教師無し学習による名詞句の言い換え

鳥澤 健太郎
北陸先端科学技術大学院大学 情報科学研究科
〒923-1292 石川県能美郡辰口町旭台 1-1
Email: torisawa@jaist.ac.jp

1 はじめに

本稿では、教師無し学習を用いて、テキストの言い換えを行う手法について述べる。広い意味での言い換えは自然言語理解に対する一手法と考えることができるが、ここでは、「A の B」の形態をもつ名詞句における AB 間に生じる意味的な関係の解析について述べる。従来より、このタスクはルールベースあるいは辞書を使うなどの手法によって試みられてきた。[4] 本研究の特徴は、大量のテキストに統計的教師無し学習である Expectation Maximization 法を適用し、得られた単語の意味クラスを用いてこれらのタスクを行うことにある。

本研究で提案する手法を別の観点から見ると、次の仮説に立脚した方法であるとみなすことができる。

- 二つの単語間で省略されている、あるいは直接的に表現されていない意味的関係は、二つの単語を結ぶ表現で最大確率で現れるものによって表現される。

本研究では、「A の B」の意味的関係の解析においては、二つの単語は A と B、ついで二つを結ぶ表現は二つの単語が現れる動詞句ないしは関係代名詞句とする。つまり、本研究で提案する手法は、以下の例で示されるような入出力関係をもつことになる。

入力 レストランのビール

出力 レストランで飲むビール

ここで、先の仮説に戻るが、この仮説、つまりは表現の頻度が、意味的に望ましいかどうかを判定する尺度になるいう仮説は自明ではなく、さまざまな実験により検証する必要がある。しかしながら、もしこの仮説がある程度の妥当性で成立するのであれば、構文解析、形態素解析などでその有効性が示されている統計的手法が、意味の領域にも適用可能である可能性がでてくる。なお類似の仮説、方法については、村田らの研究がある。[5]

技術的な困難は、コーパスから確率を得る方法にある。単語間の共起を直接カウントしたのでは、現状で利用できるコーパスの量から見て、データスパースネスの問題が生じる。本研究では、統計的な教師無し学習の一手法である Expectation Maximization 法を用いて単語の意味的クラスを生成し、データスパースネスの問題に対処する。この手法の詳細については次節で述べる。

2 EM法による単語クラス並びにクラス間の関係の推定

Expectation Maximization (EM) 法は未知の確率変数を含む確率分布に対する最尤推定の一種であり、繰り返し計算によって確率分布を推定する。本研究で用いる手法は、Rooth, Hofmann らによる単語クラスタリング [6, 2] の手法を拡張したものである。彼等の手法は以下の確率分布を仮定する。

$$P(\langle v, rel, n\rangle) =_{def} \sum_{a\in A} P(\langle v, rel\rangle|a)P(n|a)P(a)$$

ここで n は単語、v は動詞、rel は n が v のどの補語、付加詞になったかを指定するシンボルである。具体的には、subj, obj, あるいは助詞の一つを値としておる。つまり、$P(\langle v, rel, n\rangle)$ は、動詞 v に単語 n が助詞 rel を解して「係っている」確率を表す。Rooth らの手法は、この共起確率をコーパスから学習する手法である。より正確に言えば、彼らの手法で求まるのは、上式の右辺にある $P(w|a)$,$P(\langle v, rel\rangle|a)$ のような確率値である。以下では、これらの確率をパラメータと呼ぶ。単語クラスは、このパラメータの推定によって得られる。式にある a は単語クラスの identifier であるシンボルであり、本研究ではこれをクラスシンボルと呼ぶ。上式ではクラスシンボルは集合 A の要素となっているが、分類すべき単語クラスの数はこの集合の濃度によってきまる。なお、この集合は前もって人手で与える必要がある。

さて、パラメータの推定と単語クラスの関係であ

$$P_j(a|\langle v, rel, n\rangle) = \frac{P_j(a)P_j(\langle v, rel\rangle|a)P_j(n|a)}{\sum_{a' \in A} P_j(a')P_j(\langle v, rel\rangle|a')P_j(n|a')}$$

$$P_j(a, b|\langle v, rel_1, rel_2, n_1, n_2\rangle) = \frac{P_j(\langle v_i, rel_1, rel_2\rangle|a, b)P_j(n_1|a)P_j(n_2|b)P_j(a, b)}{\sum_{a', b' \in A} P_j(\langle v_i, rel_1, rel_2\rangle|a', b')P_j(n_1|a')P_j(n_2|b')P_j(a', b')}$$

$$P_{j+1}(a) = \frac{1}{Z_{class}} \sum_{\langle v_i, rel_i, n_i\rangle \in L_1} P_j(a|\langle v_i, rel_i, n_i\rangle),$$

$$P_{j+1}(a, b) = \frac{1}{Z_{joint}} \sum_{\langle v_i, rel_i^1, rel_i^2, n_i^1, n_i^2\rangle \in L_2} P_j(a, b|\langle v_i, rel_i^1, rel_i^2, n_i^1, n_i^2\rangle)$$

$$P_{j+1}(n|a) = \frac{1}{Z_N} \left\{ \sum_{\langle v_i, rel_i, n\rangle \in L_1} P_j(a|\langle v_i, rel_i\rangle, n) + \sum_{\langle v_i, rel_i^1, n, rel_i^2, n_i\rangle \in L_2, b \in A} P_j(a, b|\langle v_i, rel_i^1, rel_i^2, n, n_i\rangle) + \sum_{\langle v_i, rel_i^1, rel_i^2, n_i, n\rangle \in L_2, a' \in A} P_j(a', a|\langle v_i, rel_i^1, n_i, rel_i^2, n\rangle) \right\},$$

$$P_{j+1}(\langle v, rel\rangle|a) = \frac{1}{Z_{VP1}} \sum_{\langle v, rel, n_i\rangle \in L_1} P_j(a|\langle v, rel, n_i\rangle),$$

$$P_{j+1}(\langle v, rel^1, rel^2\rangle|a, b) = \frac{1}{Z_{VP2}} \sum_{\langle v, rel^1, rel^2, , n_i^1, n_i^2\rangle \in L_2} P_j(a, b|\langle v, rel^1, rel^2, n_i^1, n_i^2, \rangle)$$

Z_{class}, Z_{VP1}, Z_N, Z_{joint} 及び Z_{VP2} は、各々のパラメータの総和が 1 になるための正規化項である。

図 1: EM 法のための漸化式.

るが、パラメータ $P(w|a)$ にベイズの定理を適用すると $P(a|w)$ という確率を計算することができる。この確率は、単語 w が出現したときに、その用法がクラス a に属する確率と解釈できる。単語のクラスはこのような確率によって表現することができる。例えば、$P(a|$ ビール$)$ が大きな値を持つクラス a にたいして、「酒」が「ビール」と類似した意味を持つ単語であるとするならば、$P(a|$ 酒$)$ も同様に大きな値を持つということである。実際に日本語の新聞記事にこの手法を適用すると、意味を反映した単語クラスを得ることができることが実験の結果分かっている。また、この手法は単語の持つ多義性を考慮できるソフトクラスタリングとなっている。例えば、「磐田」という語は、都市名であると同時に、ジュビロ磐田というサッカーチームを指すために使われるが、実験の結果はそのような事実を反映した確率分布が得られている。

本研究で使用する手法は、以上の手法を拡張したものである。より具体的には、先に述べた確率分布に加えて次の確率分布を仮定する。

$$P(\langle v, rel_1, rel_2, n_1, n_2\rangle) =_{def} \sum_{a,b \in A} P(\langle v, rel_1, rel_2\rangle|a, b)P(n_1|a)P(n_2|b)P(a, b)$$

この拡張の意味は、Rooth らの手法が基本的に動詞一つと単語一つのペアを扱っていたのに対して、動詞一つと単語二つの三つ組みの関係を確率分布で捕らえることである。上の式にあらわれるクラスシンボル a と b はそれぞれ、単語 n_1、n_2 の単語クラスに対応するクラスシンボルである。また、$\langle v, rel_1, rel_2\rangle$ は、二つの空いたスロットを持つ動詞句のテンプレートと考えることができる。$P(\langle v, rel_1, rel_2, n_1, n_2\rangle)$ は、単語 n_1 が助詞 rel_1 を介して動詞 v に係り、単語 n_2 が助詞 rel_2 を介して動詞 v に係っている確率を表す。つまり、動詞句テンプレート $\langle v, rel_1, rel_2\rangle$ の二つのスロットが単語 n_1 と n_2 によって埋められている確率である。このような拡張によって、二つの単語を含む動詞句の出現確率の推定が可能となる。

さて、パラメータの推定手法に話を移す。重要な点は、確率分布にあらわれるクラスシンボル a,b がコーパス中では観測されないということである。本手法では、以下にある二つの観測された共起データの列（つまりはコーパスから抽出された学習データ）の出現確率が上で定義した確率分布に従って、最大になるようにパラメーターを調整する。（実際には、局所解の問題があり、必ずしも最大にはならない。）

$$L_1 = \langle\langle v_0, rel_0, n_0\rangle, \langle v_1, rel_1, n_1\rangle, \cdots, \langle v_m, rel_m, n_m\rangle\rangle$$

$$L_2 = \langle\langle v_0, rel_0^1, n_0^1, rel_0^2, n_0^2\rangle, \cdots, \langle v_k, rel_k^1, n_k^1, rel_k^2, n_k^2\rangle\rangle$$

これらの学習データの出現確率は以下の式で与えられる。

$$\prod_{\langle v_i, rel_i, n_i\rangle \in L_1} P(\langle v_i, rel_i, n_i\rangle) \times \prod_{\langle v_i, rel_i^1, rel_i^2, n_i^1, n_i^2\rangle \in L_2} P(\langle v_i, rel_i^1, rel_i^2, n_i^1, n_i^2\rangle)$$

実際のパラメータの推定は、次のように行われる。まず、パラメータの初期推定値を決定する。この推定値はランダムに決めても良いが、本研究では平均相互情報量に基づいた単語クラスタリング手法 [1] に基づいて、とりあえず、単語の属するクラスを推定し、これをもとに初期推定値を決定する。（このクラスタリング手法では、各単語はただ一つのクラスに「強制的」に属させられ、単語の持つ多義性は考慮されない。）例えば、平均相互情報量に基づくクラスタリングによって、単語 w がクラス a に属すとされた場合には、$P(w|a)$ が高くなるように、パラメータを設定する。以下ではこのように決めたパラメータをそれぞれ、$P_0(w|a)$,$P_0(\langle v, rel\rangle|a)$ などとあらわす。ついで、図 1 にある漸化式に従って、パラメータ（$P_j(dot)$ であらわされる）を順次更新していく。この漸化式は、既に仮定した確率分布に EM 法の標準的な手続きを適用することで得られる。この更新の繰り返しが一定の回数に達したところで、得られた各パラメータの値をもとに必要な確率分布を計算する。

- CLASS 1
  薬 0.628 新薬 0.631
  抗生物質 0.613 剤 0.523
- CLASS 2
  支援 0.692 バックアップ 0.644
  後押し 0.623 救援 0.619
- CLASS 3
  心 0.674 内面 0.169
  闇 0.157 心身 0.094
- CLASS 4
  勧業 0.813 日興 0.804
  大和 0.740 住友 0.733
- CLASS 5
  トヨタ 0.711 ダイハツ 0.663
  日産 0.647 本田 0.632

図 2: 単語クラスの例

以上が本研究で提案する EM ベースの手法の概要であるが、実際にはまだ問題がある。我々の言い換えタスクではかなり木目の細かい意味クラスが要求される。クラスタリングにおける意味クラスの数は、前もって指定する必要があるが、クラス数を 500 とすると、単語クラスが比較的大雑把なものとなり、例えば、「飲み物」と「食べ物」が一つの単語クラスに集まってしまうといった結果が得られる。この結果、「レストランのビール」の解析結果として「レストランで食べるビール」という望ましくない出力が出てしまう。こういった問題を回避するためには、クラス数を増加させればよいのだが、推定するパラメータ数がクラス数の二乗に比例するため、計算に要する時間、メモリが非現実的なものとなる。本研究ではこの問題に対処するため EM 法の近似手法を開発し、クラス数が増加しても計算が可能となるようにした。また、クラス数が増加するとデータスパースネスの影響が現れるため、これを避けるために、上述の確率モデルにさらなる拡張を加えた。

前述の確率モデルにおいて、学習に先だって与えられたクラスシンボルの集合 A の要素を含むパラメータ全ての集合を一つの単語分類と定義する。新たに定義された確率モデルでは、二つの互いに疎なクラスシンボルの集合を学習に先だって与え、二つの独立な単語分類を同時に計算する。動詞句テンプレートによって表現される単語クラス間の関係は、異なるクラスシンボルの集合に対応する単語クラス間の関係として捕らえられる。これにより、例えば、一つのクラスシンボルの集合としては、500 個のシンボルからなるものを与え、もう一つのクラスシンボルの集合として、2500 個という大量のシンボルを設定し、比較的おおざっぱな単語クラスと細かい単語クラスの間の関係を計算することができるようになる。この場合、必要なパラメータの数は、二つのクラスシンボルの集合の濃度の積となり、細かい単語クラスを考慮しつつ、パラメータ数を抑えることが可能となる。これ以上の詳細についてはここでは述べないが、これによりデータスパースネスの影響が緩和され、「A の B」の解釈でより良いパフォーマンスが得られた。

3 実験

以上に述べた EM 法をベースとする手法により、新聞記事 14 年分を既存の構文解析器 [3] をつかってパーズし、その結果から確率分布の推定を行った。学習の対象としたのは、単語が 18,360 語、助詞と動詞の組が 25,473 個、及び動詞テンプレート 19,704 個であり、いずれも閾値を上回る高頻度を持つものである。クラス数が 2500 の場合の得られた単語クラスの一部を図 2 に示す。図中の確率値は確率 $P(a|w)$ である。

「A の B」の言い替えの実験では、前節で定義した確率 $P(\langle v, rel_1, rel_2, n_1, n_2\rangle)$ を最大とする動詞句を A と B の間の意味的関係をあらわすものととして出力させ、その結果を人手で評価した。[1]

実験に使われた名詞句は、学習データに含まれない新聞記事中の名詞句 300 個である。この内、学習時に考慮されなかった低頻度語を含む名詞句が 55 例あり、言い換えを生成することができたのは、残りの 245 個であった。結果は表 1 にある。表中の精度は言い換えを生成することができた名詞句に対する妥当なものの割合である。表中で Model とあるのは、クラスの数であるが、2500,500 とあるのは、前述の拡張を加えた確率モデルを用いたもので、それぞれ 2500 個と 500 個の単語クラスからなる二つの独立な単語分類をもちいた場合の性能である。Top は最大の確率を持つ動詞句が A と B の関係として妥当であるケース、Top 5 は確率の上位 5 個が、妥当な関係を表す動詞句を含んでいるケースを表す。

この結果から、クラス数が 2500 になるとデータスパースネスの影響が強くなり、また、複数の単語分類を使用するモデルは単一の単語分類のを利用するモデルよりも、良い結果が出ていることが分かる。ここで、複数の単語分類を使用するモデルで得られた正解の例を図 3 に示す。正解については判断の微妙なものもあるが、比較的多くの状況で成立するような関係であれば正解とした。また、表の最後にある「2500,500

[1] 我々の先行研究 [7] では、純粋な確率ではなく ad-hoc なスコアを用いたが、EM 法の近似の精度をあげたところ、純粋な確率でもほぼ同等の精度が得られた。

Model	*Best*	*Top* 5
500	89 (36.2%)	150 (61.2%)
2500	74 (30.2%)	132 (53.9%)
2500,500	102 (41.6%)	170 (69.3%)
2500,500 バイアス	340 (67.7%)	465 (92.6%)

表 1: 言い換えの精度

名詞句	言い換え後の 動詞句/関係代名詞句
周りの人々	周りにいる人々
得票の模様	得票を獲得する模様
出馬の挨拶	出馬を表明する挨拶
１日の談話	１日に発表する談話
ゆかりの県下	ゆかりがある県下
学校の友達	学校に通う友達
本部のＸ氏	本部に勤務するＸ氏
全国の河口堰	全国にある河口堰

図 3: 言い替えの例

バイアス」は、「A の B」で頻繁に表される関係に対応している動詞句、例えば、「A が B を持つ」「A が B にある」などを 28 個選び出し、それらに対してバイアスを人手で与え、2500,500 のモデルで得られる確率にそれらバイアスを掛けた結果得られたものである。この方法で、25%程度の精度の向上を見た。実験では、これまでの実験とは異なるランダムに選ばれた 600 個の名詞句を使用した。そのうち、未知語を含む名詞句は 98 個で、502 個の名詞句に対して言い換えが生成された。

4 まとめ

本稿では自然言語理解に対する一手法と考えることができる言い換えをコーパスからの教師無し学習によって行う手法について述べた。より具体的には、「A の B」という形態の名詞句において、A と B の間に成立する意味的関係の推定について述べた。今回提案した手法は注釈付きコーパスを必要としないため、より大量のコーパスに容易に適用可能である。WWW 上のテキストなどに適用することにより、より高い精度を実現することができるのではと期待している。また、得られた確率分布は格の解析などにも利用することができ [8]、現在は動詞句の間の意味的類似を示す尺度として利用する研究を勧めている。動詞句の間の意味的類似を定義することは難しいが、「石油をサウジから輸入する」と「石油をサウジから購入する」が類似である、あるいは「レストランでビールを飲む」と「レストランがビールを出す」の間に強い類似性があることなどが現在のところ計算可能である。

参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):31–40, 1992.

[2] Thomas Hofmann and Jan Puzicha. Unsupervised learning from dyadic data. Tr-98-042, International Computer Science Institute, Berkley, CA, 1998.

[3] Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. A hybrid japanese parser with hand-crafted grammar and statistics. In *Proceedings of COLING 2000*, pages 411–417, 2000.

[4] Sadao Kurohashi and Yasuyuki Sakai. Semantic analysis of japanese noun phrases : A new approach to dictionary-based understanding. In *Proceedings of 37th Annual Meeting of the ACL*, pages 481–488, 1999.

[5] Masaki Murata and Hitoshi Isahara. Universal model for paraphrasing - using transformation based on a defined criteria. In *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications*, Tokyo, Japan, 2001.

[6] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of 37th Annual Meeting of the ACL*, 1999.

[7] Kentaro Torisawa. A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases. In *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications*, Tokyo, Japan, 2001.

[8] Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of 6th Natural Language Processing Pacific Rim Symposium*, Tokyo, Japan, 2001.