

コーパスからの同義語の獲得(1) — 近傍単語頻度統計によるアプローチ —

相川 勇之 伊藤 山彦 鈴木 克志

三菱電機株式会社 情報技術総合研究所

4Q-3

1. はじめに

我々は、従来のシステムでは困難だった高度な業務支援を可能とするヘルプデスク支援システムの構築を検討中である。同システムでは、過去の相談事例を整理して事例ベースを構築し、類似事例検索や分類集計の機能を提供する予定である。我々は、既存データを効率よく利用するために、自然言語で記述されているテキストそのものを事例として扱おうと考えている。このようなテキスト事例を整理するためには、類似文の照合処理が必要である。

類似文照合処理には、同義語辞書、および類義語辞書が必要だが、汎用の辞書を用いた実用化は困難であり、個々の開発事例にそって、処理対象に特化した辞書を個別に開発しなくてはならない。同義語関係を既存の大量データから自動獲得できれば、辞書開発にかかるコストを軽減できる。

本稿では、文書検索や文書自動分類において用いられている統計的手法を取り入れ、対象部門で蓄積された文書集合から、同義語および類義語を半自動的に抽出する手法について検討する。また、試作プログラムによる実験結果について報告する。

2. 背景

我々が処理対象とする文書集合の一例として、当社のお客さま相談センターの問合せ記録がある。これは、顧客からの電話相談内容を記録し整理したもので、入電内容および対応処置の部分が日本語で記録されている。入電内容に類似する事例を問合せ記録から効率良く検索できれば、電話で対応するオペレータの業務を支援することができる。

この記録は、電話で対応した各オペレータが自由に入力するため、表記の揺れや用語の揺れを多く含んでいる。類似する事例を検索するためには、これらの揺れを吸収するための同義語辞書が必須である。

文書の電子化が進んでいる昨今、上記の問合せ記録のような業務固有のコーパスが数多く存在すると考えられる。

しかし、同義語のような意味的な情報を扱う場合、既存の汎用辞書では、きめ細かな処理は困難である。したがって、各業務に応じた同義語辞書を個別に開発する必要がある。

そこで我々は、応用範囲の広いシステムを構築するため、統計的手法を用いてコーパスから同義語候補を自動抽出して提示する同義語辞書作成支援ツールを開発し、辞書開発コストの軽減をはかることにした。

3. 統計的手法による同義語関係の推定処理

コーパスから抽出した統計情報により、文書中に出現する単語の多義性を解消したり、未知の語義を獲得するという研究は従来から行われている。

福本ら[福本96]は、分類精度を向上するため、辞書語義文を用いて単語の多義性を解消し、共起ベクトルを用いて多義解消後の名詞間の同義語関係をとらえてリンク付けしている。

多義解消を拡張して未知語の語義推定を行なう研究には、コーパスから係り受けの2項関係を抽出し、その統計情報をもとにシソーラスに未知語を配置する研究[浦本96]や、コーパスから「を格」に関する統計情報を抽出して、分類語彙表の未知語義を推定する研究[新納97]などがある。

4. 近傍単語頻度統計による類似度計算

福本らが行なっている名詞間リンク付けは、同義語関係の獲得そのものであり、浦本らの行なった未知語の語義推定も、未知語と既知語の同義関係の獲得である。しかし、福本らの研究対象は英語(Wall Street Journal)であり、特に名詞間の関係に着目したものである。また、浦本、新納の研究では日本語を対象としているが、コーパスとして係り受け解析までなされたものを想定しているため、コーパス作成コストの点で問題がある。

我々は、処理対象としたコーパス（上記問合せ記録）に現れる日本語が、以下の性質をもつことから、形態素解析結果の自立語頻度統計のみをもとにして同義語関係を抽出できると考えた。

Acquisition of Synonyms (1)

— Approach using Co-Occurrence Word Frequency —

Takeyuki AIKAWA, Takahiro ITOH, Katsushi SUZUKI
Mitsubishi Electric Corporation.

5-1-1 Ofuna, Kamakura, Kanagawa 247, JAPAN

性質 A: 非常に短い文で記述されており、その大部分が単文である。

性質 B: 限られた領域での対話記録であるため、多義語は稀にしか現れない。

性質 C: 非常に短い文で記述されているため、表層格情報の欠落なども見受けられる。

図 1 に、入電内容および対応処置の一例を示す。図中 Q が入電内容で、A が対応処置である。

Q: 浴室換気扇を買替たい、店紹介してほしい
A: ○×電機紹介
Q: 2台追加購入したい電力容量を増やしたい
A: □△電力に相談して
Q: ハイビジョンみたい
A: M/Nコンバーター必要と説明

図 1 入電内容と対応処置の文例

性質 B より、多義性の解消は不要と判断した。また、性質 C より、自動的な係り受け解析は困難と判断した。

形態素解析の精度は、係り受け解析と比較するとかかなり良く、ロバストな処理も可能なので、生のテキストデータから自動的に自立語の頻度統計をとることは容易である。コスト最小法[吉村 89]に基づくパーザにより形態素解析を行ない、自立語のみ抽出して、前後 2 語ずつ計 4 語との共起頻度統計をとった。前後 2 語というのは少なく見えるが、上記性質 A より小さめの値を設定した。また、福本らとは異なり、名詞以外の語についても、近傍単語の頻度統計を抽出した。

次に、共起頻度統計から品詞情報によりノイズの原因となる不要な統計情報（接続詞、連体詞などの機能語）を除去した。各単語の近傍に共起する単語の頻度を要素とする近傍共起単語頻度ベクトル同士の内積値から、入力語と頻度 10 以上の全単語との類似度を計算し、上位 30 語を入力語の同義語候補として抽出した。

さらに、同義性の高い単語同士は共起しにくいというヒューリスティックにより、上記 30 語のフィルタリングを行ない、上位 10 語を最終的な同義語候補として提示し、目視チェックにより同義語および類義語を抽出した。

5. 実験結果

実験には、上記の問合せ記録から抽出した約 9 万文を使用し、評価用の単語集合として同問合せ記録に含まれるカタカナ語および英字列のみで構成される未知語 455 語を用いた。実験の結果、略記法および表記の揺れを中心とする 98 語(68 組)の同義語関係を抽出することができた。表 1 に実験結果を示す。表中の 3 行め以下は、抽出した共起単語統計情報の合計頻度の値による内訳である。

予想通り、頻度の大きな語の方が抽出精度が高くなっている。

表 1 実験結果

	同義語 関係あり	類義語 関係あり	関係なし	その他
全単語(455 語)	21.5 % (98 語)	23.7 % (108 語)	52.1 % (237 語)	2.6 % (12 語)
100~(98 語)	33.7 %	29.6 %	34.7 %	2.0 %
50~100 (113 語)	23.9 %	20.4 %	52.2 %	3.5 %
30~50 (121 語)	19.0 %	25.6 %	52.1 %	3.3 %
10~30 (123 語)	12.2 %	20.3 %	65.9 %	1.6 %

同義語関係として、主に 2 種類の結果が得られた。ひとつは略記法に関するもので、もうひとつは表記の揺れに関するものである。

前者の例として「リモコン」と「RC」、「パソコン」と「PC」のような関係がある。後者の例として「S-VHS」と「SVHS」のような関係がある。予想通り、多義語はほとんど存在せず、「NO」（「番号」と「いいえ」）および「CH」（「チャンネル」と「クリーンヒータ」）という語以外には見つからなかった。

類義語関係として、「BS ANT（BS アンテナの略記）」と「ANT（アンテナの略記）」などの上位下位の関係にあるものと、「PHS」と「携帯」のようにシソーラス上での兄弟関係にあるものが抽出された。その他として、擬音、形態素抽出ミスなどがあつた。

6. まとめ

実験に使用した全単語のうち約 20%強の単語から同義語関係を抽出できた。コーパス中に同義語が出現しない単語をいかに分離するかが今後の課題である。そのためには、形態素解析の精度を向上し、よりロバストな解析ができるようにすることである。また、現在は形態素解析結果として、複数の単語として分割されている複合語の構造解析により、正確な複合語の頻度を得る必要がある。

また、相互情報量による重み付けを行なった場合との、抽出精度および速度性能の比較などを行ない、本手法との優劣を評価する予定である。

参考文献

- [福本 96] 福本: コーパスに基づくシソーラス統計情報を用いた既存のシソーラスへの未知語の配置, 情報処理学会論文誌, Vol. 37, No. 12, pp. 2182-2189.
- [新納 97] 新納: コーパスを利用した未登録語義の発見, 情報処理学会論文誌, Vol. 38, No. 5, pp. 953-961.
- [福本 96] 福本, 他: 辞書の語義文を用いた文書の自動分類, 情報処理学会論文誌, Vol. 37, No. 10, pp. 1789-1799.
- [吉村 89] 吉村, 他: 未登録語を含む日本語文の形態素解析, 情報処理学会論文誌, Vol. 30, No. 3, pp. 294-301.