訳語選択における手掛かり語の選択手法

綱川 隆司[1]　　丸岡 良徳[2]　　梶 博行[1]
[1]静岡大学情報学部情報科学科 [2]静岡大学大学院情報学研究科情報学専攻
tuna@inf.shizuoka.ac.jp, gs09051@s.inf.shizuoka.ac.jp, kaji@inf.shizuoka.ac.jp

1　はじめに

本稿では、機械翻訳において翻訳対象語に対する訳語の選択時に手掛かりとなる語を選択する手法について比較検討を行う。機械翻訳における翻訳対象語の訳語の適切な選択は、翻訳結果の妥当性、可読性を向上させるのに不可欠である。訳語選択に影響する要素の一つとして、翻訳対象語周辺の語の分布がある。従来の機械翻訳手法においても周辺の語は部分的には考慮されているが、本稿では翻訳対象語の周辺に現れやすい語と訳語候補との間のスコアを用い、スコアの高い語を手掛かり語として選択し、これを用いて訳語の選択を試みる。このスコアを求めるための単語間関連度として相互情報量をはじめいくつかの指標を比較し、実際に訳語選択を行って評価を行う。

2　関連語－訳語関連行列の推定と訳語選択

翻訳において、複数の意味を持つ語の翻訳を行う場合、それぞれの意味に対応する訳語が異なるため、語義曖昧性の解消をし、適切な意味を持つ訳語を選んで出力する必要がある。機械翻訳においてこの問題に対処するには、入力文やその分野等の情報を考慮しなければならない。しかし､ルールに基づく機械翻訳システムでは、どの訳語を選択するかのルールを基本的に人手で記述しなければならず、多大な労力を要するだけでなく、相互に干渉する複雑なルールを管理するのも困難である。一方、統計的機械翻訳などのテキストデータに基づく方法では、データから得られた確率や言語モデル等の情報により訳語の選択を行っているが、より広範囲の情報は考慮されておらず、またあまり用いられない意味の訳語が選択されにくい傾向がある。

Kaji and Morimoto (2002) では、入力語について、ある訳語候補をすべきときに周辺に現れる手掛かり語（関連語）を求め､「関連語－訳語関連行列」を求めることでこの問題に対処している。本研究では、この手掛かり語の絞り込み手法について比較および評価を行う。

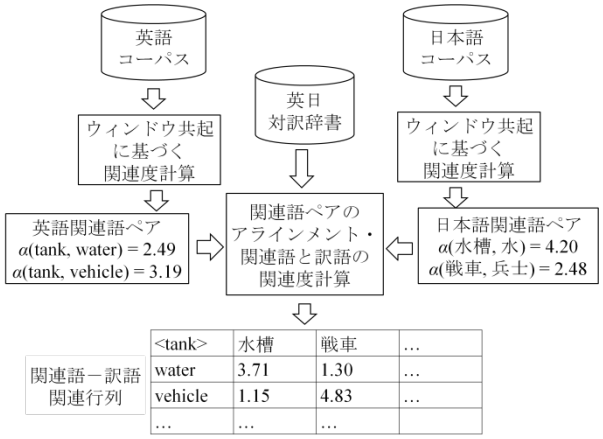

図 1　関連語－訳語関連行列の計算

2.1　関連語－訳語関連行列

図 1 に関連語－訳語関連行列の計算手法を示す。

まず入力言語および出力言語それぞれの単言語コーパスに含まれる名詞を列挙し、全ての名詞対について共起頻度に基づく指標 α を計算する（2.3 節参照)。各名詞について、指標が閾値 θ 以上でかつ上位 t 個の名詞を関連語として抽出する。

次に、関連語－訳語関連行列を以下のように計算する｡「相互に関連のある関連語は同じ訳語を支持する」という仮説に基づき、対象語 f の第 i 関連語 $f'(i)$ と、f の j 番目の訳語 $e(j)$ の関連度$C_f^{(n)}\left(f'(i), e(j)\right)$を以下の反復計算で得る。

$$C_f^{(n)}\left(f'(i), e(j)\right) = \alpha(f'(i), f) \times \frac{\sum_{f'' \in A(f, f'(i))} \alpha(f'(i), f'') \cdot C_f^{(n-1)}(f'', e(j))}{\max_k \sum_{f'' \in A(f, f'(i))} \alpha(f'(i), f'') \cdot C_f^{(n-1)}(f'', e(k))}.$$

ただし、n は反復計算のサイクル、$A(f, f'(i))$は対象語 f と関連語 $f'(i)$に共通の関連語の集合である。すなわち、

$$A(f, f'(i)) = \{f'' \mid \alpha(f, f'') \geq \theta, \alpha(f'(i), f'') \geq \theta\}.$$

反復計算の初期値は以下の式で求める。

$$C_f^{(0)}(f'(i), e(j)) = \begin{cases} \dfrac{a(f'(i), e(j))}{\sum_k a(f(i), e(k))} & (\Sigma_k a(f(i), e(k)) \neq 0), \\ 0 & (\text{それ以外}) \end{cases}$$

$$a(f'(i), e(j)) = \begin{cases} 1 & (\exists e'.\ (f, f'(i)) \approx (e(j), e')) \\ 0 & (\text{それ以外}) \end{cases}.$$

ここで、$(f, f'(i)) \approx (e(j), e')$は、$f$ と $e(j)$、および $f'(i)$と e'がそれぞれ対訳辞書に訳語対として存在することを示す。

以上から、曖昧性なく対訳関係にある関連語ペアの組が種となって、関連語と訳語の間の関連度が反復計算される。

2.2 訳語選択

ある文$\mathbf{f} = f_1 f_2 \cdots f_I$に含まれる単語$f_i$に対して訳語 $e(1), e(2), \ldots, e(J)$があるとき、訳語 $e(j)$に対するスコアを以下で定義する。

$$\mathrm{Score}(f_i, e(j)) = \sum_{\substack{l=i-w_1 \\ (l \neq i)}}^{i+w_1} r(i, l) C_f(f_l, e(j)),$$

$$r(i, l) = \begin{cases} 1 & (|i - l| \leq w_0) \\ \dfrac{1}{\sqrt{|i - l|}} & (\text{それ以外}) \end{cases}.$$

ただし、w_0, w_1はウィンドウサイズとする。また、ある$f_l, e(j)$の組について$C_f(f_l, e(j))$の値が求められていない場合は 0 を代入する。

訳語のうち、最もスコアが高い訳語を選択する。また、全ての訳語についてスコアが 0 の場合は出力なしとする。

2.3 手掛かり語の選択

本研究では、訳語選択の手掛かり語である行列における関連語の選択方法を、共起頻度に基づく指標および共起頻度の計算方法の観点から比較検討する。

2.3.1 共起頻度に基づく指標

コーパスに含まれる名詞 x, y について、x と y が内容語 $w - 1$ 語以下を挟んで出現するとき、x と y が共起すると定義する（wはウィンドウサイズ）。x と y の出現回数をそれぞれn_1, n_2、共起回数を m、コーパスに含まれる単語数を N、2-グラムの総数を M とする。共起頻度に基づく指標αに以下を用いる。

相互情報量 (MI) (Church and Hanks, 1990)

$$\mathrm{MI}(x, y) = \log_2 \frac{m/M}{(n_1/N)(n_2/N)}$$

対数尤度比 (LLR) (Dunning, 1993)

$$\mathrm{LLR}(x, y) = -2\big(\mathrm{logL}(m, n_1, r) + \mathrm{logL}(n_2 - m, N - n_1, r) - \mathrm{logL}(m, n_1, r_1) - \mathrm{logL}(n_2 - m, N - n_1, r_2)\big);$$

$$\mathrm{logL}(k, n, r) = k \log_2 r + (n - k) \log_2 (1 - r),$$

$$r_1 = \frac{m}{n_1}, r_2 = \frac{n_2 - m}{n_1}, r = \frac{n_2}{N}.$$

t-スコア (TScore) (Church et al., 1991)

$$\mathrm{TScore}(x, y) = \frac{m - n_1 n_2 / N}{\sqrt{m}}$$

Dice 係数 (Smadja, 1993)

$$\mathrm{Dice}(x, y) = \frac{2m}{n_1 + n_2}$$

Jaccard 係数 (Smadja et al., 1996)

$$\mathrm{Jaccard}(x, y) = \frac{m}{n_1 + n_2 - m}$$

Pearson's χ^2指標 (Manning and Schütze, 1999)

$$\chi^2(x, y) = \frac{N(mN - n_1 n_2)}{n_1 n_2 (N - n_1)(N - n_2)}$$

また、複数の指標の特長を組み合わせるため、相互情報量と対数尤度比、および相互情報量とt-スコアを組み合わせた指標を以下で定義する。

MI&LLR

ある名詞xについて、各名詞yの相互情報量、対数尤度比での順位をそれぞれr_1, r_2とする。各名詞を$\max(r_1, r_2)$の昇順で並び替え、上位 t 個以外を取り除く。指標の値として相互情報量の値を用いる。

MI&TScore

MI&LLR について、対数尤度比を t-スコアで置き換えたもの。

2.3.2　共起頻度の計算方法

出現回数が少ない語は共起する名詞も少なくなり、ウィンドウサイズが一定の条件下ではデータが疎になる。本研究では出現回数に応じてウィンドウサイズを変化させて共起回数を計算する以下の２つの手法について比較を行う。

ウィンドウ拡張

出現頻度 n の名詞について、ウィンドウサイズw'を以下の式で定義する。

$$w' = \begin{cases} \lfloor w + \log_2(n_t - n) + 1 \rfloor & (n < n_t) \\ w & (n \geq n_t) \end{cases}$$

ただし、n_tは出現頻度の閾値とする。

ウィンドウ拡張（頻度重み付けあり）

ウィンドウ拡張後、２つの名詞が距離 d で共起する際、共起回数を１回の代わりに重み付きの回数$\min(w/d, 1)$回をカウントする。

3　実験

3.1　実験設定

英語および日本語のコーパスから関連語ー訳語関連行列を各手法により求め、英文に含まれる各名詞の訳語を出力して評価する実験を行った。

関連行列を計算するためのコーパスには、English GigawordコーパスのNew York Times（2004 年, 277MB）および毎日新聞コーパス（2004 年,140MB(UTF-8)）を用いた。関連行列の反復計算に用いる対訳辞書にはEDR電子化辞書の英日・日英対訳辞書、EDICT (Breen, 1995) および英辞郎を組み合わせたものを用いた。評価対象としてNew York Times（2005 年 1 月）の 157 パラグラフに含まれる延べ 1448 語に対して訳語の出力を行った。1448 語のうち、周辺に手掛かりとなる関連語が一語も出現しない場合、および関連行列が存在しない語（191 語 [1]）については出力なしとした。

訳語の出力結果に対して、正解（1 点）、一部正解（0.5 点）および不正解・出力なし（0 点）の 3 段階で人手による評価を行った。評価は 1 語につき 2 名で行い平均を最終的なスコアとした。実験に用いたパラメータは以下の通りである。

$$w = 10, n_t = 20, t = 400, w_0 = 5, w_1 = 25$$

表 1　実験結果

手法	平均スコア	出力語	関連語なし
相互情報量	0.31	714	543
対数尤度比	0.43	1190	67
t-スコア	0.41	1203	54
Dice	0.49	1166	91
Jaccard	0.48	1166	91
Pearson's χ^2	0.26	714	543
MI&LLR	0.30	690	567
MI&TScore	0.12	359	408
ウィンドウ拡張	0.31	717	540
頻度重み付け	0.31	716	541

3.2　実験結果

2.3 節で述べた各手法を適用した結果を表 1 に示す。Dice 係数および Jaccard 係数を用いた場合に最も高いスコアを得た。

3.3　実験結果の検討

相互情報量や Pearson's χ^2指標を用いるのに比べ、他の指標を用いることで関連語が存在しないために出力できなかった語が大幅に減少した。相互情報量では低頻度語に対して高い値が割り当てられる傾向があるため、訳語選択の際に手掛かりとなる周辺の語と関連語が一致しないためと考えられる。

平均スコアでは Dice 係数および Jaccard 係数が最も高く、およそ 39%の語について正解の訳語を選択できた。また、出力語数では t-スコアが最も多かった。

相互情報量を対数尤度比と組み合わせる手法では、スコアの変化はみられなかった。また、t-スコアとでは悪化した。組み合わせに用いた閾値、および Dice 係数など他の手法との組み合わせは今後の課題である。

共起ウィンドウの拡張、および頻度の重み付けについては、結果にほぼ変化が見られなかった。ウィンドウ拡張の対象となった低頻度語が

[1] MI&TScore については 681 語。

結果にほぼ寄与できなかったと考えられるが、より広いウィンドウや分野情報の導入、また Dice 係数での拡張が改善案として挙げられる。

4 関連研究

単言語における語義曖昧性解消は、辞書やコーパス等のデータを使って教師なしで学習を行う手法が提案されてきた (Ide and Veronis, 1998)。品詞等の文法的情報、構文的関係にある語、および周辺にある分野に関する語を曖昧性解消の手掛かりとして用いている。本研究ではこれらの一部を利用して、単語の翻訳の曖昧性解消に用いている。

Li and Li (2002) は単語の翻訳の曖昧性解消を対訳辞書の対応付けをもとにブートストラッピングによって分類器を構築し行っている。本研究では辞書の対応関係は反復計算の種として用いコーパスから求めた手掛かり語を考慮に加えている。

Vickrey et al. (2005) では統計的機械翻訳に文脈を考慮した訳語選択を素性として導入し、訳語の選択を試みている。文全体の統計的機械翻訳システムへの導入が大きな課題である。

5 おわりに

本稿では関連語ー訳語関連行列を用いた訳語選択において、手掛かり語となる関連語の選択手法の比較検討を行った。共起頻度に基づく指標では Dice 係数および Jaccard 係数を用いた場合に最もよい結果が得られた。また、指標の組み合わせや共起ウィンドウの拡張手法では性能の向上が見られなかった。

今後の課題としては、Dice 係数を中心とした指標の組み合わせや低頻度語への対処による性能の改善、および文全体の機械翻訳システムへの導入と文全体での翻訳結果の評価が挙げられる。

謝辞

本研究の一部は、科学技術振興調整費・重要課題解決型研究等の推進「日中・中日言語処理技術の開発研究」の助成を受けています。

参考文献

Breen, J.W. 1995. Building an Electronic Japanese-English Dictionary. In *Proc. of the Japanese Studies Association of Australia Conference*.

Church, Kenneth W., William Gale, Patrick Hanks and Donald Hindle. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115-164.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22-29.

Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1-40.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora. In *Proc. of the 19th International Conference on Computational Linguistics*, pages 411-417.

Li, Cong and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 343-351.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.

Smadja, Frank, Kathleen R. McKeown and Vasileios Hatzivassiloglou. 1996. Translating Collocations for Bilingual Lexicons: A Statistical Approach. In *Computational Linguistics*, 22(1):3-38.

Vickrey, David, Luke Biewald, Marc Teyssier and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proc. of the Conference on HLT/EMNLP*, pages 771-778.