情 報 学 基 礎　68－9
自 然 言 語 処 理 151－9
（２００２．９．１８）

類義語のオンライン検索

伊東 秀夫
(株) リコー　ソフトウェア研究所
hideo@ic.rdc.ricoh.co.jp

与えられた任意の語Qおよび文脈Cの組みに対し，文脈Cにおける語Qの類義語群をオンラインでランキング検索するための手法を提案する。オフラインでのシソーラス構築とは異なり、予め対象とする語群を収集する必要がなく、また、文脈を限定した上で類義語を抽出するので、コーパスワイドな文脈に基づく従来手法に比べ、良い精度が得られる可能性がある。NTCIR-3 の特許タスクで用いられたデータを用いた類義語検索の実験結果を示す。

On-line Retrieval of Synonyms

Hideo Itoh
RICOH COMPANY, LTD.
Software Research Center
hideo@ic.rdc.ricoh.co.jp

We present a method for on-line retrieval of synonyms. This method enable us to retrieve synonyms of the target term in the context specified by the user. Moreover the method is free from vocabulary setting which is required in off-line methods. We show experimental results of synonym retrieval using NTCIR-3 patent retrieval test collection.

1　はじめに

自然言語において語と語の関係は、共出関係と選択関係に２分できる。一般にシソーラスでは前者は関連語として表現され、後者は広義語、狭義語、同義語などで表現される。

本稿では、語と選択関係にある語を、その語の類義語と呼ぶ。そして、与えられた任意の語Qおよび文脈Cの組に対し文脈Cにおける語Qの類義語群をオンラインでランキング検索するための手法を提案する。

類義関係を幅広く正確に捉えることは、意味レベルの自然言語処理を構築する際の基礎となる。

2　従来技術

シソーラス構築として、関連語と共に、類義語をコーパスから自動収集する研究は数多くある。山本らの研究 [1] は、オフラインでのシソーラス構築ではあるが、語用論的な類義語を収集する点で本研究に近い。

Jing らは、語毎の出現文脈を擬似文書として文書検索システムに登録しておき、文書検索のクエリ拡張に利用している [2]。質問語と直接共起する語が得られる点で関連語のオンライン検索に相当する。また語の出現文脈はコーパス全体から収集する点で本研究と異なる。

Schutze らは、類義関係にある語が同一文書に共起しにくいことに着目し、類義語を得る方法として、直接共起ではなく、間接共起に基づく類義語抽出方法を提案している [3]。すなわち語彙毎にその語と直接共起する語群を出現文脈としてコーパスから収集する。そして、それら出現文脈の類似性を基に語間の類義度を得てランキングする。しかし、出現文脈をコーパス全体からオフラインで構築しておく点で本研究とは異なる。

Xu らは質問に対する文書検索で上位にランクされた文書群を、その質問の局所文脈として捉え、その文書群を解析することで関連語を得ている [4]。本研究は、局所文脈解析 (Local Context Analysis) を類義語検索に発展させたものと言える。

3　類義語の検索

類義語のオンライン検索問題を以下に定義する。予め文書集合Uが与えられているものとする。このときユーザが与えた質問語Qおよび文脈Cに対し、文脈Cにおける語Qの類義語群Sを文書集合Uからランキング検索する。ただし文脈Cは自然言語テキストで表現されるものとする。

例えば語Qとして “小泉” を与え、文脈Cとして “日本の総理大臣” を与えた場合、新聞記事の集合から “森”, “小渕” などを類義語としてランキング検索する問題である。

この類義語検索を実現するための基本的アイデアを図 1 を用いて説明する。

図は文書集合に関するベン図である。外側の四角は、予め与えられた文書集合U、すなわち全体集合を表す。図中のQのサークルは、語Qをクエリとして文書ランキング検索した結果として得られる文書集合を表す。一方、文脈Cをクエリとして文書ランキング検索して得られる文書集合がCのサークルに対応する。この時、語Qの文脈Cにおける類義語は図中でC－Qで表した範囲の文書集合中に出現する場合が多いと仮定し、このC－Qの文書群を類義語の抽出元とする。この仮定は、ある語とその類義語は文書中で直接共起しにくいという傾向を明示的に利用することに対応する。図中QCで表された部分、すなわちQとCの共通集合から語Qの出現パタンを収集し、それを用いて、C－Qから類義語候補を収集する[1]。

最後に類義語候補を、適切に定義された選択値を基にしてランキングし出力する。

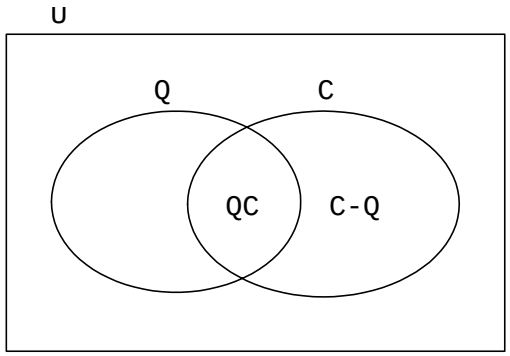

図 1: 類義語検索における文書集合

[1]ユーザが与えた語Qと文脈Cの内容に依っては図 1 において、QとCの共通部分が無い場合も有り得る。このような場合にはCに関する質問拡張技術により対処する。

ＱＣの文書は語Qの出現パタンの抽出元となり、Ｃ－Ｑの文書は類義語の抽出元となる。これらをシード文書群と総称する。シード文書群は、その全体を利用するのではなく、その文書ランキングで上位にランクされた比較的少数の文書群のみを利用する。

以上説明した類義語検索の処理の流れは以下のようになる。

1. クエリQによる文書検索
2. クエリCによる文書検索
3. シード文書群の決定
4. Qの出現パタンの収集
5. Qの類義語候補の収集
6. 類義語候補のランキング

以降は上記各処理の実現法を各説する。オンライン検索では応答速度も重要である。そこで実用的な処理効率を考慮した実現法としている。

3.1 文書検索

予め与えられランキング検索の対象となる文書集合Uにはその為の索引づけを行い文書ランキング検索システムに登録しておく。

我々の文書ランキングシステムＦＴＳ [5] は、文字 n-gram を索引単位として用いることができる。これにより、日本語文書を対象とする場合、技術用語など、形態素解析が困難な質問語Qが入力された場合でも、Qの出現文書を完全に漏れなく検索することができる。

文脈Cをクエリとした文書ランキング検索では、まず入力テキストCに対し形態素解析および異表記正規化等を含む質問処理により検索語群を生成し、その後、各検索語毎に検索が実行され結果がマージされてゆく。

3.2 シード文書の決定

前述の文書ランキングシステムＦＴＳの検索結果は、文書番号順に並んだ文書スコア付きの文書リストである。最終的なランキングは、その文書スコアをキーとして、上記文書リストをソートすることで得られる。

よって文書番号順に並んでいる段階で、それらを付き合わせることで、文書リストQとCから、その共通部分ＱＣおよびＣ－Ｑ部分を高速に決定できる。

その後、ＱＣおよびＣ－Ｑを、文書スコアをキーとしてソートし、それらの上位 n 文書（ここで n は調整用パラメタ値）を、各々、質問語Qの出現パタン収集用、および、類義語候補の抽出用のシード文書群として決定する。

3.3 質問語の出現パタン収集

ＱＣから決定したシード文書群からの質問語Qの出現パタンを収集する。ここで、出現パタンは、具体的には文書中でQの左側に d 語以内で出現する語の集合（左側出現語）と右側に d 語以内で出現する語の集合（右側出現語）により表現する。ここで d は調整用パラメタ値である。従来の語彙文脈窓の幅に相当する。

各シード文書毎に、その文書を１回走査することで、左側出現語の集合Ｌと右側出現語の集合Ｒを簡単かつ高速に求めることができる。集合ＬとＲ中の各語には、出現したシード文書数を記録してゆく。最終的に、出現したシード文書数が m 以下だった左側出現語、および、右側出現語は用いない。ここで m は調整用パラメタ値である。

3.4 類義語候補の抽出

前述したようにオンライン検索であるから応答性を重視したい。左右出現語の組み合わせをパタンとすると、パタン数が増大し、処理効率が低下する。そこで次のように、左右別々に処理を行う。

類義語抽出のシード文書毎に、各左側出現語の右側に d 語以内で出現する語の集合ＬＲ、および、各右側出現語の左側に d 語以内で出現する語の集合ＲＬを求める。集合ＬＲとＲＬ中の各語には、出現したシード文書数を記録してゆく。

最終的に集合ＬＲとＲＬが求まった後、その共通集合ＬＲ ∩ ＲＬに属する語のみを類義語候補とする抽出する。

3.5 類義語のランキング

類義語候補 s の各々に対し、次の選択値 TSV を計算し、この値の降順にランキングする。

$$TSV = (l + r) \cdot balance \cdot weight \qquad (1)$$

ここで l, r は類義語候補 s のＬＲおよびＲＬでのシード文書数である。$balance$ は左右の出現語との共起数が近いほど大きな値を取るよう以下に定義される。

$$balance = \begin{cases} l/r & \text{if } l < r \\ r/l & \text{if } l \geq r \end{cases}$$

頻出語の TSV が大きくなるのを抑制するため類義語 s 出現文書頻度 n および文書集合 U 中の総文書数 N で定義される $weight$ を用いる。

$$weight = \log(N/n + 1) \qquad (2)$$

4 実験

前節で説明した方法を実装し、類義語のオンライン検索の実験を行った。使用データは NTCIR-3 の特許タスクにおいて配布されたテストコレクションである [6]。このテストコレクションは公開公報約７０万件を検索対象とし、31 の検索課題（トピック）および、課題毎の正解データが付属している。

各検索課題には以下に例示するように DESCRIPTION など検索要求を表現するフィールドの他に、CONCEPT フィールドが記述されている。

```
<DESCRIPTION>
ステッピングモータの微小角誤差を小さくする駆動制御
装置または制御方法
</DESCRIPTION>
<CONCDEPT>
ステッピングモータ　微小角　駆動 装置
</CONCEPT>
```

本実験では、この CONCEPT フィールドに記述された複数の語の内、先頭に記述された語を質問語Ｑとし、残りを文脈Ｃとした。上記例では “ステッピングモータ” が質問語Ｑであり、“微小角 駆動 装置” が文脈Ｃに相当する。添付資料として、全３１課題について、類義語を検索した結果を示した。

今回の実験結果を見るかぎり、現状では精度が出ていないものの、見通しの良いシンプルな手法であり、応答時間もワークステーション上で 10 秒程度と実用レベルに届く範囲にある。この結果の誤り分析を基に、各種改良に取り組んでゆくつもりである。

5 おわりに

与えられた任意の語Ｑおよび文脈Ｃの組みに対し文脈Ｃにおける語Ｑの類義語群をオンラインでランキング検索するための手法を提案した。

今後は英語および日本語の様々なコレクションを対象とした実験を基に手法を改良する。特に類義語の選択値の設定には改良の余地が多分にある。例えば左右出現語が類義語候補にスコアを与えるようにすることでより木目の細かい選択ができる可能性がある。

類義語検索は定量的な性能評価が難しい。また文脈に応じた類義語の検索という課題自体が新規であり、評価用データが見当たらない。語はその語自身と類義関係にあるという自己類義性（self-synonymy）を利用することで定量的に性能を評価できる可能性がある。

また対話型文書検索における質問拡張への適用や特許マップ作成用のオントロジー構築など、具体的な応用のもとでの評価も行ってゆく。

参考文献

[1] 山本, 梅村. 辞書を用いない関連語リストの構築方法. 自然言語処理研究会報告, Vol. 148-12, pp. 81–88, 2002.

[2] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO'94*, pp. 146–160. New York : Rochefeller University, 1994.

[3] H. Schutze and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, Vol. 33, No. 3, pp. 307–318, 1997.

[4] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on In-*

formation Systems, Vol. 18, No. 1, pp. 79–112, 2000.

[5] Y. Ogawa and H. Mano. RICOH at NTCIR-2. *Proc. of NTCIR Workshop 2 Meeting*, pp. 121–123, 2001.

[6] M. Iwayama, A. Fujii, A. Takano, and N. Kando. Patent retrieval challenge in NTCIR-3. *IPSJ SIG Notes*, Vol. 2001-FI-63, pp. 49–56, 2001.

付録

01:乳化
微細化 撹拌 分散 撹拌混合 成形 解繊 混合 懸濁 凝集

02:種子
作土 モミ 籾 孔隙 播種 施用 水田 苗箱 苗 菌 植物

03:ステッピングモータ
ヨウモウタ 俯仰モウタ ステツプモウタ 推力 旋回モウタ

04:符号
バーコード バーコードパターン グループ コードコード体系

05:キチン
脂質 担子菌 抗生 担子 酵素 反応 グラフト化 燐酸塩

06:レンズ付きフィルム
プリペイドカード パチンコ パチンコ台

07:ガソリン直噴エンジン
燃料 機関 インジェクタ 希薄燃焼 内燃機関 エンジン

08:シリコーン
抗酸化 アミノ 変成 保湿 抗菌 付加 タンパク 溶解 組成

09:硬貨
地絡 光ビーム 感光体 センサ 校正 返却 同期 判定

10:ホルムアルデヒド
悪臭 アセトアルデヒド 脱臭剤 光触媒 吸着剤 アルデヒド

11:茶
カキ肉 カキ エキス 焙煎 グルカン粉末 糠エキス たばこ

12:発光波長
発光 紫外 紫外光 波長 素子 黄緑 黄緑色 青紫色

13:かんばん方式
（類義語は得られず）

14:振動
返答 返答信号 電波 信号 帯 タイミング 調節

15:ポーラス
共存 鋳型 TIN 発泡 発泡体 ロール 粉体 粉 銅 金属

16:水質汚染
酸素 水質 全窒素 アルカリ 窒素 含有 全

17:衣類
艶 艶だし 含水 水分 配合 樹脂 加熱

18:カラオケ
通信 システム CTI 着呼 自律的 自律

19:耳式
外耳道 放射 赤外線 外耳道 耳孔 耳

20:使用済み油
脱硫

21:印刷
刺繍 紐 紐体 マイコン 模様 テイブル シート シート状

22:窒素酸化物
NOx 空燃比 酸素 濃度 燃焼 リン 放出 吸蔵 NOx 濃度

23:亜鉛溶融めっき
（類義語は得られず）

24:イオン交換膜
固体高分子 ナフイオン 高分子電解質 電解質 イオン解離

25:光触媒
耐摩耗 親水性 析改質剤 有機チタネイト ラジカルカツプリング

26:天然ガス
ダイヤフラムポンプ パレツト プラズマ フツ素 水素 ガス

27:薄型
発泡 材質 有機 電子機器 位置

28:携帯電話
外来 分子配向 迅速 電磁波 吸収体 基板 加工 検出位置

29:米びつ
（類義語は得られず）

30:太陽
宇宙 宇宙船 マイクロ アンテナ マイクロ波 受信 機器

31:飲料
金属クロム 水酸化 水酸化ナトリウム 缶 鋼板 水 固溶