重要複合語の自動判別方式の評価

類似例尺度を重要複合語収集というタスクに応用し、その有用性を検証した。類似例尺度がある閾値を越えるものを探索対象から除くというヒューリスティクスを導入することにより、作業がどの程度効率化され、また副作用としてどの程度の重要複合語のロスが見込まれるかを実験で評価した。

1　実験方法

今回の実験には日経新聞 (CD-ROM 版) ９７年 [4] の約２０万記事のタイトルと本文を利用し、形態素解析器 ANIMA [3] を施して単語＋品詞の列に分解し、品詞が名詞または未知語で構成されるすべてのバイグラムとその頻度をカウントした。形態素総数は４４M 個、(名詞／未知語)バイグラムの総数は 6.9M 個、種類数は 965,923 種類であった。

［第１段階 : 対数尤度による順位付け］最初に対象バイグラムを簡略化版の対数尤度をスコアとして順位付けを行った。その定義は以下で与えられる。

$$LL(AB) = fr(AB) \cdot \log \frac{fr(AB) \cdot N}{fr(A) \cdot fr(B)}$$

主尺度として対数尤度（の簡略版）を用いることにした理由は Daille らによる比較実験で相互情報量に対する優位性が示されていたからである [1, 2](1994)。上位１０万個を取り、それらに対して類似例尺度を計算した。計算に要した時間は Compaq Alpha server (300MHz) で３時間弱であった。

［ヒューリスティクスによるふるい落し］次に以下に示すようなヒューリスティクスによりふるい落しを行い、さらに両方の構成語が２文字以上という条件を付けて残ったものの上位１０００個（対数尤度による順位）を実験対象として用いた。
構成語が２文字以上という条件を付けた理由は、１文字語を含む場合には形態素解析誤りの場合が多数含まれていて今回の実験の目的には適さないと判断したためである。

否定性因子としては、強汎用語因子（左右２種類)、部分性因子（左右２種類)、分解表現への換言性因子（補う助詞の種類で５種類）の合計９種類を用いた。以下各々の否定性因子について説明する。本実験とは直接には関係しないので詳しい説明を省略する。

2　予備実験

初めに予備実験として類似例尺度が陽性のものと陰性のものを取り、それらが主観的な重要性判断とどのように関係するかを調べた。

類似例因子の陰陽の判定基準となる閾値を 0.25 に設定した。その場合、上記のようにして選んだ１０００個のバイグラムには類似例因子が陽性となるもの（類似例尺度 が 0.25 以上）が 159 種類、また陰性のものが 841 種類ということになる。

表 1. 類似例尺度の閾値を0.27とし、それを越える（類似例因子が＋）のバイグラムとそれ以下のもの各３０個に重要度の主観判断を下し、主観判断別にまとめた結果

主観判断	類似例因子	A-B	類似例尺度	対数尤度
+1	-	付加-価値	0.11	10001
+1	-	医療-保険	0.24	8520
+1	-	自己-破産	0.13	4865
+1	-	数値-目標	0.19	4364
+1	-	衛星-放送	0.21	4310
+1	-	カラ-出張	0.12	3594
+1	-	在宅-介護	0.16	2968
+1	-	使い-勝手	0.11	2133
+1	-	コスモ-石油	0.13	1237
+1	-	所得-隠し	0.13	944
+1	+	第一-勧銀	0.29	883
-1	-	クリントン-大統領	0.24	6469
-1	-	全会-一致	0.03	2096
-1	-	和平-交渉	0.23	1309
-1	-	特例-措置	0.19	1141
-1	-	シアヌーク-国王	0.08	1114
-1	-	休業-日数	0.19	1003
-1	+	医療-機関	0.31	5566
-1	+	同日-午前	0.29	4209
-1	+	ポイント-改善	0.29	2422
-1	+	処理-能力	0.31	2079
-1	+	メーカー-各社	0.30	1972
-1	+	ドル-前後	0.37	1701
-1	+	財政-状況	0.28	1458
-1	+	流通-業界	0.38	1352
-1	+	前期-推定	0.27	1278
-1	+	推進-会議	0.34	1092
-1	+	除去-作業	0.28	1017
-1	+	チェルノムイルジン-首相	0.30	953
0	-	信用-取引	0.19	2472
0	-	経常-収益	0.22	2128
0	-	水道-メーター	0.11	1851
0	-	商品-ファンド	0.19	1691
0	-	財政-支出	0.23	1539
0	-	実効-税率	0.16	1446
0	-	航空-貨物	0.23	1431
0	-	宇宙-飛行士	0.07	1260
0	-	パネル-討論	0.14	1120
0	-	下値-不安	0.13	1119
0	-	中華-料理	0.15	1119
0	-	緊急-融資	0.23	1094
0	-	土砂-崩れ	0.05	893
0	+	増収-増益	0.30	7050
0	+	為替-相場	0.28	5469
0	+	部品-メーカー	0.32	5254
0	+	衆院-本会議	0.30	2260
0	+	金融-収支	0.27	1847
0	+	電機-メーカー	0.27	1815
0	+	債券-先物	0.31	1684
0	+	家電-メーカー	0.32	1665
0	+	当期-利益	0.36	1600
0	+	回収-作業	0.31	1532
0	+	給与-振り込み	0.30	1342
0	+	行革-本部	0.29	1161
0	+	調整-局面	0.28	1021
0	+	株式-含み益	0.31	919
0	+	買い-意欲	0.31	900
0	+	自民党-執行部	0.30	894
0	+	税率-引き下げ	0.36	887
e	-	共産-党大会	0.21	1357

陰陽各３０個ずつをランダムに取り、それらがどちらであるか分からないように混ぜ合わせて並べ、筆者の判断で重要複合語として取りたいか、捨てたいか、どちらでもよいかで +1, -1, 0 を降った。少しでも迷った場合はすべて０としたので０の数が約半数になった。

表1は主観判断別に結果をまとめたものである。第２コラムの＋／－は類似例因子の陰陽を示す。

主観的に重要と判定したもの１１個の内、１０個は類似例因子が陰性であり、類似例因子がプラスで重要と判定されたのは「第一/勧銀」一つであった。この結果は類似例因子がプラスのものを捨ててしまってもロスが少ないこと、すなわち類似例因子を否定的なフィルターとして利用できそうなことを示している。

3　作業量削減効果と重要語損失率の関係

類似例因子の閾値を小さく設定すれば、それだけ類似例因子がポジティブになりやすくなり、探索対象から除かれるバイグラムの個数が大きくなる。またその分副作用として多くの重要な複合語が失われる可能性もある。

図1 は類似例因子の閾値θを変化させた場合の作業量削減率と重要複合語損失率の推移を示したグラフである。

このグラフで重要なことはグラフの最初の部分がフラットになっていることである。すなわちある程度までは作業量削減しても損失が非常に軽微に押えられることをしめしている。 例えばこの場合のフラットになっている右端は閾値を0.23に設定した場合で、３４％の削減率で損失が０％であることを示している。しかしもう少し安全に見積もって閾値を0.25に設定した場合でも、２６％の削減率を達成している。

4　結言

4.1　成果のまとめ

コーパスから抽出した多数の日本語単語バイグラムを対象とし、統計的なスコアによる順位付けの後、さらに重要な複合語の濃度を高める手段として、種々のヒューリスティクスに加えて顕著な類似例のある複合語は平凡である可能性が高いので除くという方法を提案した

新聞から抽出した単語バイグラムのセットを対象とした実験によりその効果を確認した。統計的スコア上位１０００位を対象とした場合、重要複合語をほとんど損失することなく２５～３０％の作業量を削減できた。

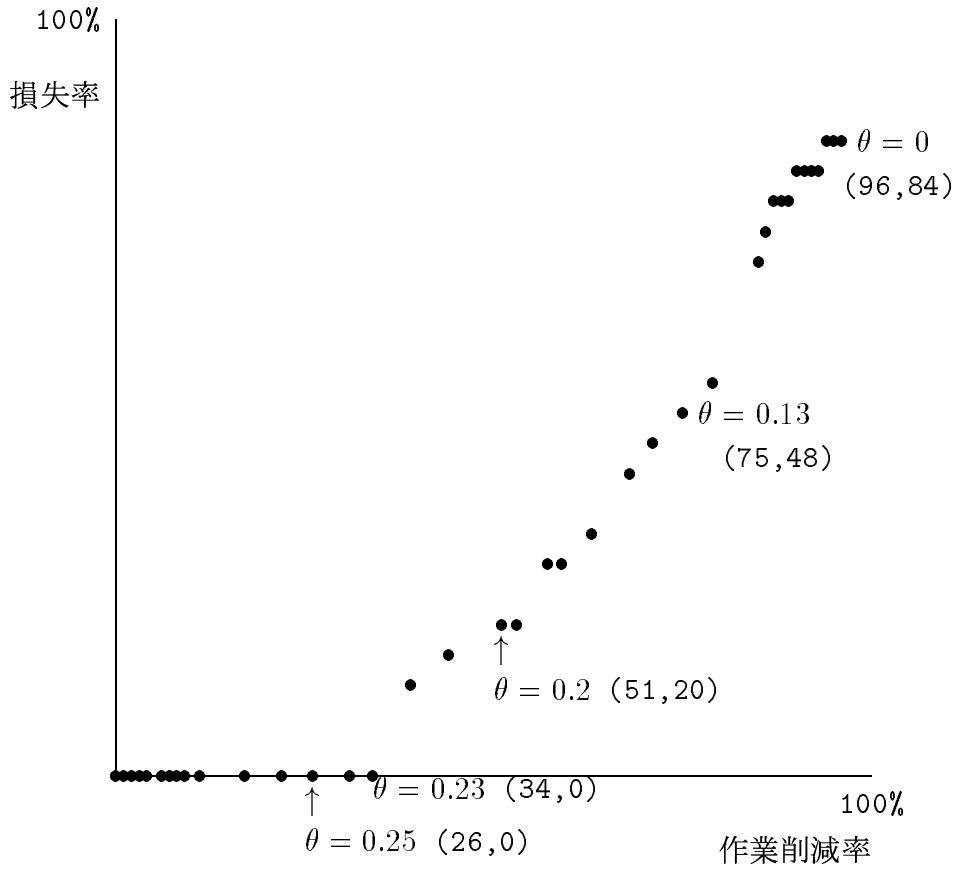

図 1. 類似例因子の閾値(θ)を変えた時の作業量削減効果と重要複合語損失率の推移

4.2 今後の課題

本提案手法は単語類似度の測り方に大きく依存する。従ってその最適化が課題となる。本実験ではバイグラムの右側にくる語としての類似度を計算する場合には左側に伴う語の分布のみを用いた。
しかしながら純粋に単語間の類似度を測る目的であれば、左右片方より両方用いた方が（さらには遠距離の共起関係なども用いた方が）良い結果がでるはずであろう。
ただし、本研究の目的である重要な複合語を抽出するという目的に用いる場合には、どちらが良い結果になるかは明らかとは思われず、実験的に確認する必要がある。

共起分布とシソーラスベースの類似度を用いた場合の比較も興味深い。後者はネットワークタイプの語彙データベース上での単語間距離に基づく類似度であるが、その場合未登録語の影響が不利に働くのではないかと予想される。

5 参考文献

[1] Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. In *The Balancing Act (Workshop at the 32nd Annual Meeting of the ACL)*, pp. 29–36, New Mexico, 1994.

[2] Béatrice Daille, Éric Gaussier, and Jean-Marc Langé. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of COLING'94*, pp. 515–521, Kyoto, 1994.

[3] Hirofumi Sakurai and Toru Hisamitsu. A data structure for fast lookup of grammatically connectable word pairs in japanese morphological analysis. In *Proceedings of ICCPOL'99*, pp. 467–471, 1999.

[4] 日本経済新聞社. 日本経済新聞ＣＤ－ＲＯＭ１９９７年版, 1998.