英語文章における単語間の距離を基にした構文の特徴抽出

An charactaristic extraction of parsing based on distance between words in English sentences

児玉 憲造
Kenzo KODAMA

Abstract: When we make a monograph and start a new research, we always make an investigation into the literature written in English or the other language. But a translation software can't translate a technical term and a characteristic expression, accurately. For translation of technical terms, We need to know which word is a technical term. In this paper, I tried to extract a technical terms and a characteristic expressions from literature written in English.

1　概要

論文の作成や新たな研究を始める際には，英語を初め他国語の文献を調査，参照することが頻繁にある．その際，対象とする分野の専門用語や固有の表現に注意して訳さなければ正しく翻訳することはできない．

しかし，専門用語に関する知識が不足している場合，どの単語が専門用語であるか判別が困難である．翻訳ソフトは専門用語の判別ができないため，正しい翻訳を得られることが少ない．

また，テキストデータから有用な知識，規則等を発見する手法として「テキストマイニング [1)] 」という技術が，近年注目を集めている．そこで本研究では，この「テキストマイニング」の技術を利用して，対象とする英語文献中から専門用語，固有の表現を抽出することを試みた．

2　抽出手法

2.1　単語の頻度

専門分野について書かれた文献には，その分野に関する重要な専門用語が高頻度で出現すると考えられる．そこで対象とする英語文献から単語ごとに頻度をとることで専門用語の抽出を試みた．(Fig. 1)

A | genetic | algorithm | starts | with | a | population | of | strings

word	frequency
a	2
genetic	1
⋮	⋮
strings	1

Fig. 1　単語頻度データの抽出例

2.2　単語間の距離データ

単語ごとの頻度だけでは，熟語や決まった表現などが抽出できないため，単語間の距離という概念を用いた．これは，ある単語から次にでてくる単語を距離 1，さらに次の単語を距離 2 として，距離 100 まで求め，その単語の組合せと距離を基に頻度を計算したものである．(Fig. 2)

A genetic algorithm starts with a population of strings

front	back	distance	frequency
a	genetic	1	1
a	algorithm	2	1
⋮	⋮	⋮	⋮
of	strings	1	1

Fig. 2　単語距離データの抽出例

これらの手法を用いて，出てきた結果から専門用語や固有の表現の抽出について考察を行う．

3　実験

3.1　対象文献

本実験では，「遺伝的アルゴリズム (Genetic Algorithm：以下 GA)」という専門分野について書かれている「GENETIC ALGORITHMS in Search，Optimization，and Machine Learning 」という文献を対象として実験を行った．

3.2　実験結果 (単語頻度)

対象文献に対し，第 1 章の頻度をとった結果を Table 1 に示す．実験結果は”the”や”of”など専門用語とは直接関係のない語 (接続詞，前置詞など：以下不要語) が上位頻度に多数確認されたため，これらを省く処理を行った結果を示している．Table 1 における有効性は実

際に GA の研究を行っている知的システムデザイン研究室 GA 研究グループ (以下 GA グループ) による専門用語の判定を示している.

Table 1　上位頻度単語 (第 1 章)

順位	単語	頻度	有効性
1	string	156	○
2	genetic	111	○
3	algorithm	102	
4	search	80	
5	population	56	○

実験の結果から上位頻度に GA の専門用語を多数確認することができた.

3.3　実験結果 (単語間の距離データ)

対象文献に対し，単語間の距離データをとった結果を Table 2 に示す．この結果に対しても front と back のいずれかに「不要語」を発見した場合にその結果を省く処理を行っている．また，この結果に対しても有効性は GA グループによる判定である．このデータから，1 単語の頻度では上位に現れなかったが重要な熟語である”black box”などが抽出できた．この結果から，GA の文献において重要な表現である熟語などを抽出する事ができたと言える.

Table 2　単語間距離の上位頻度データ (第 1 章)

順位	前単語	後単語	頻度	距離	有効性
1	genetic	algorithm	94	1	○
2	black	box	16	1	○
3	objective	function	12	1	○
4	fitness	value	11	1	○
4	roulette	wheel	11	1	○

3.4　実験結果 (距離の離れているデータ)

距離の離れているデータに対しても上位頻度から重要な表現が出現するかの実験を行った．Table 3 に距離 4 の場合におけるデータを示す.

Table 3　単語間距離大の上位頻度データ (第 1 章)

順位	前単語	後単語	頻度	距離
1	schemata	population	4	4
2	consider	schemata	3	4
2	guide	search	3	4

この結果における頻度最上位の組合せに対して，文章中の出現例を調査した.

- schemata contained in the population
- schemata in a particular population
- schemata depending upon the population
- schemata contained in a population

この結果から，距離 4 においてはすべて異なった表現が出現し，特異な表現を発見することはできなかった．距離 4 以上の組合せに対してはすべてこのような結果が出たため，今回対象とした文献においては距離 4 以上の結果は有効ではなかった.

3.5　実際の使用例

本実験において，抽出することのできた単語について，以下のような例文を用いて有効性を検証した.

> A genetic algorithm starts with a population of strings and thereafter generates successive populations of strings

この例文を「j-London2000EJ(KODENSHA)」という翻訳ソフトで翻訳を行った結果，以下のような翻訳を得た.

> 「遺伝のアルゴリズムはストリングの人口で始まりその後ストリングの相次いだ人口を生成します」

本手法において，専門用語として”genetic algorithm”，”population”，”string”がわかっているので，ここに正しい翻訳を当てはめると以下のようになる.

> 「遺伝的アルゴリズムは遺伝子列の母集団で始まりその後遺伝子列の相次いだ母集団を生成します」

こうして翻訳することによって，GA の文献として意味のある文章となることがわかった.

4　結論

対象文献に対して不要語を除いた頻度上位の単語，およびある程度の距離までの頻度上位の組合せに着目することによって専門用語，固有の表現の抽出が可能であった.

本手法を翻訳システムに用い，全文におけるデータから専門用語を辞書に追加登録し翻訳を行えば，専門分野の英文をある程度きれいに翻訳することが可能である事がわかった.

参考文献

1) 松澤裕史．テキストデータからの頻出パターンのマイニング．知識発見のための自然言語処理シンポジウム, 1999.