言語内・言語間単語共起情報を併用した対訳文書の文アラインメント

熊野 正　　田中 英輝
NHK 放送技術研究所
{kumano.t-eq, tanaka.h-ja}@nhk.or.jp

1　はじめに

対訳文書内の1対1、あるいは複数対複数文対応（以後単に文対応と呼ぶ）を推定する手法は、主に機械翻訳の学習データを準備する用途のために、古くから数多く研究されてきた。その代表的な手法の1つに、対訳辞書を知識に用い、辞書による単語対応の個数に基づいた指標によって各文対の対訳度を求め、文書全体にわたって指標値を最尤とするよう文対応を決める方法 [2, 4, 3] があり、広く使われている。また、対訳文書対コーパスなど（必ずしも文単位の対応が付与されていなくてもよい）から言語間の単語共起情報を収集し、統計指標によって共起の有意性を検定することで、対訳辞書（統計的対訳辞書と呼ぶこととする）を獲得することも広く行われており、これらの組み合わせにより、トピックなどの単位で対応づけられたある程度の規模の対訳コーパスに対して、その他の言語資源や人手による知識資源を別途必要とすることなく、文アラインメント問題を解くことができる。

単語共起情報から作られた統計的対訳辞書は、言語間で「結びつきの強い」「単語」対の一覧である。このため、複数単語（形態素）表現が含まれる場合や、ある単語と結びつきの強い単語（訳語とは限らない）が相手言語側に複数含まれている場合などに、単語対応数の適切な決定が難しく、対訳度の見積もり精度が悪化する。単語対応数による対訳度の見積もりに代わり、この対訳コーパスから統計機械翻訳の翻訳モデルを学習し、文対の翻訳確率を計算することでこの問題を改善できるが、現状、翻訳モデルの学習には文単位の対訳コーパスが必要であるため、文対応が付与されていないコーパスに対して他の資源を用いずに文アラインメントを推定する場合にはこの方法は採用できない。

本稿では、統計的対訳辞書を獲得するのと同様に収集した言語間単語共起情報と、加えてコーパスの各言語側から収集した言語内単語共起情報を用いることで、他の資源を必要とせずに文対の翻訳確率を近似的に計算し、これを対訳度指標としてより精度の高い文アラインメント推定を行うことができることを示す。我々が保持する、NHK が放送原稿として執筆・翻訳した日英ニュース原稿コーパスに対して、本稿で提案する対訳度指標を用いて対応の交差を許した複数対複数文アラインメントの推定を実施し、得られた文対応データより学習した統計機械翻訳モデルの性能評価を行うことで、提案手法が有用であることを示す。

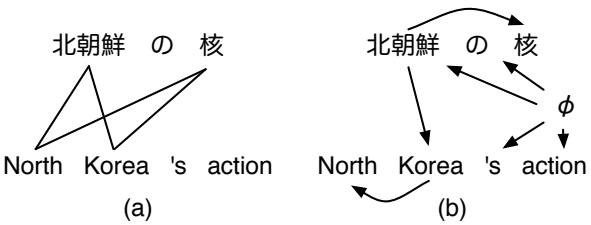

図1　対応度計算のアイディア

2　アイディア

辞書による単語対応個数に基づく2言語単語列対間の対応度の一例として、Utiyama らは以下の指標 $\mathrm{SIM}(J,E)$ を提案している [3][*1]。

$$\mathrm{SIM}(J,E) = \frac{2\sum_{j\in J}\sum_{e\in E}\dfrac{\mathrm{map}(j,e)}{\sum_{e'\in E}\mathrm{map}(j,e')\sum_{j'\in J}\mathrm{map}(j',e)}}{|J|+|E|} \quad (1)$$

ただし、J,E は単語列対の各々を構成する単語からなる重複を許した集合、$\mathrm{map}(j,e)$ は辞書を参照して (j,e) が対訳対であるならば1、さもなくば0を返す関数である。この指標は、辞書に基づいて対間に1対1単語対応を付与した場合に、対応相手ありとなる単語数が対の総単語数に占める割合を意味していると考えられる。

この指標を用いて例えば以下の文対の対応度を求めてみる。

北朝鮮|の|核 ⇔ North|Korea|'s|action

仮に辞書による単語対応が図 1(a) のようであった場合、対応度は 2/7 となる。しかし、

- 北朝鮮 は North|Korea 全体と対応しているのであり、North は対応相手ありと数えるべきでは?
- 仮に 北朝鮮–North|Korea であるならば、これと核–North や 核–Korea は両立するか? もし両立

[*1] Utiyama らは本手法に基づいたものとして公開している文アラインメントプログラム align (http://mastarpj.nict.go.jp/~mutiyama/software.html) において、論文の記述と異なる計算を用いている。本稿では、彼らの提案指標としてこちらを参照する。

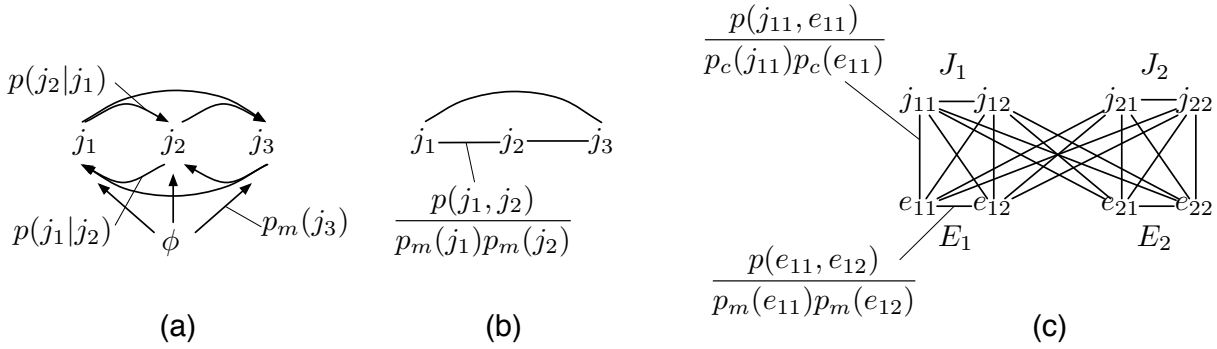

図 2　最大全域木探索による文/対訳対生起確率の計算

しないのならば、それは 北朝鮮–North と 北朝鮮–Korea が両立することと何が違うのか?

を考えると、対訳辞書情報のみを用いて適切な対応度を得ることが難しいことが分かる。

そこで、単語共起情報から統計的対訳辞書を作成する代わりに共起確率を計算して利用することとし、また同一言語の単語間についても単語共起情報を収集して共起確率を求めることで、図 1(b) のような条件つき共起確率でリンクされたグラフの確率を計算することを考える。これは従来の辞書による単語対応づけの確率的な拡張になっているおり、また 2-gram 言語モデルによって単語列生起確率を近似するのと同様に文対生起確率の近似となっているので、この確率値を指標に最尤な文アラインメントを求めることができる。

3　単語共起確率に基づく対訳度指標

3.1　共起頻度情報と単語共起確率

N_J 文の単一言語コーパスの各文について、含まれる単語の異なりからなる単語集合を作成し、同一言語の 2 単語 j_1, j_2 の共起回数 $c(j_1, j_2) =$（j_1, j_2 が共に含まれる単語集合の個数）と定義する。このとき、ある文が与えられたときにその文中に j_1, j_2 が共に含まれる確率（言語内単語共起確率）$p(j_1, j_2)$ は以下の通り。

$$p(j_1, j_2) = c(j_1, j_2)/N_J \tag{2}$$

同様に、N_P 個の 2 言語共起単位対（文対、文書対など）からなる対訳コーパスの各対について、その各々に含まれる単語の異なりからなる単語集合対を作成し、言語の異なる 2 単語 j_1, e_1 の共起回数 $c(j_1, e_1) =$（一方に j_1 が、他方に e_1 が、それぞれ含まれる単語集合対の個数）と定義する。このとき、ある 2 言語対が与えられたときにその対中に j_1, e_1 が共に含まれる確率（言語間単語共起確率）$p(j_1, e_1)$ は以下の通り。

$$p(j_1, e_1) = c(j_1, e_1)/N_P \tag{3}$$

また、$p_m(j_1) = \sum_{j'} p(j_1, j')$, $p_c(j_1) = \sum_{e'} p(j_1, e')$ とする。これらはそれぞれ、（言語内共起情報の共起単位である）1 文内に j_1 が含まれている確率、（言語間共起情報の共起単位である）文書などの中に j_1 が含まれている確率を表す。

3.2　統計検定の利用

j_1, j_2 の生起が独立であるとき、同時生起確率 $p(j_1, j_2) = p_m(j_1)\ \ p_m(j_2)$ である。両値が異なるとき、j_1 と j_2 の生起間には相関があるわけだが、それが有意であるかどうかを統計検定を用いて検定することができる。例えば、対数尤度比 [1] が閾値より大きい時に有意な相関を認める。2 言語単語 j_1, e_1 の相関についても同様である。

ある閾値において対数尤度比検定にて有意な相関が認められ、かつ $p(j, e) > p_c(j)\ \ p_c(e)$ である全ての単語対 (j, e) を、本稿では統計的対訳辞書として用いる。また提案手法においては、同様にある閾値で有意な相関が認められなかった言語内・言語間単語対については、共起確率に独立であるときの値を用いる。これは、低頻度単語の確率が含まれる計算の頑健性を高めるためである。

3.3　単語共起確率による文/対訳対生起確率の近似計算

ある文に含まれるすべての単語の異なりから構成される単語集合が J であるとき、1 文中に J の全要素が含まれる確率 $p(J)$ をこの文の生起確率と見なし、これを単語共起確率を用いて近似する。J の各要素は一般に複数の他要素に依存するが、これを各々 1 要素との関係だけで近似する、すなわち $p(J)$ を以下のように計算することとする。

$$p(J) = \arg \max_{\vec{J}\ (j_1, j_2, \ldots, j_m)} p_m(j_1) \prod_{i=2}^{m} \max(p(j_i|j_1), \ldots, p(j_i|j_{i\ -1})) \tag{4}$$

ただし、$\vec{J}$ は J の全要素をある順番に並べたベクタで、$p(J)$ を最大にする $\vec{J}$ の要素の並びを探索することになる。この探索は、図 2(a) のような有向グラフの全てのノードを覆う最大確率の有向木の探索と等値である。ここで、$m(J) = p(J)/\prod_{j \in J} p_m(j)$、すなわち J が 1 文に含まれる確率の J の各要素が独立に含まれる確率に対する比率を考えると、

$$m(J) = \arg \max_{\vec{J}\ (j_1, j_2, \ldots, j_m)} \prod_{i=2}^{m} \max\left(\frac{p(j_1, j_i)}{p_m(j_1)p_m(j_i)}, \ldots, \frac{p(j_{i\ -1}, j_i)}{p_m(j_{i\ -1})p_m(j_i)}\right) \tag{5}$$

となり、図 2(b) のような無向グラフの最大全域木探索問題となる。これはプリム法などの動的計画法によって高速に計算可能である。

対訳対についても同様に、1 個以上の同一言語の文があって各文の異なり単語集合がそれぞれ $J_1, \ldots, J_m$ であり、また 1 個以上の他言語の文があって各文の異なり単語集合がそれぞれ $E_1, \ldots, E_n$ であったとき、これが 1 つの対訳対として生起する、つまり $J_i (i = 1 \ldots m)$ 中の各要素の存在が J_i 中の他要素もしくは $E_1, \ldots, E_n$ のいずれかの中の要素に依存するとして計算できる確率 $p(\boldsymbol{J} = \{J_1, \ldots, J_m\}, \boldsymbol{E} = \{E_1, \ldots, E_n\})$ を、この複数対複数文対の生起確率と見なす。こちらも同様に、各文の各要素がすべて独立である確率に対する比率 $m(\boldsymbol{J}, \boldsymbol{E}) = p(\boldsymbol{J}, \boldsymbol{E}) / \{\prod_{i=1}^{m} \prod_{j \in J_i} p_c(j) \prod_{i=1}^{n} \prod_{e \in E_i} p_c(e)\}$ を考えると、図 2(c) のような無向グラフの最大全域木がその値となる。

文対の生起確率の各文の生起確率の積に対する比率を、この文対の「対訳度」と呼ぶことにする。すなわち、対訳度 $t(\boldsymbol{J}, \boldsymbol{E})$ は以下のように定義される。

$$\begin{aligned} t(\boldsymbol{J}, \boldsymbol{E}) &= p(\boldsymbol{J}, \boldsymbol{E}) / \{\textstyle\prod_{J \in \boldsymbol{J}} p(J) \prod_{E \in \boldsymbol{E}} p(E)\} \\ &= m(\boldsymbol{J}, \boldsymbol{E}) / \{\textstyle\prod_{J \in \boldsymbol{J}} m(J) \prod_{E \in \boldsymbol{E}} m(E)\} \end{aligned} \tag{6}$$

また、$t(\boldsymbol{J}, \boldsymbol{E}) / \{\sum_{J \in \boldsymbol{J}} |J| + \sum_{E \in \boldsymbol{E}} |E|\}$ を「単語あたり対訳度」と呼び、文対の対訳のよさを表す指標とすることにする。

4 文アラインメントの推定

我々が手法の適用対象に想定している NHK の日英ニュース原稿コーパスでは、日英記事間の情報提示順序の入れ替わりが大きく、従来手法の多くが採用してきた、文対応の交差がないことを前提とした動的計画法による最尤アラインメント探索戦略を採ることができない。本稿では、各原稿が日英たかだか数文程度であることから、簡易な厳密解探索による解法を示すが、より大きな文書のアラインメントに際しては、何らかの効率的な近似解法を用いる必要があろう。

4.1 文対候補の列挙

各々 m, n 文の集合からなる 2 言語文書対 $(\{J_1, \ldots, J_m\}, \{E_1, \ldots, E_n\})$ に対して、その各言語側の任意の部分集合の対（ただし少なくとも一方は空でない）が、この文書対の文アラインメントを構成する文対となりうる。従って、最尤アラインメントの探索に先立ち、まず可能な文対候補を列挙し、各々の対訳度を計算する。なお、一方が空である文対候補は対応先のない文の認定に必要であり、その対訳度は 1 である。

提案手法の対訳度の性質より、複数対複数文対の対訳度は、常にそれをいくつかに分割した各文対の対訳度の積を下回らない。従って、次節に述べる最尤アラインメントの探索を可能な全ての文対候補を用いて行うと、常に文書対全体を 1 つの対とするアラインメントが最尤となる。これは妥当な結論だが、文アラインメントを行う趣旨からは無意味な結果であるので、目的に応じたヒューリスティックスを導入して文対候補をふるいにかけることにする。

我々が用いたヒューリスティックスは以下の通りである。

1. 各言語側の最大文数を制限する。
2. 文対候補 $(\boldsymbol{J} = \boldsymbol{J}_1 + \boldsymbol{J}_2, \boldsymbol{E} = \boldsymbol{E}_1 + \boldsymbol{E}_2)$ が以下の条件を満たさないならば、候補から外す: $m(\boldsymbol{J}, \boldsymbol{E}) / m(\boldsymbol{J}_1, \boldsymbol{E}_1) m(\boldsymbol{J}_2, \boldsymbol{E}_2) > t_m$ かつ $m(\boldsymbol{J}, \boldsymbol{E}) / m(\boldsymbol{J}_1, \boldsymbol{E}_2) m(\boldsymbol{J}_2, \boldsymbol{E}_1) > t_m$

4.2 最尤アラインメントの探索

前節で列挙した文対候補集合の部分集合のうち、文書対を過不足なく被覆できるものが、この文書対の文アラインメントであり、可能な文アラインメントのうち、各文対の対訳度の積が最大のものを最尤アラインメントとする。最尤アラインメントの探索は、以下のグラフ経路探索問題に帰着できる。

- 文書対の各文が文対によって被覆済かどうかの状態を考え、全ての状態の異なりに対するノードを作成する。
- 各ノードについて、現状態に追加配置可能な文対候補の各々に対して、それを追加した被覆状態を表すノードへのリンクを作成する。各リンクのコストは追加配置した文対の対訳度。
- 全文未被覆状態を表すノードから、全文被覆済状態を表すノードへの経路のうち、各リンクコストの総積が最大となるものを探索する。

我々は、A^* 探索を用いてこの問題を解いている。ヒューリスティック関数には、全ノード共通で、十分に大きな定数を与えている。

5 実験・考察

NHK の日英ニュース原稿コーパスを用いて、従来手法と提案手法の文アラインメント性能を比較する実験を行った。性能の評価は、アラインメントの結果得られた文対集合から統計機械翻訳モデルを学習し、翻訳結果を客観評価した結果を比較することで行う。

■文アラインメント実験条件 NHK 日英ニュース原稿 100,000 対より学習データとしてランダムに選んだ 80,000 対（1 記事対あたり平均で、日本語側 5.6 文/313 形態素[*2]、英語側 6.8 文/160 単語[*3]）を参照可能データとし、その中からさらに選ばれた 10,000 記事対（日本語側 55,317 文/英語側 68,171 文）に対して、以下の条件で従来手法と提案手法でそれぞれ文アラインメントを行った。

- 80,000 記事対全てを用い、言語内共起は文単位、言語間共起は記事対単位で言語内/言語間単語共起情報を収集し、統計的対訳辞書および言語内/言語間単語共起確率の一覧を作成した。対数尤度比閾値にはどちらも 100 を用いた。
- 可能な文対の大きさは、コーパスの性質を勘案して、日本語側 2 文、英語側 4 文以内とした。また日英とも連続しない複数文が 1 つの対応単位となることを許した。
- 従来手法では、式 (1) の SIM の値を対訳度指標に用い、上記条件の可能な文対全てを要素候補とし

[*2] 形態素への分割は MeCab+Unidic (`http://www.tokuteicorpus.jp/dist/`) を用いた。
[*3] 統計翻訳ツールキット Moses 付属の tokenizer と lowercaser を用いて前処理を行った。

従来手法				提案手法			
閾値	文対データ規模		BLEU	BLEU	文対データ規模		閾値
	文数	J:形態素数/ E:単語数			文数	J:形態素数/ E:単語数	
0	49,761	2,878,228 / 1,174,000	8.55				
0.009	31,290	1,573,229 / 702,553	7.95				
0.0105	26,526	1,287,799 / 588,349	5.69	10.17	22,976	1,307,281 / 580,835	0
0.0135	18,579	843,485 / 402,637	5.91	9.87	16,379	875,130 / 398,068	0.02
複数対複数文対を再度 1 対 1 文対にアラインメント							
0	51,888	3,085,858 / 1,242,053	8.36	10.03	40,775	2,441,588 / 1,017,390	0

表 1　評価実験結果

て、4.2 節の手順と同様に最尤アラインメントを求めた。提案手法では、4 章の内容に従い、$t_m = 1.2$ の条件で選択した文対を要素候補として最尤アラインメントを探索した。

実験の結果、従来手法では 40,050 個（1 対あたり平均、日本語 1.11 文/英語 1.45 文）、提案手法では 51,802 個（同、日本語 1.00 文/英語 1.03 文）の、両言語側とも空でない複数対複数文対を得た。

■翻訳実験条件　文アラインメント実験の結果得られた文対のうち 1 対 1 文対のみを用い、統計翻訳ツールキット Moses を用いて phrase-based 翻訳モデルを学習した。言語モデルには、学習データ 80,000 記事対の英語側文書から SRI 言語モデルツールキットを使って作成した 5-gram 言語モデルを用いた。各モデル学習後、日英ニュース翻訳用に構築した 1 対 1 文参照訳セットの開発用データ 84 文対を用いてパラメタを BLEU 値でエラーレート最小化チューニングし、同データセットの評価用データ 83 文対（1 文あたり日本語側平均 37 形態素）を用いて BLEU 値による翻訳性能比較を行った。なお、この参照訳セットは、上記日英ニュース原稿で学習データに選ばれなかった 20,000 記事対からランダムに記事対を選択し、1 対 1 文の直訳対になるよう人手で翻訳を調整して作成したものである。

■実験結果　実験結果を表 1 に示す。従来手法、提案手法ともに、アラインメントで得られた 1 対 1 文対を全て用いた場合（閾値 0）に加えて、文対の「対訳のよさ」を表す指標（従来手法では Utiyama らの提案する評価指標 SntScore、提案手法では単語あたり対訳度）がある閾値以上のもののみを用いた場合の結果も示してある。これは、提案手法において単語あたり対訳度が実際に対訳のよさを適切に表す指標であるかを検証するためと、翻訳モデルの学習データが同一規模であるときの BLEU 値を比較するため（表中、両手法でほぼ同規模のデータ規模から学習した結果を横に並べてある）である。

さらに、アラインメントで得られた 1 対 1 文でない対を、改めて同じアルゴリズムを用いて 1 対 1 文の対応に分割し、得られた文対を加えたデータから学習した結果も示す。これは、最初の文アラインメントで得られた複数対複数文対が妥当である（1 対 1 文対に分割すべきでない）かを検証するためである。

表 1 より、提案手法は全てのケースにおいて、従来手法より高い BLEU 値を得ている。文アラインメント実験の規模が小さく得られた学習データが少ないため、BLEU 値が全般に小さく、結果として手法間の差は顕著であるとまでは言えないが、提案手法がより適切な対応を推定する傾向にあると言ってよいと考える。

「対訳のよさ」閾値を変えた場合の性能の変化は、今回の実験結果では BLEU 値が小さすぎるため顕著な傾向が読み取れなかった。より大規模な実験を行い、再度検証したい。

複数対複数文対を 1 対 1 に分割して加えた場合は、従来手法、提案手法のいずれも加える前より性能が低下している。従って、両手法が推定した複数対複数文対は妥当である可能性が高いと言える。

6　おわりに

本稿では、トピックなどの単位で対応づけられたある程度の規模の対訳コーパスに対して、その中から獲得した言語内/言語間単語共起情報を利用することで、他の言語資源や知識資源を必要とせずに高精度の文アラインメントを行う手法を提案した。文アラインメント実験結果から学習した統計翻訳モデルは、従来より広く用いられてきた、統計的対訳辞書による単語対応数に基づく対訳度指標を用いた手法から学習したものに比べて、より高い性能を達成した。このことから、提案手法は有用な文アラインメント手法であると考える。

本稿では非常に小規模な実験結果を示すのみであったが、今後大規模な実験を行って再度検証を行いたい。また、アラインメント結果の人手による分析を行うことで、その傾向を把握し今後の改良につなげたい。

参考文献

[1] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74, 1993.

[2] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, Vol. 19, No. 1, pp. 75–102, 1993.

[3] Masao Utiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of ACL 2003*, pp. 72–79, 2003.

[4] Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, and Yuji Matsumoto. Bilingual text matching using bilingual dictionary and statistics. In *Proceedings of COLING '94*, pp. 1076–1082, 1994.