大域的な統語情報を用いた単語アラインメントの改善

坂口智洋 †　　中澤敏明 ‡　　黒橋禎夫 †
† 京都大学大学院 情報学研究科　‡ 科学技術振興機構
{sakaguchi, kuro}@nlp.ist.i.kyoto-u.ac.jp　nakazawa@pa.jst.jp

1　はじめに

コーパスベース機械翻訳では対訳コーパスにおける単語アラインメントの結果から翻訳モデルを構築するため、高精度の単語アラインメントが必要不可欠である。現在広く使われている単語アラインメントツールとして、IBMmodel[1] の単語アラインメントモデル部分を実装した GIZA++[2] が知られており、英仏など言語構造の近い言語対では高精度にアラインメントを行うことができる。

しかし、GIZA++では文を単なる単語列と見なして処理し依存関係などの言語情報は用いないため、日英や日中など言語構造が大きく異なる言語対では十分な精度を達成できていない。先行研究ではこれらの言語対に対して文の構造を利用する手法が多く提案されている。Riesa ら [3] は、片方の言語の構文木を用いた bottom-up の単語アラインメントモデルを提案した。また、Nakazawa ら [4] は両言語の依存木を利用することで単語同士の依存関係を組み込んだ単語アラインメントモデルを提案している。

こうした研究により内容語に対しては高精度にアラインメントを行えるようになったものの、機能語では対応付けを誤ることが多い。その原因の一つとして、これらのモデルは一単語から多くても数単語を基本単位としている事が挙げられる。そこで本研究では句や節などの大域的な統語情報に注目することで、単語アラインメントの精度向上を目指す。

2　ブロックアラインメント

日本語と英語など言語構造の大きく異なる言語対では、対応する単語の順序も大きく異なることが多いが、こうした単語対も句や節などの大きなかたまり同士の対応は満たしていることが多い。そこで本研究ではかたまりのことをブロック、対応するブロック同士をブロック対、対訳文中の全ブロック対をブロックアラインメントと呼ぶこととし、このブロックアラインメントを手がかりとすることで誤った単語アラインメントを修正することを目指す。例えば図 1 では、正解の単語アラインメントと整合性のある 3 つのブロック対を生成することができる。この例では、「は」が誤って「was」と対応付けられているが、ブロックアラインメントの情報を用いることで正しく「in」への対応に修正されることが期待できる。

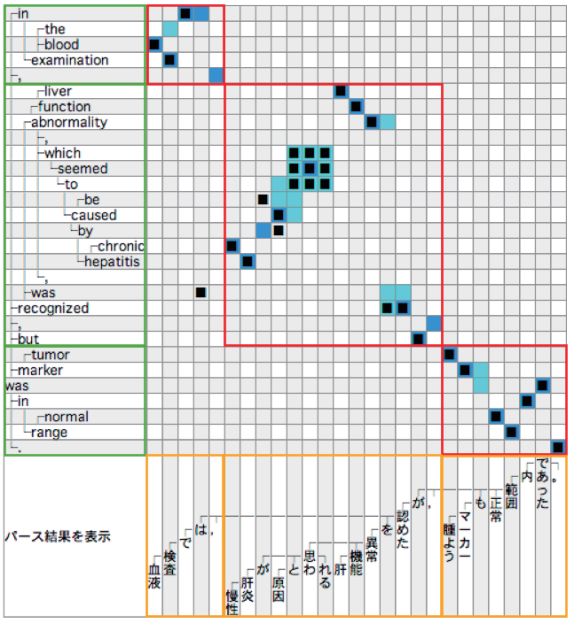

図 1: 統語情報から正解と整合性のある 3 つの大きなブロックを生成できる。このブロックを満たすように単語アラインメントを生成する。[青：必要な正解、水色：あっても間違えではない正解、黒：Nakazawa らの実験結果]

ブロックアラインメントは、ブロックの分割の仕方とそれらの対応付け方により複数の候補が存在するため、ブロックアラインメントにスコアを付け、スコアが最大となるブロックアラインメントを選択する。ブロック対を $\langle E, F\rangle$、ブロックアラインメントを $\{\langle E, F\rangle\}$ と表記するとき、スコア $S_b(\{\langle E, F\rangle\})$ が最大となるブロックアラインメント $\arg\max_{\{\langle E,F\rangle\}} S_b(\{\langle E, F\rangle\})$ を出力する。

本研究では、まず単語アラインメント結果を用いてブロックアラインメントを生成し、次にブロックアラインメント結果を用いて単語アラインメントを修正する、というサイクルを繰り返すことでより良い単語アラインメントを発見する手法を提案する。

2.1 ブロックアラインメント候補の生成

ブロック候補は句や節同士の依存関係のつながりの強さを元に作成したルールを利用して生成する。例えば英語の場合、according to や because of などから始まる前置詞句は他の部分との関係が弱いので一つのブロックを構成していると判断できる。

日本語文では形態素解析器 JUMAN と構文解析器 KNP を利用し、KNP の出力で連用形、level:C、level:B、level:A、外の関係、のいずれかの属性を満たす単語と、助詞「は」の直前を分割候補とした。level とは、南 [5] による従属節の分類の詳細化であり、A は「同時」の表現、B は「原因」、「中止」の表現、C は「独立」の表現を表している。外の関係とは、連体修飾のうち、修飾節と主名詞との間に格関係が認められないもののことである。また英文では nlparser を用いて句構造解析を行った結果に対して、フレーズのヘッドを定義するルールを適用し単語依存構造に変換したものを利用し、PP、SBAR、S、VP のいずれかの句の主辞となる単語の直前を分割候補とした。

上記のルールに従って対訳文中で分割候補を見つけ、各候補に対して分割するかしないかの全組み合わせをブロック候補として列挙した。なおブロック対は両言語でブロックが一対一対応するように生成するという制約を設け、ブロック同士が同数になるようにした。

2.2 ブロックアラインメントのスコア

ブロックアラインメントのスコアは、単語アラインメントとの一貫性と、ブロックの大きさの均一性を反映したものである。但し、ここで単語アラインメントは一単語同士とは限らず、数単語同士の対応を含む。対訳文 $\mathsf{S_e}$ と $\mathsf{S_f}$ において、対応する単語対を $\langle e, f\rangle$、対応するブロック対を $\langle E, F\rangle$、単語アラインメントを $\{\langle e, f\rangle\}$、ブロックアラインメントを $\{\langle E, F\rangle\}$ としたとき、ブロックアラインメントのスコア $S_b(\{\langle E, F\rangle\})$ を次のように定義する。

$$S_b(\{\langle E, F\rangle\}) = S_p \times S_w \times S_u$$

$$S_p = \frac{\sum_{\{\langle E,F\rangle\}} \sum_{\{\langle e,f\rangle\}} w(\langle e \subset E, f \subset F\rangle, (\mathsf{S_e}, \mathsf{S_f}))}{\sum_{\{\langle e,f\rangle\}} w(\langle e, f\rangle, (\mathsf{S_e}, \mathsf{S_f}))}$$

$$S_w = \frac{\sum_{\{\langle E,F\rangle\}} \frac{\sum_{\{w_e \in E\}} s(w_e)}{\mathrm{length}(\{w_e \in E\})} \times \frac{\sum_{\{w_f \in F\}} s(w_f)}{\mathrm{length}(\{w_f \in F\})}}{\mathrm{length}\{\langle E, F\rangle\}}$$

$$S_u = \sqrt{t(\{E\}, \mathsf{S_e}) \times t(\{F\}, \mathsf{S_f})}$$

S_p は現在の単語アラインメントがブロックアラインメントとどの程度一貫性があるかを表すスコアである。データ中の $\langle e, f\rangle$ の出現回数を $n_{\langle e,f\rangle}$ と表すとき、w を次のように定める。

$$w(\langle e, f\rangle, (\mathsf{S_e}, \mathsf{S_f})) = \frac{n_{\langle e,f\rangle}}{\sum_{e' \in \mathsf{S_e}} n_{\langle e',f\rangle}} \times \frac{n_{\langle e,f\rangle}}{\sum_{f' \in \mathsf{S_f}} n_{\langle e,f'\rangle}}$$

S_p では対応付けられた単語対しか考慮されない。しかし対応付けられていない単語もあり、こうした単語もどのブロックに属しそうかは判断できることが多い。S_w は各単語がどのブロックと対応付けられる確率が高いかを考慮したものである。ここでは内容語のみを扱い、これを w_e、w_f と記述する。データ中で w_e と w_f が対応する回数を $n_{\langle w_e,w_f\rangle}$ とするとき、$s(w_e)$ を次のように定める。

$$s(w_e) = \frac{\sum_{w_f \in F} n_{\langle w_e,w_f\rangle}}{\sum_{w_f \in \mathsf{S_f}} n_{\langle w_e,w_f\rangle}}$$

S_u は各ブロックが均一な大きさであるかを評価するスコアである。ここでは $t(\{E\}, \mathsf{S_e})$ を次のように定める。

$$t(\{E\}, \mathsf{S_e}) = 1 - \sum_{\{E\}} \frac{\mathrm{length(E)}}{\mathrm{length}(\mathsf{S_e})} \times \log \frac{\mathrm{length}(E)}{\mathrm{length}(\mathsf{S_e})}$$

なお、F の場合も同様に定義する。

3 モデル

提案モデルは Nakazawa ら [4] のモデルの拡張である。この先行モデルは依存木を用いることで、大きく語順の異なる単語対の対応付けを高精度で行うことに成功した。しかし、機能語などの明確な対訳が存在しないものは、誤って全く関係のない単語と結び付けられてしまうことが多々あるため、ブロックアラインメントを組み込むことでこれを改善する。

提案手法の流れを図 2 に示す。ブロックアラインメントステップでは、ルールに基づいて文をブロックに分割しブロック同士の対応付けを行い、ブロックアラインメントを生成する。単語アラインメントステップでは Nakazawa らのモデルに、ブロックアラインメントから導出されるブロック確率を導入し単語アラインメントの修正を行う。

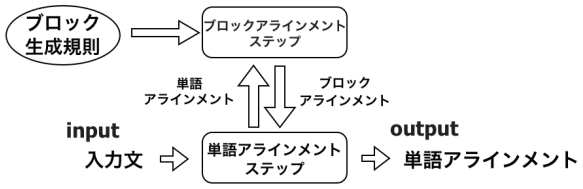

図 2: 提案手法の流れ

3.1 生成モデル

Nakazawa らのモデルは l 個の概念を生成し、それぞれの概念から両言語で木を作り、各言語で木を結合させることで対訳文を生成する生成モデルである。概念の数 $l(>0)$ は定数 $p_\$$ を用いた幾何分布 $P(l) = p_\$ \cdot (1-p_\$)^{l-1}$ によって調整される。各概念は分布 θ_T に基づいて両言語で部分木を生成する。この部分木のペアをコア・アラインメントと呼び、$\langle e_C, f_C \rangle$ と表記する。部分木の片方は NULL でもよいが、簡単のためこれは 1 つの単語からなるものとする。また、各部分木 e_C と f_C はそれぞれ、分布 ϕ に基づいた機能語からなる派生木 $\{d_{e_C}\}$ と $\{d_{f_C}\}$ を持つ。ここでは $\langle e_C, f_C \rangle, \{d_{e_C}\}, \{d_{f_C}\}$ を含めたものを $\langle e, f \rangle$ と表記する。

本研究ではこれに加え、複数の部分木の集合からブロックアラインメント $\{\langle E, F \rangle\}$ を生成することを考える。簡単のためブロックアラインメントは e_C と f_C のみに依存すると近似し、確率分布 ψ で表わされるものとする。

最後に、それぞれの言語で部分木を結合し、結合した木を D と表す。このとき、文が生成される同時確率は次のように定式化される。

$$P(l, \{\langle e, f \rangle\}, D, \{\langle E, F \rangle\})$$
$$= P(l) \cdot P(D|\{\langle e, f \rangle\}) \cdot$$
$$\prod_{\{\langle e,f \rangle\}} [\theta_T(\langle e_C, f_C \rangle) \cdot \phi(\langle e, f \rangle) \cdot \psi(\langle e_C, f_C \rangle)]$$

3.2 ブロック確率

ψ はブロックアラインメントと単語アラインメントの一貫性を表す確率分布であり、各単語が NULL と対応付くか否かで場合分けを行う。NULL と対応付かない単語対の場合、対応する単語対が同じブロック対に含まれているか否かを離散的に表す。e_C と f_C が同じブロック対にある場合は 0.8、異なるブロック対にある場合は 0.05、ブロックが部分木をまたぎ、e_C や

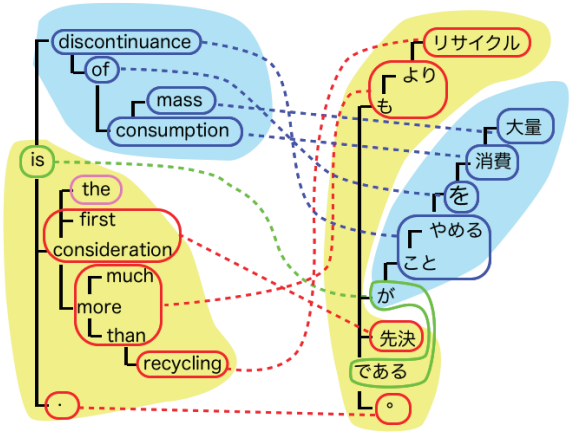

図 3: ブロックアラインメントの例

f_C がどのブロック対に対応するか判断を付けられない場合は 0.15 とするとき、ψ は次のように表せる。

$$\psi(\langle e_C, f_C \rangle) = \begin{cases} 0.8 & (e_C \subset E, f_C \subset F) \\ 0.05 & (e_C \subset E, f_C \subset F' \neq F) \\ 0.15 & (\text{otherwise}) \end{cases}$$

また、片方の単語が NULL に対応する場合はこれとは分けて考え、次のように表す。

$$\psi(\langle e_C, f_C \rangle) = 1.0$$

図 3 に例を示す。黄色と水色で塗られたかたまりがブロックであり、黄色同士、水色同士が対応するブロック対である。赤色と青色の点線で結ばれた単語対は同じブロック対にあるので $\psi = 0.8$、緑色の点線は日本語文でブロック対が単語をまたいでいるため $\psi = 0.15$ となる。また、桃色で囲まれた単語「the」は NULL と対応付けられており $\psi = 1.0$ となる。

3.3 モデルの学習

モデルの学習は Nakazawa ら [4] と同様の手法で行う。これはギブスサンプリングを用いた学習手法である。まず GIZA++を用いてアラインメントを初期化し、その後木構造を組み替える 6 つの操作を用いてサンプリングを行う。

4 単語アラインメント実験

4.1 手法

提案手法の有効性を検証するために、対訳コーパスを用いてアラインメント実験を行った。利用した対訳コーパスは JST 日英論文抄録コーパス 30 万文である。

	Pre.	Rec.	AER
ベースライン	91.00	83.15	12.78
提案手法	**91.22**	**83.37**	**12.56**

表 1: JST 論文コーパス 30 万文での実験結果

Nakazawa らのシステムをベースラインとして、提案手法と比較した。

アラインメントの評価には、人手で正解を付与した 500 文に対して、以下の式で計算される Precision、Recall、Alignment Error Rate(AER) を用いた。

$$\text{Precision} = \frac{|A \cap P|}{|A|} \qquad \text{Recall} = \frac{|A \cap S|}{|S|}$$

$$\text{AER} = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

但し、A はシステムの出力、S は必要な正解、P は日本語の接頭辞や英語の冠詞のような、あっても誤りではない正解である。

4.2 結果と考察

表 1 に実験結果を示す。実験の結果、Precision、Recall、AER の全てで改善が見られた。図 4 に改善された例を示す。緑色の四角がブロックアラインメントである。ベースラインのモデルでは「による」が「in」と誤って対応していたが、ブロックアラインメントを用いることで正しく「by」と対応するよう修正することができた。一方図 5 は似たような状況にも関わらず改善されなかった例である。この例では「用いた」が「studies」と誤って対応付いたままとなってしまっている。これを改善するためには、ブロックアラインメントの影響をより強くしてブロックに引き込む必要があるが、現状ではブロックアラインメントの精度が十分でないため悪影響を及ぼしてしまう。ブロックアラインメントの精度の更なる向上が必要である。

5 まとめと今後の課題

統語情報により生成されたブロックを考慮した単語アラインメントモデルを提案し、実験の結果精度が改善されることが確認できた。今後の課題として次のようなものがある。現在はブロック同士は一対一でしか対応付けていないが、多対多対応を許すようにモデルを拡張する。また、現在は分割ルールを予め決めて利用しているが、これを自動で生成できるようにする。

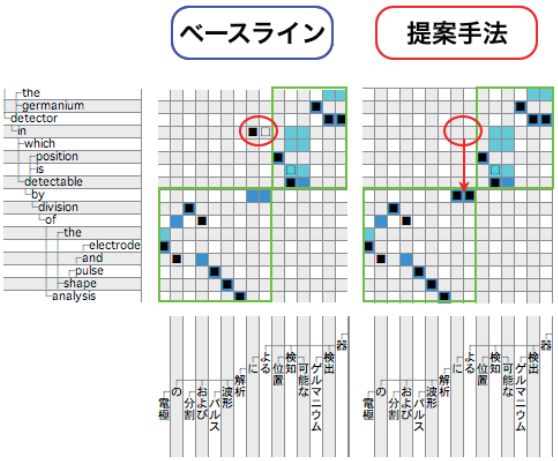

図 4: 改善例

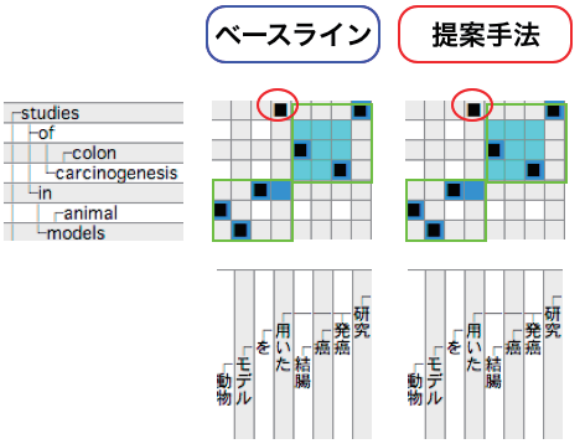

図 5: 改善できなかった例

参考文献

[1] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. Conputational Linguistics, 19(2):263-312.

[2] F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics. 29(1):19-52. MIT Press. Cambridge, MA. USA.

[3] J. Riesa and D. Marcu. 2010. Hierarchical Search for Word Alignment. In Proceedings of the 48th Annual Meeting of the ACL, Uppsala, Sweden

[4] T. Nakazawa and S. Kurohashi. 2012. Alignment by bilingual generation and monolingual derivation. In Proceedings of COLING 2012, pp. 1963-1978

[5] 南不二男. 1993. 「現代日本語文法の輪郭」, 大修館書店