3G-7

対訳テキストを利用した訳語選択のための共起関係の自動抽出

中島 弘之 ・ 梶 博行
日立製作所 関西システムラボラトリ

1．はじめに

機械翻訳システムの実用性向上のため、翻訳実行時に共起関係知識を自動的に獲得する機能の実現が望まれている。共起関係は日本文の係り受け解析・深層格解析や英文生成での訳語選択に有効である[1]。しかし、共起関係は市販の辞書などに整理されていない。このため、翻訳システムで共起関係を利用して、実際に効果を発揮するには、学習機能が必須であるといえる。

本報告では、後編集結果から訳語選択のための共起関係を学習する機能を提案し、日英対訳テキストから英語共起関係を自動抽出する実験を行なった結果について述べる。

2．訳語選択のための共起関係の自動抽出

機械翻訳での訳語選択には、共起関係の利用が最も有効な方法の１つである。しかし、膨大な量の共起関係を翻訳システム作成時にすべて用意しておくことは不可能であるので、システムが共起関係を自動的に学習する機能を実現することが重要である。

訳語選択に用いる共起関係は目的言語内の知識である。目的言語に関する学習には、出力文の後編集結果が利用できる。たとえば、後編集結果を解析し、曖昧性のない共起関係を抽出する方式[1]を適用すれば、目的言語の共起関係が自動抽出できる。また、西田らは、日英翻訳システムの出力英文の後編集結果を英日翻訳システムで解析して、英文生成に関する知識を獲得する手法を提案している[2]。

しかし、後編集文を解析する方式には、次のような問題点がある。第一に、翻訳システムは一般に単方向であるので、目的言語の解析機能は持っていない。知識獲得のためだけに目的言語解析機能を付加することは、システム開発にとって大きな負担である。第二に、双方向システムであっても、後編集文の解析による学習を行なえば、解析・生成に要する時間は等しく、かつ、言語に依存しないと仮定しても、翻訳に要する時間は、学習をしない場合の１．５倍になる。

後編集文を解析せずに訳語選択の知識を獲得する方法として、システムの与える訳語候補をユーザが選択し、選択された訳語をそれ以後の翻訳で利用する、という学習機能がある。しかし、実際の後編集では、単語を単に置き換えるだけではなく、出力文を文字列として書き換えることが通常行なわれる。すなわち、実際的な後編集文では、原文との単語間の対応関係が失われてしまうので、この方法は実用的ではない。

3．英語共起関係自動抽出方式

本報告では、①後編集文を解析せず、②原文と後編集文の単語間の対応関係を復元し、復元した対応関係を利用して、原言語の共起関係を目的言語に変換するという共起関係学習方式を提案する。

日英対訳テキストを利用して、提案方式の原理実験を行なった。英語共起関係の抽出は、対訳英文と日英対訳辞書を利用して、対訳文中の単語間の対応関係を同定し、日本文解析結果から抽出した共起関係を英語に変換することにより行なう。具体的には、図１に示すように、以下の（１）～（３）の処理を実行する。

（１）日本語共起関係の抽出：日本文を解析し、解析結果から曖昧性のない日本語共起関係を抽出する[1]。

（２）対訳英文の単語分割：（１）で解析した日本文の対訳英文を単語に分割する。

（３）共起関係の日英変換：日本語共起関係を構成する２つの単語の訳語候補を、日英対訳辞書から検索する。これらの訳語候補と（２）で分割した単語とのマッチングをとり、単語の対訳関係を同定する。マッチした単語の見つかった訳語候補を正しい訳語と判断して、この英単語によって英語共起関係を構成する。

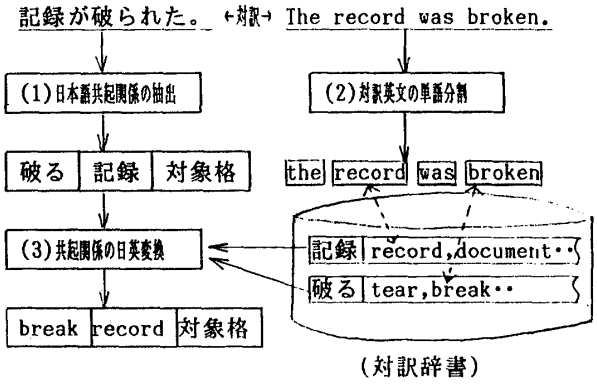

図１．英語共起関係自動抽出方式

Automatic Acquisition of Knowledge of Word Co-occurrences from Bilingual Texts
Hiroyuki NAKAJIMA and Hiroyuki KAJI
Kansai Systems Laboratory, HITACHI Ltd.

4．実験の設定条件と結果

4．1　設定条件

①例文：英語の文法書中の日本語・英語それぞれ２００文の対訳文を用いた。

②対訳辞書：平均訳語登録数（日本語１単語当りの平均訳語数）が約５個の日英対訳辞書を使用した。

4．2　実験結果

表１に実験結果を示す。表１には、日本語共起抽出機能で抽出した共起関係のうち、英語に正しく変換されたものの割合と、変換されなかった共起関係をその原因別に分類した結果を示している。

4．2．1　共起関係抽出の失敗

（1）訳語の未登録：「目撃する」－‘see’ のような対訳関係が日英対訳辞書に登録されていなかった。

（2）連語：‘put on’のように、訳語が連語であって、かつ英文中では‘put the record on ’のように分離して使われるものは、抽出できなかった。本方式では、連語は英文中で連続して現われるもののみを扱っているためである。

（3）誤変換：対訳辞書に登録された日本語単語の複数の訳語候補が対訳文中に含まれているため、誤った訳語に変換されてしまうケースが想定されるが、本実験では、このようなケースはなかった。

4．2．2　共起関係抽出の対象外

以下の２つのケースは、日英間で共起関係の対応関係がない場合であり、共起関係抽出の対象外である。

（1）慣用句：「手に入らない」－‘short off ’のように、「手に入る」という慣用句が辞書に未登録であったため、「手」と「入る」の共起関係が抽出されたが、対応する共起関係は英文中には存在しない。

（2）文生成における日英間での発想の相違：

・先生がいなかったので、生徒はいたずらを始める機会を得た。

・ The teachers absence provided a oppotunity for the pupils to get into mischief.

上記の例では、「なる型」の動詞「得る」を用いた日本文が、「する型」の動詞‘provide’を用いた英文に対応している。このような場合は、「得る」の訳語として‘provide’を対訳辞書に登録するのは不適切であるので、共起抽出の対象外とする。

5．考察

5．1　英語共起関係抽出能力

本方式の英語共起関係抽出能力は、英文を解析して共起関係を抽出する方式と同等であると結論できる。その理由は以下のとおりである。

表１より、対訳辞書に登録された訳語の範囲内では、単語のマッチング方式を連語に対処できるよう改良すれば、ほぼ１００％の単語の対訳関係が同定できることがわかる。単語間の対応関係が、対訳辞書に登録されている訳語の範囲内で１００％同定できれば、本方式の共起関係抽出能力は、目的言語を解析する方法と実質上等価である。目的言語解析を行なって共起関係を抽出しても、その単語が対訳辞書に登録されていなければ訳語が選択できないからである。

5．2　訳語の未登録についての対策

文中の他の単語の対訳関係や、品詞の対応などから対訳関係をある程度自動的に決定できる。また、翻訳システムでは、対象とする分野によって辞書を調整する必要がある。本方式は、実テキストからの学習方式なので、単語の対訳関係決定機能を付加すれば、必要十分な訳語のみを学習できるという効果が期待できる。

5．3　逐語訳翻訳の限界

表１より、共起関係抽出の対象外のケースは、全体の約２０％になることがわかる。したがって、対象とするテキストに依存するが、約８０％の文は、慣用句や意訳を用いない逐語訳的な方法で機械翻訳可能であると見積もることができる。

6．おわりに

日英対訳テキストと対訳辞書を利用して、英文解析を行なわずに英語共起関係を抽出する方式の実験を行なった。その結果、①本方式は、英文解析を行なって共起関係を抽出する方式と同等の能力を持つこと、②約８０％の文は逐語訳によって翻訳可能であることがわかった。以上の結果から、後編集結果からの共起関係学習機能の実現可能性が確認できた。

表１．　英語共起関係抽出実験結果

		件数	成功/失敗率(%)	割合(%)
成功		102	82.3	65.4
失敗	訳語の未登録	19	15.3	12.2
	連語	3	2.4	1.9
	誤変換	0	0.0	0.0
	計	22	17.7	14.1
成功・失敗の合計		124	100.0	79.5
対象外	慣用句	15		9.6
	発想の相違	17		10.9
	計	32		20.5
総計		156		100.0

参考文献

[1] 中島、梶：テキストからの共起関係自動抽出の試み、情報処理学会第３８回全国大会論文集、pp. 325~326 (1989).

[2] F. Nishida, et al. : Feedback of Correcting Information to a Machine Translation System, COLING-88, pp. 476~481 (1988).