Institute of Electronics, Information, and Communication Engineers

D-3-1

要求文書中の同義語推定手法の開発

Development of the detection method of synonyms in the requirements documents

平尾英司* 久野綾子* 五藤智久* 伊藤晃** 小林大輔** 川井康示** 田村一樹** 吉川大弘**古橋武**
Eiji Hirao Ayako Kuno Tomohisa Gotoh Akira Ito Daisuke Kobayasi Yasushi Kawai Kazuki Tamura Tomohiro Yoshikawa Takeshi Furuhashi

*日本電気株式会社 サービスプラットフォーム研究所
Service Platforms Research Laboratories, NEC Corporation

**名古屋大学大学院 工学系研究科
Graduate School of Engineering, Nagoya University

1. はじめに

要件定義書や仕様書等の SI 上流工程の要求文書は複数の担当者による分筆が多く、同一の内容を異なる表記にした同義語が埋め込まれやすい。同義語は同一のものを異なるものと誤解させ、不要な機能や誤った動作の作りこみに繋がるため、早期に検出する必要がある。

本稿では、SI 上流工程の要求文書特有の同義語に着目し、同義語候補を推定する手法の提案と提案手法を実データに適用した評価結果を報告する。

2. 提案手法

翻訳や検索などの分野では、同義語の推定は精度向上に関わる重要な課題として、様々な方法が提案されている[1]. これらの手法の多くは、同義語は文脈情報が類似するという仮説に基づき、共起語の出現頻度を用いた共起ベクトルのコサイン類似度などを抽出に利用している。しかし、SI 上流工程の要求文書は、①案件に関する文書内というスモールコーパスでのみ成り立つ同義語が存在する、②辞書に未登録の未知の複合語が大量に含まれる、という点で一般の文書と異なる。このため、既存の同義語推定手法をそのまま適用しても高い抽出精度は期待できない。そこで我々は以下の複数のアプローチを組み合わせた抽出手法を考案した。

- シソーラスを用いた概念集約：スモールコーパスで共起ベクトルを作成すると、共起語が少ないため疎なベクトルとなり、単語間の類似性を考慮しにくい。そこで、各共起語のシソーラス上の概念が同じ共起語を同一語とみなし、共起ベクトルを概念で集約した概念ベクトルを作成することで、類似性を考慮しやすくする。
- 未知語への概念付与：シソーラスに未登録の未知語は、概念での集約ができないため、概念ベクトルの作成が困難になる。そこで、未知語の大半を占める複合語について、複合語を構成する熟語（以下、構成語）に分解し、構成語毎の概念を付与する。また、概念付与の際、より概念上の支配力の大きい構成語ほど重視されるように重み付けを行う。
- 共起語の除外：同義語が同文中で利用される事例が少ないことから、共起語を同義語候補から除外する。

3. 実験

1)適用データ

Web 上で公開されている仕様書[2]を評価用データとして利用した（文字数：46,945）。また、同義語候補の抽出性能を把握するために、抽出すべき正解情報として、文書内の表 1 に示す 6 個の単語の約半数を他の語に置き換えることで、正解情報となる同義語対を埋め込んだ。なお、この置き換えは同一文に同義語対が混在しないよう制限をかけた上でランダムに行った。

2)実験条件

以下の条件で適用データの全単語間の類似性を算出し、正解の同義語対の抽出順位を算出した。

- 共起語：句点および連続する改行を区切りとした一文を共起範囲とし、名詞、動詞、形容詞を共起語とした。
- 概念集約：概念分類が大分類、中分類、小分類の各粒度で集約した場合について比較した。
- 未知語への概念付与：構成語の支配力は同一の構成語を持つ複合語群の共起ベクトル間の類似性が高い構成語ほど支配力が高くなる指標で算出した。
- 類似性の指標：コサイン類似度をベースにしたが、共起語が多く一致するほど類似性が高まるよう共起ベクトルに直交ベクトルを付加した。

4. 結果

適用データにおける 2,208,151 組の単語対（名詞のみ）の中で、共起ベクトルの類似度のみに基づく既存手法に比べ、概念集約を行った場合の正解の同義語対の抽出順位は、約 1/2～1/76 と大幅に上昇した。集約の粒度は小概念より中概念や大概念での集約の方が順位を向上させた。さらに、未知語の概念付与の際、支配力で重み付けにより、均等に重み付けた場合より半数の組合せに関して 1/2 程度の順位向上が見られ、また共起語の除外により 1/3～1/280 程度の順位向上が見られた。既存手法と最適条件での提案手法の正解の同義語対の抽出順位の比較を表 1 に示した。

表 1 正解情報とした同義語対の抽出順位比較

正解情報とした同義語対(出現数)		既存手法	提案手法
管理(89)	監督(75)	4,459	11
運営受託者(61)	運用受託者(56)	4,633	2
作成(16)	策定(23)	8,359	79
障害(17)	弊害(16)	9,353	141
検討(8)	思案(6)	15,880	207
保管場所(10)	保存場所(5)	4,459	981

表 1 より、提案手法では、上位 200 件程度をチェックすることで、除外すべき同義語候補の大半を抽出できることが確認された。課題として出現数が 10 個以下の語の順位が低い傾向が見られるため、改善を検討してゆく。

参考文献

[1] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL" In Procs. of the Twelfth European Conference on Machine Learning (ECML-2001), pages 491-502, Freiburg, Germany(2001)

[2]がん対策情報センターシステム運用業務委託仕様書、厚生労働省(2009)

Copyright © 2012 IEICE

NII-Electronic Library Service