

用例に基づいて英語前置詞句の係り先決定を行う英日翻訳システム

5B-2 隅田 英一郎*, 土井 伸一**, 飯田 仁*, 山端 潔**

*ATR自動翻訳電話研究所 **NEC C&C情報研究所

1はじめに

自然言語処理における最も困難な問題の一つに、構造的曖昧性の解消がある。例えば、前置詞句の係り先は典型的な構造的曖昧性を引き起こす。

(1) I present a paper at the conference.

例文(1)の前置詞句「at the conference」は、動詞「present」、名詞「paper」の両方を修飾しうる。一般に前置詞句の係り先は構文規則による解析だけでは一意に決定することが困難である。この例では前者の方が自然であるとしてよいだろう。例文(1)の翻訳を考えてみる。二つの係り先の相違は、それぞれ「会議で論文を発表する」、「会議での論文を発表する」と異なった日本語訳になるが、やはり前者の方が尤もらしい。このように前置詞を含む文を正しく翻訳するためには前置詞句の係り先の曖昧性を解消する必要がある。既に、構文的な情報、統計的な情報などを利用して複数の係り先候補から尤もらしいものに絞り込む手法が数多く提案されている。

筆者らは対訳コーパスから用例（原言語表現とその対訳の対）を抽出し、入力表現と用例との間の意味的距離をシソーラスに従って計算し、最小距離の用例に基づいて訳語選択を行なう手法（Example-Based Machine Translation、EBMT）を提案している。同手法は英日翻訳における前置詞などの機能語に対して高い翻訳正解率を達成している[1]。本稿では同手法を応用して、コーパスから抽出した用例

（前置詞句とその係り先）とシソーラスに従った意味的距離（の総和）・頻度に基づいて最適な前置詞句の係り先を決定する手法（Example-Based Disambiguation、EBD）を提案する。EBDは、ATRコーパスを用いた実験で良好な結果を得た。また、基本的な解析・生成は規則主導で行ない、英語前置詞の翻訳（係り先決定・訳語選択）は用例主導で行なう英日機械翻訳システムREALIST（Rule + Example = A LinguIST）を作成した[2]。

2 従来手法

2.1 構文的な手法

なるべく近くの句への係りを優先するRight Associationとより大きな構造をカバーする規則を優先するMinimal Attachmentの二つの構文的な情報のみを使う方法[3]が広く知られている。これらの手法は単純で一般的な記述となつ

ており、特別な知識を必要としない利点はあるが、正解率が悪いことが最近、指摘されている[4,5]。

2.2 文法規則や辞書に基づく手法

現在、広く行なわれている手法は、文法規則や辞書に意味マーク等を用いて、係り先の優先情報を指定する方法である。しかし、この手法では、動詞の必須格は別として、動詞の自由格や名詞に前置詞句が係る場合を扱いきれない。

2.3 共起頻度に基づく手法

構造的な曖昧性の解消のために単語の共起頻度を利用する方法がいくつか提案され、構文的な方法より高い正解率が報告されている。堤等[6]はコーパスから三つ組、動詞（あるいは名詞）・前置詞・名詞の共起頻度を収集し、解析木中に現われる三つ組の頻度の総和をその解析木の尤度とする方法（以下ではStatistically-Based Disambiguation、SBDと呼ぶ）を提案している。

本稿では正解率が高いSBDとEBDを比較する。

3 用例に基づく係り先決定手法

係り先の候補数をn、係り先の候補（動詞または名詞）を $x_i (1 \leq i \leq n)$ 、前置詞をp、pの目的語（名詞）をyとする。問題は入力「 x_1, \dots, x_n, p, y 」に対して「p, y」の最尤の係り先「 x_k 」を選ぶことである。例文(1)では、入力は「present, paper, at, conference」で、正解は「present」である。以下のアルゴリズムはTDMT[7]の構造的曖昧性解消のアルゴリズムを頻度情報によって改良したものに相当する。

3.1 係り先決定のアルゴリズム

0)次の処理を繰り返す ($1 \leq i \leq n$)。

全用例と x_i, p, y の距離をEBMTの式[1]に従って計算し最小距離の用例を検索する（近似照合）。最小距離 d_i と同一距離の用例の頻度 f_i を記憶する。

SBDの場合：全用例から x_i, p, y と同一の用例を検索する（完全照合）。距離 d_i （用例あれば0なければ∞）と用例の頻度 f_i を記憶する。

- 1) d_i が最小の x_i が1つなら x_i を係り先として返し終了する。
- 2) d_i が最小の候補で f_i が最大である x_i が1つなら x_i を係り先として返し終了する。
- 3) 係り先未決定として終了する。

3.2 データ[1,2]

実験対象は頻度順上位9個の前置詞、「of」、「to」、「for」、

Example-Based Disambiguation of PP-Attachment in REALIST

Eiichiro SUMITA*, Shinichi DOI**, Hitoshi IIDA* and Kiyoshi YAMABANA**

*ATR Interpreting Telephony Research Laboratories

**NEC C&C Information Technology Research Laboratories

"in"、"on"、"at"、"from"、"by"、"with"である。用例は訳し分け用に収集した約3000件の対訳用例の英語側を用いる。シソーラスはLONGMAN LEXICONに準拠した。

4 実験結果

4.1 統計的な手法との比較

表4に比較実験結果を示す。三つ組「x p y」の代わりに二つ組「x p」だけ参照する場合も比較した。

第3.1節のアルゴリズムは候補を一意に決定できない場合もあるので、2段階の評価基準、すなわち、決定できたか（決定率）、正解か（正解率）を用いた。

まず、SBDを見ると、三つ組の方が二つ組より決定率が低い。共起単位が大きいほど類度データが集まりにくいからである。

EBDはSBDに比べて決定率、正解率共に優れている。これはEBDは近似照合の効果で相対的に少ないデータで良い性能が出せることを意味する。

表4 SBDとEBDの正解率（決定率）

	x p y	x p
SBD	24%(25%)	73%(82%)
EBD	86%(98%)	83%(95%)

4.2 失敗例の分析と対策

- ・類似用例の欠落。これは用例の追加で解消できる。
- ・用例単位の問題。現在の三つ組「x p y」単位では、格の必須性、他の格、修飾語の影響が捉えられない。対策は2つある。1)大きな用例単位の使用、2)辞書や規則による処理のハイブリッド化。
- ・多義語の問題。多義性を除去した用例を用意すれば解消できると思われる。

4.3 規則主導システムとの統合実験

第4.1節の実験はEBD単独のシミュレーションであった。次に、筆者等は規則に基づいた解析システムにEBDを組み込んだ[2]。係り先候補の抽出が文法規則、尤度計算がEBDという役割分担である。すべての前置詞の距離と類度およびバーザが出力する他の尤度（ギャップの有無など）とを総合的に評価し、最尤の係り先を決定する実験を行なった。ここでも同様の高い正解率を得た。

5 考察

EBDは、REALISTで示したように、従来のバーザに容易に組み込むことができる。係り受け非交差などの大局的な制約や動詞の必須格など語彙的な制約と整合させることができる。

EBDは前置詞の係り受けに特化した方法ではないので、用例を収集すれば、他の構造的な曖昧性、例えば、to-不定詞、関係節、従属節などの係り受けにも有効である。

Jensen等[8]の辞書定義文を利用する手法では構文木のパターンから意味関係と信頼係数を引き出すヒューリスティクスを前置詞および意味関係毎に開発する必要があるが、EBDでは用例さえ集めれば距離計算で処理できる。

さらに、知識源としてコーパスと辞書が考えられるが、EBDでコーパスを利用したのはドメイン固有の共起関係の必要性からである。既存の辞書情報は一般的で分野固有の情報がない、辞書作成者の興味に従って例文が採録されるなど、偏向があると指摘されている[5]。

今回は用例の収集を自動化していない。梶等[9]は対訳用例の収集を自動化する研究を行なっている。筆者等もこの方向の研究を進める予定である。

6 おわりに

本稿では、前置詞句とその係り先からなる用例とその間の意味的距離に基づいて前置詞句の係り先の曖昧性を解消する手法EBDを提案し、従来手法との比較実験によって正解率の観点から本手法の優位性を示した。また、規則に基づいた解析システムにEBDを組み込んで効果を確認した。

今後は、REALISTの規則・用例の充実をはかると共に、用例収集の自動化を試みる。

参考文献

- [1] Sumita, E., Iida, H.: "Example-Based NLP Techniques - A Case Study of Machine Translation -," *Statistically-Based Natural Language Processing Techniques*, Technical Report W'92-01, AAAI Press, (1992).
- [2] 土井伸一, 隅田英一郎, 飯田仁: "用例に基づいて英語前置詞の訳し分けを行う英日翻訳システム," 情報処理学会第46回全国大会SB-1, (1993).
- [3] Frazier, L., Fodor, J.: "The Sausage Machine: A New Two-Stage Parsing Model," *Cognition* 6, pp.191-325, (1979).
- [4] Whittemore, G., Frrara, K., Bruner, H.: "Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases", 28th ACL, pp.23-30, (1990).
- [5] Hindle, D., Rooth, M.: "Structural ambiguity and lexical relations," 30th ACL, pp.229-2236, (June 1991).
- [6] 堤豊, 堤泰治郎: "統計データに基づいた構文解析のあいまいさ解消方式," 電子情報通信学会論文誌, J72-D-II, 9, pp.1448-1458, (1989).
- [7] Furuse, O., Iida, H.: "Cooperation Between Transfer and Analysis in Example-Based Framework," *Coling '92*, pp.645-651, (1992).
- [8] Jensen, K., Binot, J.: "Disambiguating Prepositional Phrase Attachments by Using On-Line Dictionary Definitions," *Computational Linguistics*, 13, 3-4, pp.251-260, (1987).
- [9] Kaji, H., Kida, Y., Morimoto, Y: "Learning Translation Templates from Bilingual Text," *Coling '92*, pp.672-678, (1992).