日本語ニュースからの定型文の自動抽出

1 Q-1

加藤 直人 江原 暉将

NHK 放送技術研究所

{katonao, eharate}@strl.nhk.or.jp

1 はじめに

日本語ニュースの中で、経済ニュースの文 (経済ニュース文) は、特殊な文型をしており、また、翻訳する場合には経済独特の訳が要求される。一方で、経済ニュース文は定型パターンで多くカバーされる文 (定型文) である場合も多く、このような定型文のみを収集し、処理したほうが有効である。

本稿では、定型文を自動抽出する手法について述べる。提案する手法では、定型パターンに単語 n グラムを用い、各文に対して単語 n グラムの列を求め、その類似度に基づいてクラスタリングする。類似した定型文は 1 つのクラスにまとめられ、それぞれのクラスは同じ種類の定型文であるとして、自動抽出される。

2 定型パターンの含有率に基づく定型文

我々は定型文を定型パターンで大部分被覆された文と定義し、図 1 に示した手順 (アルゴリズム 1) によって求めた [加藤 95]。

STEP1 では、形態素解析した後、連続して出現する単語列 (単語 n グラム) の頻度をまず求める。そして、ある頻度以上出現した単語 n グラムを集めて単語 n グラム辞書を作成する。

STEP2 では、単語 n グラム辞書を引き、1 文中に含まれる単語 n グラムが最大何% 含まれるか (含有率 r) を各文 s に対して計算する。含有率 r は次式のように定義した。詳細は [加藤 95] を参照のこと。

$$\text{含有率}(r) = \frac{\bar{x} \text{ 中の文字数}}{\text{文 } s \text{ 中の文字数}} \times 100 \quad (1)$$

$$\text{ただし,} \quad \bar{x} = \operatorname{argmax}_x \sum_x \text{Point}(x^i) \quad (2)$$

$$X = \{(x^1, \dots, x^i, \dots, x^m) \mid x^i \text{ は文 } s \text{ に含まれ, 他と重ならない単語 } n \text{ グラムの番号}\}$$

$$\text{Point}(x^i) = 3^n(x^i) \quad (3)$$

($n(x^i)$: 単語 n グラム x^i における n の値)

上式では文 s になるべく多くの単語 n グラムで被覆

An automatic method for extracting fixed sentences from Japanese news corpora.
Naoto Katoh and Terumasa Ehara
NHK Science and Technical Research Laboratories

STEP1 単語 n グラム辞書の作成

STEP2 含有率を計算

STEP3 含有率がしきい値 (r_0) 以上であればその文を出力

図 1 定型文を抽出する手順 (アルゴリズム 1)

し、また、より長い単語 n グラム (すなわち、より大きい n) が採用されるように定義している。

STEP3 では、あらかじめしきい値 (r_0) を決めておき、コーパスの各文に対して含有率を計算し、 r_0 以上の文を定型文として抽出する。

以上の手法に基づいて、実際に日本語経済ニュースから単語 n グラム辞書を作成し、定型文を抽出した結果の例を図 2 に示す。ただし、日本語経済ニュースとして '99, '00 年度の NHK ニュース原稿 (約 200 万文) を、単語 n グラム辞書は出現頻度 20 回以上の単語 3 グラム (約 1 万 7 千個) を使った。

得られた定型文は図 2 のように、類似した文が多数出現していた。このような類似した文を 1 つのクラスにまとめておくことができれば、共通した特徴などを発見しやすい。そこで、次に述べる方法で定型文のクラスタリングを行う。

27:2306:きょうの/東京
84:1184:市場/、主要
441: 417:銘柄/の/平均

(a) 単語 n グラム辞書の例

(ただし、辞書中の番号: 出現頻度: 単語 n グラム)

1.00:27,441,84,...:きょうの東京市場、主要銘柄の平均株価、午前の終値は、きのうより 1 3 8 円 9 6 銭高い 1 万 7 8 1 6 円 5 2 銭でした。
1.00:27,441,84,...:きょうの東京市場、主要銘柄の平均株価、午前の終値は、きのうより 2 8 6 円 5 5 銭安い 1 万 7 6 2 2 円 7 4 銭でした。
1.00:27,441,84,...:きょうの東京市場、主要銘柄の平均株価、午前の終値は、きのうより 4 3 円 7 3 銭安い 1 万 7 7 1 2 円 7 8 銭でした。
1.00:27,441,84,...:きょうの東京市場、主要銘柄の平均株価、午前の終値は、きのうより 1 5 0 円 7 5 銭安い 1 万 7 6 5 1 円 7 3 銭でした。
1.00:27,441,84,...:きょうの東京市場、主要銘柄の平均株価、午前の終値は先週末より 1 7 2 円 4 4 銭高い 1 万 7 8 0 2 円 4 3 銭でした。

(b) 定型文の例

(ただし、含有率: 単語 n グラムの列: 定型文)

図 2 単語 n グラム辞書と定型文の例

3 クラスタリングによる定型文の自動抽出

定型文をクラスタリングするには、図1のアルゴリズムでは単語 n グラムの番号の列を使って類似度を計算する。すなわち、ある2つの定型文が類似しているか否かは、それぞれの定型文を抽出する際に用いた単語 n グラムの類似度で比較する。類似度には単語 n グラムの番号の一致とともにその順序も考慮している。

定型文をクラスタリングする手順（アルゴリズム2）を図3に示す。STEP1とSTEP2は旧アルゴリズム（図1）と共通である。新アルゴリズムでは、クラスタリングの計算量を減らすために、基本的には[Brown 92]と同様に、あらかじめクラス数 C を決めておき、そのクラスに1つ文を足して、その中で最も類似度が大きいクラス同士を1つにまとめている。

STEP3'では、定型文を抽出するとともに、その抽出に用いられた単語 n グラムの列を保存しておく。

STEP4では、STEP3'で得られた定型文を集めて定型文の集合 S を作り、その中から $C+1$ 個の定型文を取り出して、これらを1つの初期クラスとする。そして、互いのクラス（定型文）それぞれの類似度を計算する。ここで、クラス s_i （単語 n グラム $\bar{x}_i$ ，単語 n グラムの番号の生起回数 f_i ）クラス s_j （同様に $\bar{x}_j$ ， f_j ）の文類似度は $Sim(s_i, s_j)$ は次式で定義する。

$$Sim(s_i, s_j) = \frac{\sum_{x_i^r = x_j^s} (f_i^r + f_j^s)}{\sum_{r=1}^p f_i^r + \sum_{s=1}^q f_j^s} \quad (4)$$

ここで、

$$\bar{x}_i = (x_i^1, x_i^2, \dots, x_i^p), f_i = (f_i^1, f_i^2, \dots, f_i^p)$$

$$\bar{x}_j = (x_j^1, x_j^2, \dots, x_j^q), f_j = (f_j^1, f_j^2, \dots, f_j^q)$$

例えば、

$$\bar{x}_1 = (27, 441, 56, 78, 84), f_1 = (1, 1, 1, 1)$$

$$\bar{x}_2 = (27, 441, 78, 84), f_2 = (2, 2, 1, 2)$$

の場合には、

$$Sim(s_1, s_2) = \frac{(1+2) + (1+2) + (1+2)}{(1+1+1+1) + (2+2+1+2)}$$

となる。単語 n グラムの回数は、初期値（すなわち、1つの定型文からなるクラスで）はすべての値は1である。

STEP3' 含有率がしきい値 (r_0) 以上であれば
その文とその単語 n グラムの列を出力

STEP4 定型文の集合 S 中から $C+1$ 個の文を取出し、互いの類似度 $Sim(s_i, s_j)$ を計算

STEP5 類似度が最大となるクラス同士をマージする。

STEP6 定型文の集合 S から1文を取出し、1つのクラスとし、STEP5で得られた C 個のクラスとの類似度を計算し、STEP5へ。

STEP7 定型文の集合 S が空になるまで、STEP5～STEP6をくり返す。

図3 クラスタリングによる定型文の自動抽出
(アルゴリズム2)

STEP5では、類似度が最大となるクラス同士をマージする。マージ後のクラスの単語 n グラムはマージ前のクラスを合わせたものとして求める。例えば、先のクラス s_1 と s_2 をマージしたクラス s_3 では、

$$\bar{x}_3 = (27, 441, 56, 78, 84), f_3 = (3, 3, 1, 1, 3)$$

となる。

STEP6では、定型文の集合 S から新たに定型文を1つ取出し、これを新しいクラスとし、STEP5で求められた C 個のクラスとの文類似度を計算する。そしてSTEP5に戻る。

STEP7では、定型文の集合 S が空になるまでSTEP5とSTEP6をくり返す。

以上の手順によって類似した定型文が1つのクラスにまとめらる。

4 おわりに

定型パターンの類似度により定型文をクラスタリングし、定型文を自動抽出する手法について述べた。今後は本手法を使って定型文を自動抽出し、日本語経済ニュースの文の機械翻訳精度の向上などをはかりたい。

参考文献

- [Brown 92] Brown, P.F. et al., "Class-based n-gram models of natural language," Computational Linguistics, Vol.18, No.4, pp.467-479, 1992.
- [加藤 95] 加藤直人, "定型パターンを含む文の機械翻訳手法," 情報処理学会論文誌, Vol.36, No.9, pp.2081-2090, 1995.