異なる辞書を利用した意味コードの自動付与

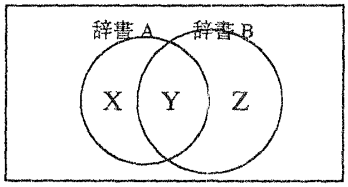
1N－7

鷹尾 和享　柏岡 秀紀　白井 諭
ATR 音声翻訳通信研究所
E-mail: {ktakao, kashioka, shirai}@itl.atr.co.jp

1　はじめに

単語の意味を利用した処理が重要視されているが、単語に意味コードを付与する作業は、人的・時間的コストが大きな負担となる。従来使用していた辞書とは語彙数やカバレッジの異なる辞書を利用し、付与作業を自動的に行うことができれば、負担を大幅に軽減できる。その場合意味コード体系が異なっているので変換が必要になる。本稿では意味コードの自動付与の手法を検討した結果について具体例を交えながら報告する。

2　手法



従来利用していた辞書Aとは異なる辞書Bを利用する場合、両者に共通して存在する語(図のY)を元に意味コードの対応関係を得ることができる。意味コードの対応付けができれば、辞書Bの語(図のZ)で辞書引きをし、辞書Bの体系を経由して、辞書Aの体系の意味コードが得られる。従来のXとYでの辞書引きの他にZが加わることになる。作業手順は次のようになる。

(1)辞書B→辞書Aのマッピング表をプログラムで自動作成。
(2)人手によるチェックを行い、マッピング表を完成する。
(3)辞書Bとマッピング表を使って付与対象の語に意味コードを自動付与。
(4)対象を絞り込んだ上で人手によるチェックを行う。

Automatic Addition of Semantic Codes Using a Different Dictionary
Kazutaka Takao, Hideki Kashioka, Satoshi Shirai
ATR Interpreting Telecommunications Research Laboratories

3　具体例

筆者らは、科学技術用語の英日対訳辞書（JSTの日英対訳辞書から作成）の約３５万語に意味コードを付与する実験を行った。これは、英単語とその日本語訳から構成される。我々は従来より角川類語新辞典[1]の意味コード体系を用いてきたが、日本語語彙大系[2]を利用して意味コード付与を効率的に行う方法を検討した。

3.1 マッピング表の自動作成

意味コードの組み合わせ毎に、両方に共通する語を拾い出し、意味コードの対応関係を抽出した。さらに、そのうち妥当そうなものを残して残りを切り捨てる処理、および、人手によるチェックの対象の絞り込みを行った。

3.2 マッピング表の人手によるチェック

意味コードによっては共通して存在する語がない場合があるので、手作業で意味コードの対応を補った。また、分類体系が違うため意味コードの対応関係を個別にチェックする必要がある場合は"要ﾁｪｯｸ"の印を付けた。その際、次の段階で予備情報なしに辞書で探すよりも、挙がっているものから間違いを削除する方が簡単なので、対応の可能性のあるもの全てを残すようにした。(例：化 → 腐敗 強化 変動 変質 凝固 美化)

3.3 意味コードの自動付与

付与対象の語数が多いことから、従来作業者が行っていた一連の手順を全てプログラムで自動的に行うこととした。なお、英語の単語に意味コードを付与するのが目的であるが、英語で辞書を引く処理の他に、対訳の日本語で辞書を引く処理も併せて行った。すなわち、

(1)過去に手作業で付与した語を利用して付与できる場合(英語で参照)
(2)定型パターンにより辞書を見なくてもわかる場合（英語の例：○○ network、日本語の例：○○剤)
(3)角川を参照して付与できる場合：図のXとY(日本語で参照)
(4)日本語語彙大系とマッピング表から付与：図のZ(日本語で参照)

さらに、付与対象には複合語が多いことから、以下の処理も併せて行った。

(ア)英語が2語以上の場合は意味をなす上で重要な単語に視点を当てる（例："draft budget prepared by ministry of finance" → "budget")
(イ)日本語の一部分を区切って辞書を調べる（例：亜鉛イオン → イオン)

3.4 自動付与結果の分析と人手による修正

自動付与した結果からランダムにいくつか抽出して傾向を調べたところ、以下のことがわかった。

(a)マッピングを使用したものは、正しい意味コードのほかに合わない意味コードが多数挙がっているものや、挙がるべき意味コードが欠けているものがあるが、おおむね良好である。
(b)英語の"意味をなす上で重要な単語"はおおむね正しく得られている。また、これで付与できた語の割合が比較的高い。
(c)日本語の区切れ方が間違っているものが目に付く。その場合は間違った語で辞書を参照しているので、意味コードも当然間違いである。

３５万語を全て人手でチェックすることはできない。そこで、結果が合っていそうなものはチェック対象から除外し、間違っていそうなものに重点的にチェックを入れることとすれば、少ない手間で効果的に品質を向上させることができる。すなわち、

(1)区切れ方の間違っていそうなもの（例：カタカナの並びの途中で切れている場合)
(2)マッピング表で"要ﾁｪｯｸ"の印があるもの

をチェック対象にすることとした。

3.5 考察

意味コードの決まり方は表1の通りである。日本語語彙大系を利用しない場合付与できない語が約 10 万語(91184+9580)生じるが、利用したことによって 1/10(9580 語)に減少したことがわかる。

また、人手によるチェックの対象を表2に示す。

角川(3)	68732(19.6%)
日本語語彙大系とﾏｯﾋﾟﾝｸﾞ(4)	91184(26.1%)
その他(1)(2)	180486(51.6%)
自動付与できず	9580(2.7%)
合計	349982(100.0%)

表1：意味コードの決まり方と語数

区切れ方チェック	27217
マッピングチェック	8051 ※重複あり
チェック対象合計	35072

表2：人手によるチェックの対象の語数

4 まとめ

本稿では、意味コード付与での異なる意味コード体系の辞書利用が有効であることがわかった。また、対象の語に複合語が多いことから、語を区切って辞書を引くことが有効であることがわかった。ただし、区切り方については改良の余地がある。

現在、本手法で作成した意味コード辞書を機械翻訳システムに利用することを検討中である。また、付与結果の品質の評価を機械翻訳の観点から行うことを検討している。

最後に、研究に当たってご協力頂いた NTT コミュニケーション科学研究所の諸氏に感謝いたします。また、本検討にあたりご討論いただいた JST の関係各位に感謝いたします[3]。

参考文献

[1]大野晋,浜西正人: 角川類語新辞典 CD-ROM 版, 角川書店(1989)
[2]NTT ｺﾐｭﾆｹｰｼｮﾝ科学研究所監修: 日本語語彙大系 全5巻, 岩波書店(1997)
[3]H. Ohta et.al: Applying TDMT to Abstracts on Science and Technology, MT Summit VII(1999)