論文データベースからのイディオム用例検索

森下　智史　　難波　英嗣　　相澤　輝昭
（広島市立大学　情報科学研究科）

1. はじめに

研究者が英語論文を書く時，もしある日本語に対応する英語表現がわからなければ，和英辞書を引いて調べることができる．しかし，ある日本語に対する英訳語が複数存在する場合，どれが最適であるのか判断に困ることがある．一般的な内容の英文を書く場合には，和英辞書に載っている用例を見ることで，どの訳語が適切か判断できる．しかし，学術論文の場合，用語やイディオムの用法が一般的な用法と異なることがしばしばあり，辞書の用例だけから最適な訳語を選択することは困難であると思われる．本研究では，論文データベースからイディオムの用例を検索・提示することで，英語論文を書く研究者の訳語選択を支援するシステムの開発を行う．

2. 関連研究

これまでに様々な観点から数多くの英作文作成支援システムが開発されてきたが，本節では本研究と特に関連が深いと思われる武田ら[1]の研究について述べる．武田らはあらかじめ用意してある例文集合を参照することにより，適切な英文の作成を支援するシステムを開発している．具体的な手順は次のとおりである．ユーザはまずシステムに日本語文を入力する．次にシステムは，入力文から核となる意味(アイディア)を抽出する．キー概念と日本語単語の対応表を用い，抽出されたアイディアに対応するキー概念を抽出する．キー概念を組み合わせて検索式を作成し，用例の検索を行う．

この方法は，日本語単語や英単語のキー概念との対応表データが必要であり，データに含まれていないキー概念が現れた場合，人手でキー概念と単語の対応表データを更新する必要がある．

3. イディオムの用例検索

本研究では，２語以上の単語から構成されているものをイディオムと呼ぶことにする．

イディオムは以下のように分類することができる．

- 単語の連続性による分類
  連続型…構成単語が一続きになっているもの．
  (例)choice of ~【～の選択】
  分離型…構成単語が離れているもの．
  (例)regard ~ as …【～を…とみなす】
- 品詞による分類
  動詞形…動詞句として用いられるもの．
  (例)is attributed to ~【～の原因となる】
  名詞形…名詞句として用いられるもの．
  (例)decrease in ~【～の減少】
  その他の形…副詞句や形容詞句や前置詞句をまとめたもの．
  (例)because of ~【～の理由から】

本研究では上記のようなイディオムを用例検索システムの入力とする．また，上記の分類を考慮した検索方法を提案する．

◆　単語の連続性を考慮した検索

連続型については，単純な文字列の一致を探す．分離型については 2 種類の方法で検索を行う．ひとつは，分離している構成単語間[1]に任意の文字列を含んでよいという条件で検索する方法，もうひとつは，分離している構成単語間が１つの名詞句の場合のみ適切な用例と考える場合である[2]．後者は分離している構成単語間が名詞句であるかを判定するために構文解析する必要があり，前者よりも検索のための手順が増えるが，その分検索精度の向上が期待できる．

◆　品詞を考慮した検索

原文中で動詞が活用変化したり名詞が複数形になっている場合にも検索できるようにするため，論文原文に LimaTK[2]を適用し，原文中の各単語を原形に変換（原形文）する．

4. システムの用例検索過程

図１　システムの用例検索過程

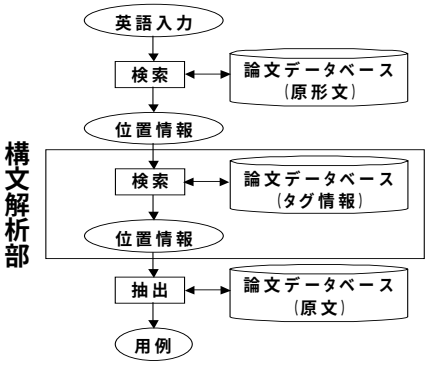

図 1 にシステムの用例検索過程を示す．図中の論文データベースについて，(原文)は原著論文，(原形

[1] 例えば regard～as の場合，regard と as の間の文字列を指す．

[2] 本研究では，例えば脚注 1 の例の場合，regard の直後に名詞句が出現するかどうかで判定する．

文)は LimaTK を用いて全ての単語を原形に直した文, (タグ情報)は原文の構文解析結果として得られた品詞タグ，位置情報はイディオム構成単語が文頭から何番目に存在するかという情報をそれぞれ示す.

5. 実験

・実験方法

自然言語処理分野のフルテキスト英語論文 700 論文から, イディオム 32 個について各イディオム最高 10 件の用例を検索した．ここで, (a)構文解析を用いない場合と(b)構文解析を用いる場合の 2 つの方法で用例検索を行った．その結果，構文解析を用いない場合で 264 個，構文解析を用いる場合で 239 個の用例が得られた．これらを人手で調べ，検出精度を求めた．なお，イディオムを構成する単語数およびブロック数毎に検出精度を算出している．ブロックとは，イディオム中にある 1 連の単語の並びを 1 集合とした場合の集合の数を表すものとして定義したものである．2 つ以上のブロックから構成されるイディオムが 3 節で述べた分離型イディオムに相当する.

・結果

検出精度を表 1 に示す．表 1(a)は構文解析を用いない場合を，表 1(b)は構文解析を用いる場合を表している．表からわかるとおり，全体的に 1 ブロックで構成されているイディオムの用例検出精度は高く，ブロック数が多くなるほど検出精度は下がっている.

また，構文解析を使用する場合と使用しない場合では，構文解析を用いると得られる用例の数が 2 ブロックの場合 12 件，3 ブロックの場合 13 件減少したが，検出精度が大幅に向上していることが分かる.

・考察

構文解析を用いた場合と用いない場合の検出精度について考察する．例えば“classify～in …(～を…に区別する)”というイディオムを構文解析を用いずに検出すると次のような用例が検出される.

> In the previous account these cases are all classified as non–parallel, resulting in the...

classified と in の間にはカンマがあり，classify～in の用例としては適切ではない．しかし，構文解析を用いれば, classified の直後に名詞句が出現しないという情報から，このような用例は検出されない.

構文解析を用いる場合の検出誤りは，主に前置詞のみのブロックがある場合に，前置詞がイディオムとしての用いられ方であるのかそれ以外の用いられ方であるのかの判断の誤りから起こるものであった.

6. おわりに

本研究ではイディオムの用例検索システムを作成した．イディオムの分類を行い，次に分類を考慮した検索手法を提案し,用例検索実験を行った．実験の結果，構文解析を用いない場合の検出精度 87.1%, 構文解析を用いる場合の検出精度 93.7%が得られ, 構文解析が用例検出に有効であることが分かった.

参考文献

[1]武田明子，古郡廷治：例文をもとにした英文書作成支援システム，情報処理学会論文誌，Vol.35, No.1，pp.53-61.（1994）

[2]山下達雄，松本裕治：言語に依存しない形態素解析処理の枠組，自然言語処理，Vol.7，No.3，pp.39-56.（2000）

[3]Satoshi Sekine: The domain dependence of parsing, In Proceedings of Fifth Conference on Applied Natural Language Processing, pp. 96-102.（1997）

表 1　用例検索精度

(a)　構文解析を用いない場合　　(単位：%)

		2 単語	3 単語	4 単語	5 単語	合計
連続型	1 ブロック	97.2　(69/ 71)	100.0 (34/ 34)	100.0 (30/ 30)	100.0 (1/ 1)	98.5 (134/136)
分離型	2 ブロック	71.2　(42/ 59)	86.4　(38/ 44)			77.7 (80/103)
	3 ブロック		10.0　(　1/ 10)	100.0 (11/ 11)	100.0 (4/ 4)	64.0 (16/ 25)
	合計	85.4 (111/130)	83.0　(73/ 88)	100.0 (41/ 41)	100.0 (5/ 5)	87.1 (230/264)

(b)　構文解析を用いる場合

		2 単語	3 単語	4 単語	5 単語	合計
連続型	1 ブロック	97.2　(69/ 71)	100.0 (34/ 34)	100.0 (30/ 30)	100.0 (1/ 1)	98.5 (134/136)
分離型	2 ブロック	80.4　(41/ 51)	92.5　(37/ 40)			85.7 (78/ 91)
	3 ブロック		100.0 (　1/　1)	100.0 (　8/　8)	100.0 (3/ 3)	100.0 (12/ 12)
	合計	90.2 (110/122)	96.0　(72/ 75)	100.0 (38/ 38)	100.0 (4/ 4)	93.7 (224/239)