ＮＩＣＴ日中対訳コーパスにおける単語アライメント

張 玉潔[*1]　馬 青[*1*2]　井佐原 均[*1]
[*1]独立行政法人　情報通信研究機構
[*2]龍谷大学
[*1]{yujie,qma,isahara}@nict.go.jp

1　はじめに

単語アライメントは対訳コーパスから翻訳知識を抽出するための重要な技術である．本稿では、NICT日中対訳コーパスにおける単語アライメントの手法について述べる.

単語アライメントは、原文とその訳文に対し、対訳関係にある単語間に対応付けを行うことである．単語アライメントされたコーパスは機械翻訳、翻訳辞書の編集、多義解消など様々な研究課題に利用できる．これまで、単語アライメントに関する研究が多数行われており、いくつかの手法が提案されてきた．統計情報に基づく手法についていえば、大量の対訳データさえあれば利用できるが、出現頻度の低い単語に対しては精度が低い．この問題に対して Ker が語彙知識を利用する手法を提案した[1]．Ker のアイディアを拡張すれば、様々な言語知識を利用する試みも考えられる．しかしながら、これまでの研究は主に英語を中心に行われてきており、日中両言語間の単語アライメントに関する本格的な研究はあまり見受けられない.

本研究は NICT 日中対訳コーパスプロジェクト[2]の一環である．本プロジェクトの目標は、単語アライメントも含めた様々な情報を付与した大規模日中対訳コーパスを構築することである．コーパスは 38,383 文対の日本語原文と中国語訳文から構成される．現在、日本語文に形態素解析と構文解析の情報を付与する作業と、中国語文に単語分割と品詞タグを付与する作業が進んでいる.

本稿では語彙情報と位置情報に基づく単語アライメント手法を提案する．提案手法は、NICT 日中対訳コーパスを対象に既存の統計手法との比較実験を通じて評価する．評価結果に基づき、両手法の長所を取り入れられるような統合的な単語アライメント手法を提案し、評価実験を行う.

2　語彙情報と位置情報に基づく手法

本手法は二段階の手順、語彙情報に基づくアライメントと位置情報に基づくアライメントからなる[3]．ここで、対訳文対に対し、日本語文の形態素列をW_J 、中国語文の単語列をW_C で表す.

2.1　語彙情報に基づくアライメント

W_Jの中のある形態素 j に対しW_C の中の（対訳関係にある）対応付けている単語 $\hat{c}$ を見つけるには、jとW_Cの中の個々の単語 cとの対応の可能性を推定して、その中からもっともらしいものを決定すればよい．ここでは、 jとc の対応可能性をスコア $S(j,c)$ で表す．このスコアの推定は、 j の中国語訳語とc の類似度を測ることにより行う．一般的に、文字列 x と y の類似度は式（1）のDice式で計算する.

$$Sim(x,y)=\frac{2\times|x\cap y|}{|x|+|y|}. \qquad (1)$$

以下では、スコア $S(j,c)$ の推定に使われる三種類の語彙情報を紹介し、推定の詳細を述べる.

翻訳辞書　辞書から j の中国語訳語を取り出す．その訳語の集合をC_j で表す． $S(j,c)$ の推定には、すべての訳語 $c'(\in C_j)$ と中国語単語 c との類似度を計算して、その中から類似度の一番高いものの値をスコアとする．この推定は式(2)で示し、このように翻訳辞書により推定したスコアをS_Dで表す.

$$S_D(j,c)=\max_{c'\in C_j} Sim(c',c). \qquad (2)$$

翻訳辞書としては、EDR 日英辞書とLDC の英中辞書から英語を介して自動的に構築した日中辞書[4]を利

用した. したがって、本研究は、自動構築した辞書の検証も兼ね合わせている.

実際、翻訳辞書はまだ完全に整備されていないため、中国語訳語のない日本語単語が数多く存在する. そのために、スコアの推定に以下に述べる漢字情報も利用することにした.

漢字の字形情報　日本語において約半分の単語に漢字が含まれている. このような漢字単語の一部は中国から伝わってきたもの、あるいは、日本から中国へ伝わったものである. このような日本語漢字単語はその中国語訳語と字形的に同じものである. 例えば、日本語単語「人民」と「国家」はその中国語訳語もそれぞれ「人民」と「国家」である. この由来から、字形情報を利用し、スコアを式(3)のように推定する. このように字形情報により推定したスコアをS_Oで表す.

$$S_O(j,c) = Sim(j,c). \qquad (3)$$

漢字の簡体字と繁体字　中国語の漢字は、その繁体字が文字改革で簡略化され、現在簡体字が使われている. 日本語の漢字単語の中の一部は、中国から伝わってきた時点からその字形が繁体字のままで変わっていない. このような漢字単語とその中国語訳語の間には繁体字と簡体字の対応関係がある. 例えば、単語「故郷」とその中国語訳語「故乡」には、一番目の文字が同じで、二番目の文字「郷」と「乡」が繁体字と簡体字の対応関係にある. 以上のような観察から、単語cはもしその繁体字がjと同じなら、cはjの訳語である可能性が高いと思われる. この漢字の簡繁関係を利用し、スコアを式(4)のように推定する. 推定したスコアをS_Tで表す.

$$S_T(j,c) = Sim(j,c^T). \qquad (4)$$

c^Tは、中国語の繁体字と簡体字の対応関係表を利用し、単語cの各文字を繁体字に直したもの（以降「繁体字単語」と呼ぶ）である.

以上で述べた三つの推定スコアから最大のものを選んで、jとcの対応可能性を表す最終スコア$S_L(j,c)$とする. すなわち、

$$S_L(j,c) = \max(S_D(j,c), S_O(j,c), S_T(j,c)). \quad (5)$$

次に、アライメントの具体的な手順を記述する.

入力: W_JとW_C

出力: 対応付けの単語対の集合A_L

(1) W_Jの各形態素jに対し、翻訳辞書からC_jを得る.

(2) W_Cの各単語cに対し、その繁体字単語c^Tを得る.

(3)すべての単語対jとcに対し、式(2)、(3)、(4)と(5)より、

S_D, S_O, S_TとS_Lを計算する.

(4)各jに対し、条件$\max_{c\ in\ W_C} S_L(j,c) \geq \theta_L$を満たすような

$\hat{c}(= \arg\max_{c\ in\ W_C} S_L(j,c))$を求め、$(j,\hat{c})$を$A_L$に出力する. θ_Lは予め設定した閾値である.

ただし、実際の処理においては上述の手順に書いてある「形態素j」は4個以内の連続した形態素の列に拡張して扱っている. また「単語c」は4個以内の連続した単語の列に拡張して扱っている.

2.2　位置情報に基づくアライメント

次は、語彙情報で対応付けできていない日本語形態素列$\overline{W_J}$と中国語単語列$\overline{W_C}$に対し、それらの対応付けを見つける手法について述べる. 原文において同じ構文成分に属する単語が訳されたとき、対訳関係にある訳語も対訳文において同じ構文成分に属することがよく観察される. この現象により、例えば単語j_1と単語j_2が同じ構文成分に属し、j_1が$c_{\hat{1}}$にアライメントされた場合、$c_{\hat{1}}$の位置から、単語j_2が対応する単語の位置を推定することが可能と考えられる. したがって、実際の単語の位置推定は、すでに得られたアライメント情報を利用することによって行うことができる[1]. 以下ではこのようなアイディアに基づきjとcの対応付け可能性のスコアの推定について述べる.

$\overline{W_J}$ の j と $\overline{W_C}$ の c の単語対に対し、A_L から以下のような四つのアライメント(対応付けている単語の対)を選んで、スコアの推定に用いる.

$a_{jL}(a_{cL})$: j (c)の左側に一番近いアライメント

$a_{jR}(a_{cR})$: j (c)の右側に一番近いアライメント

a_{jL} の日本語形態素と中国語単語のそれぞれの文中の位置を m_{jL}^J と m_{jL}^C で表す. 単語 j と c のそれぞれの文中の位置を m_j と m_c で表す. そして、j と c のアライメントと a_{jL} とのねじれ及び離れ具合を以下の二つの数値で表示する.

$$\Delta m_{jL}^J = m_j - m_{jL}^J, \quad \Delta m_{jL}^C = m_c - m_{jL}^C. \qquad (6)$$

Δm_{jL}^J と Δm_{jL}^c の値が小さければ小さいほど、離れ具合が小さく、j と c がそれぞれ a_{jL} の日本語形態素と中国語単語と同じ構文成分に属する可能性が高い. したがって、j と c が対応している可能性が高い. Δm_{jL}^J と Δm_{jL}^c が同じ整数あるいは同じ負数なら、ねじれがない、つまり a_{jL} の日本語形態素と中国語単語がそれぞれ同じく j と c の左側にある、あるいは、同じく右側にある. この場合、j と c が対応している可能性が高い. 一方、Δm_{jL}^J と Δm_{jL}^c の値の符号が相反するなら、j と c のアライメントは a_{jL} とねじれている. この場合、j と c が対応している可能性が低い. したがって、j と c の対応可能性のスコアは以下のように推定される.

$$S_{jL}(j,c) = \frac{2}{(|\Delta m_{jL}| + |\Delta m_{jL}|)e^{|\Delta m_{jL} - \Delta m_{jL}|}}. \quad (7)$$

同様に、残りの三つのアライメントから S_{jR}, S_{cL} と S_{cR} を得ることができる. そして、式(8)のように、その中から、一番高い値を選び、j と c の対応付け可能性を表す最終スコア $S_P(j,c)$ とする.

$$S_P(j,c) = \max (S_{jL}(j,c), S_{jR}(j,c), S_{cL}(j,c), S_{cR}(j,c)). \qquad (8)$$

次は、位置情報に基づくアライメント手順を記述する.

入力 : $\overline{W_J}$, $\overline{W_C}$, A_L

出力 : アラインメント A_P

(1) $\overline{W_j}$ の各形態素 j と $\overline{W_C}$ の各単語 c のすべての単語対に対し、A_L から四つのアライメント a_{jL}, a_{jR}, a_{cL} と a_{cR} を選ぶ.

(2) 選んだ四つのアライメントを用いて、式(6), (7)と(8)より、$S_{jL}, S_{jR}, S_{cL}, S_{cR}$ と S_P を計算する.

(3) 各 j に対し、条件 $\max_{c \text{ in } \overline{W_C}} S_P(j,c) \geq \theta_P$ かつ $S_L(j,\hat{c}) \geq \theta_L (\hat{c} = \arg\max_{c \text{ in } \overline{W_C}} S_P(j,c) > \theta_P)$ を満たすような $\hat{c}$ を求め、$(j,\hat{c})$ を A_P に出力する. θ_P は予め設定した閾値である.

3 評価実験

提案手法を評価するために、NICT 日中対訳コーパスから 1,127 文対を抽出し、人手で単語の対応付けを付与した. 全部で17,332個の単語の対応付けがあ

る. 閾値θ_Lとθ_Pをそれぞれ0.85と0.8に設定した[3]. 評価に正解率、再現率と*F-measure*を用いた.

表1は、提案手法についての各種の単独情報及びそれらの組み合わせを用いた実験結果を示した. この結果から以下のことが分かった. (1)単独情報を用いた場合は、字形のみを用いたとき正解率も再現率も一番高かった. (2) 字形に漢字の簡繁関係を加えた結果、その*F-measure*の値が39.8%から42.3%に向上した. (3) さらに、翻訳辞書を加えた結果、*F-measure*の値が42.3%から51.8%に向上した. この結果により、自動構築した翻訳辞書の有効性も検証できた. (4)位置情報を加えた結果、*F-measure*の値がさらに11.2%向上し63.0%に上昇した.

既存の統計手法 GIZA＋＋ツールを比較実験に用いた. 評価実験には、日本語を原言語とし、中国語を目的言語とする場合の実験(J->C)と、中国語を原言語、日本語を目的言語とする場合の実験(C->J)を行った. その結果も合わせて表 1 に示す. 表からも分かるように、C->J の場合の結果が J->C の場合のそれよりよかった. 提案手法と GIZA＋＋と比較すると、C->J の場合は前者のほうでは正解率が高く、後者のほうでは再現率が高かった. 前者のほうが高い正解率になったのは、語彙情報に基づく手法ではその利用した語彙情報が正しかったからである. 一方、現段階の語彙情報はまだ網羅されていないので、統計手法のほうが、対応付けを統計情報から推定することができ、再現率が高かった.

情報/手法	正解率 (%)	再現率 (%)	*F-measure* (%)
提案手法(語彙情報と位置情報に基づくアライメント)			
字形	98	25	39.8
簡繁関係	97	11	19.8
翻訳辞書	87	19	31.2
字形+簡繁	98	27	42.3
字形+簡繁＋辞書	92	36	51.8
字形+簡繁＋辞書＋位置	69	58	63.0
統計手法(GIZA++ツール)			
J→C	46	55	50.0
C→J	55	73	62.7
提案手法と統計手法の統合的な利用			
多数決の結果	79	63	70.1

表1 提案手法、統計手法及び統合的な手法を評価した結果

以上のことから、語彙情報と統計情報のそれぞれの長所を生かす統合的な手法を試みた. 具体的には、提案手法、GIZA++のC->JとJ->Cの多数決の結果を採用することにした. その結果、再現率が 63%に、正解率が 79%に向上した(表1). したがって、語彙情報と統計情報を同時に利用することが有効であることが分かった.

4 おわりに

本稿では語彙情報と位置情報に基づく単語アライメント手法を提案した. 提案手法は、NICT 日中対訳コーパスを対象に既存の統計手法との比較実験を通じて評価した. 評価結果に基づき、両手法の長所を取り入れられるような統合的な単語アライメント手法を提案し、評価実験を行った. 語彙情報と統計情報を同時に利用することが有効であることが分かった.

参考文献

[1] Ker, S.J., Chang, J.S(1997). A Class-based Approach to Word Alignment. *Computational Linguistics*, 23(2):313-343.

[2] 張 玉潔,馬 青,内元 清貴,井佐原 均 (2005). “NICT 多言語コーパスにおける日中対訳データの構築”,言語処理学会第 11 回年次大会発表論文集, 510-513.

[3] Yujie Zhang and Qun Liu and Qing Ma and Hitoshi Isahara (2005). A Multi-aligner for Japanese-Chinese Parallel Corpora. In 10th MT Summit Proceedings, 133-140.

[4] 張 玉潔, 馬 青, 井佐原 均 (2005). 英語を介した日中対訳辞書の自動構築. 自然言語処理, 12(2): 63-85.