高い収集性能を得ることができた．今後は本手法によって得た共起単語の頻度情報を用いて，機械翻訳の訳語品質向上の検討を進めていく．

## 参考文献

[麻野間 2000] 麻野間直樹，中岩浩巳: 翻訳ルールの意味制約と目的言語共起情報を併用した訳語選択，情処研報 NL-135-22, pp. 165-172, 2000.

[Leacock 98] C. Leacock, M. Chodorow and G. A. Miller: Using corpus statistics and WordNet relations for sense identification, *Computational Linguistics*, vol. 24, no.1, pp. 147-165, 1998.

[Marcus 93] M. P. Marcus, B. Santorini and M. A. Marcinkiewicz: Building a large annotated corpus of English: Penn Treebank, *Computational Linguistics*, vol. 19, no.2, pp. 313-330, 1993.

[Shütze 92] H. Shütze: Dimensions of Meaning, *Proc. of Supercomputing*, pp. 787-796, 1992.

[Smadja 93] F. Smadja: Retrieving collocations from text: Xtract, *Computational Linguistics*, vol. 19, no.1, pp. 143-177, 1993.