単語間結合度に基づく複単語表現のアライメントの改善

塩田 嶺明[†]　　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
[†]{shioda, kuro}@nlp.ist.i.kyoto-u.ac.jp
[‡]nakazawa@pa.jst.jp

1　はじめに

コーパスベースの機械翻訳では単語アライメントの結果から翻訳知識を獲得するため、高い翻訳精度を実現するには単語アライメントの精度を向上させる必要がある。しかし、日英などの言語構造が大きく異なる言語間では、英仏などのよく似た言語間に比べるとアライメントの精度は低い。言語構造の違いは、依存構造解析の情報を利用することである程度克服することができる。それでもなおアライメントが困難なのは、複数の単語で構成される表現、特に機能表現である。

複数の単語で一つの意味を表す表現は言語ごとに様々なものがあり、機能表現の他にも固有名詞や慣用句などがある。このうち、機能表現などは構成語単体では全体の意味と結びつかないことが多く、また対訳文において相手言語側に明確に対応する語を持たないことが多いため、アライメントがより困難になっている。特に膠着語である日本語では、長い機能表現が動詞に後続することが多いため、この問題が顕著である。フレーズベース SMT[1] では、意味のある単語の集合と単なる単語列を明示的に区別するものではないため、この問題に対処できているとは言いがたい。

図 1 は、依存構造木を利用したアライメントモデルによる複単語表現に関する誤りの例である。なお、図では黒い四角がシステムの出力を表し、濃い青と薄い青のマスは正解を表す。ここでは “に 及ぼす” が塊として扱われておらず、“及ぼす” が “effects” と対応してしまっている。

本研究では、既存の依存構造解析を利用したアライメントモデルに単語列上の情報も加え、さらに隣接する単語間のつながりの強さも加味することで、複単語表現のアライメントを改善することを目指す。

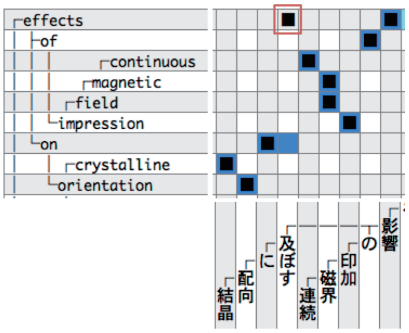

図 1: 複単語表現に関するアライメントの誤り例

2　関連研究

複単語表現を機械翻訳で扱う手法としては、まとまりやすい表現とそのスコアを獲得して利用するアプローチが多く用いられている。

Ren ら [4] は、対訳コーパスの片方の言語側から対数尤度比を用いて複単語表現を獲得し、単語アライメントの結果からその対訳を獲得する。得られた複単語表現とその対訳を使って SMT を補強している。Liu ら [2] の手法はまず、単言語コーパスの各文を複製して対訳コーパスのように扱い、単語アライメントを行ってから同じ単語同士の対応を取り除く。残った対応は単言語で共起しやすい単語同士であるから、そこから任意の二単語の共起しやすさのスコアを獲得し、アライメントや翻訳において素性として利用している。

3　提案モデル

3.1　ベースラインシステム

本研究で利用したベースラインシステムは中澤ら [3] のアライメントモデルである。このモデルではフレーズの翻訳確率に加え、フレーズの依存関係確率や単言語の派生確率を考慮することで言語間の構造の違いを

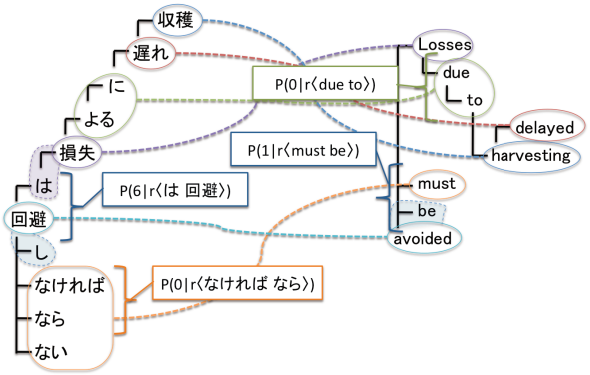

図 2: モデル概要

吸収している。既存の統計的単語アライメントモデルで推定したアライメントを出発点とし、前述の確率を計算しながらサンプリングによってアライメントを修正していく。

この手法では依存構造解析において意味主辞に基づく依存構造を用いている。意味主辞は統語主辞に比べ、内容語同士の依存関係が言語間で保持されやすい。一方、図 2 の “し なければ なら ない” のように、日本語で動詞に後続する機能語列が全て動詞に係る子になるため、単語列上の情報なしでは依存関係が同じになる動詞の前の助詞と区別しにくい。

3.2 アライメント距離確率モデル

本研究では、先行研究のように複単語表現そのものを事前に獲得して利用するのではなく、隣接する二単語間の結合度という形に一般化して複単語表現に対応する。すなわち、複単語表現を構成する単語は結合度が強く、そのような二単語間では相手言語側でも同じもしくは近接するフレーズに対応しやすいと仮定する。この仮定に基づき、**単語間結合度**に対して相手言語側での対応先の距離を確率化し、サンプリング時に利用する。なお、ここで隣接する二単語とは単語列で連続であって、かつ依存構造木上でも連続または兄弟の関係になっていることを意味する。以後、対応先の距離を**アライメント距離**、上に述べた確率を**アライメント距離確率**と呼ぶことにする。単語間結合度を $r\langle bigram\rangle$ と表し、アライメント距離確率を次式のように定義する。

$$P(d|r\langle bigram\rangle) \tag{1}$$

アライメント距離は、相手言語側で対応するフレーズ間の単語列上の最短距離とする。なお、二単語の対応先が同じである場合は、距離は 0 とする。また、二単語に NULL のアライメントが含まれる場合は、

- 一単語目が NULL 対応
- 二単語目が NULL 対応
- 両方の単語が NULL 対応

の三つに分類する。

図 2 にモデルの概要を示す。日英の対訳文が依存構造木で表してあり、色付きの破線でつながれたフレーズがアライメントを表す。また、アライメント距離確率の例をいくつか示している。これらのうち、“なければ なら” という二単語を考える。これらの二単語は同じ “must” に対応するのでアライメント距離は 0 となる。この二単語は結合度が高く、図の確率は高くなることが期待される。

3.3 単語間結合度

単語間結合度を表すスコアとして様々なものを用いた。一つ目は **Bigram** 確率であり、Bigram 確率が高いほど結合度は高いと考えた。

$$P(w_n|w_{n-1}) \tag{2}$$

二つ目は逆方向の Bigram 確率も加味したスコアである。順方向、逆方向双方の Bigram 確率が高ければ、より結合度は高いと考えられる。これらの幾何平均を結合度 (**Bidirect**) とする。

$$(P(w_n|w_{n-1}) \cdot P(w_{n-1}|w_n))^{\frac{1}{2}} \tag{3}$$

三つ目は、単語の出現確率も考慮したスコアである。機能表現などはコーパス中に高頻度で出現すると考えられるため、二単語それぞれの出現確率も加味したものを結合度 (**Uni&Bi**) とする。

$$(P(w_{n-1}) \cdot P(w_n))^{\alpha} \cdot (P(w_n|w_{n-1}) \cdot P(w_{n-1}|w_n))^{\frac{1}{2}} \tag{4}$$

なお、実験ではスコアが小さくなりすぎないように、$\alpha = \frac{1}{10}$ とした。

上で述べたスコアは、アライメント実験でも使用する JST 日英論文抄録コーパス [5]（約 100 万文）を単言語コーパスとして計算したもの（JST）と、大規模 Web コーパス（日本語約 70 億文、英語約 10 億文）から計算したもの（Web）をそれぞれ用いた。スコアの例を表 1 に示す。この中で、“なければ なら” などは全てのスコアリング手法において高いスコアを示して

表 1: 単語間結合度スコアの例

	Bigram(JST)	Bigram(Web)	Bidirect(JST)	Bidirect(Web)	Uni&Bi(JST)	Uni&Bi(Web)
なければ なら	0.845346	0.364585	0.602368	0.175835	0.097682	0.026065
なら ない	0.591857	0.206020	0.197442	0.084910	0.042676	0.014478
に よる	0.068922	0.015649	0.261949	0.123414	0.103429	0.022370
は 回避	0.000044	0.000049	0.000936	0.001620	0.000240	0.000309
due to	0.989265	0.734175	0.191227	0.060962	0.059808	0.014370
be avoided	0.003486	0.000993	0.045504	0.020610	0.008718	0.003042

いる。一方、“に よる” は Bigram 確率ではそれほど高くないが、逆方向 Bigram 確率なども考慮すると比較的高いスコアとなる。また、つながりの弱い “は 回避” は全てのスコアリング手法で低いスコアである。なお、Web コーパスから計算すると表現のバリエーションが増えるため、論文コーパスに比べると低いスコアとなっている。

アライメント距離確率の計算は、単語間結合度に対して行うとスパースになるため、スコアをいくつかの範囲に分割してその範囲内で計算している。その際、範囲に含まれる Bigram の頻度がアライメントを行うコーパス内でほぼ均等になるようにしている。また、実際の確率分布に近づけるために、隣接する範囲の確率値を用いてスムージングを行っている。

4 アライメント実験

4.1 実験設定

実験に使用したコーパスは、3.3 節で述べた JST 日英論文抄録コーパスである。このうち、文 ID 順に最初の 30 万文を用いて実験した。まず、日英それぞれの文に対して構文解析を行う。日本語に関しては、形態素解析器 JUMAN と構文解析器 KNP を利用した。英語に関しては、nlparser を用いて句構造解析を行った結果に対し、フレーズのヘッドを定義するルールを適用することで単語依存構造に変換する。

アライメントは、3.1 節で述べた中澤らのモデルをベースラインとし、これに提案モデルを組み込んだものと比較した。単語間結合度として 3.3 節で述べたスコアを使用する方法では、アライメント距離確率を計算する範囲の分割数は 10 または 20 とした。また、式 1 のアライメント距離確率を単語間結合度によらず、$P(d)$ として計算する方法でも実験を行った。

アライ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。

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

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

A はシステムの出力（図 1 の黒の四角）、S は必要な正解（図 1 の濃い青の部分）、P は日本語の接尾辞や英語の冠詞のようにあっても誤りではない正解（図 1 の薄い青の部分）である。

表 2: アライメント実験の結果（30 万文）

	Pre.	Rec.	AER
ベースライン	91.40	82.87	12.74
単語間結合度なし	91.47	84.07	12.10
Bigram(JST,10 分割)	91.84	83.80	12.06
Bigram(JST,20 分割)	91.85	83.89	12.00
Bigram(Web,10 分割)	91.62	83.77	12.18
Bigram(Web,20 分割)	91.62	83.83	12.16
Bidirect(JST,10 分割)	91.41	84.03	12.15
Bidirect(JST,20 分割)	91.91	84.11	11.86
Bidirect(Web,10 分割)	91.42	83.90	12.21
Bidirect(Web,20 分割)	91.25	84.02	12.23
Uni&Bi(JST,10 分割)	91.82	**84.18**	11.87
Uni&Bi(JST,20 分割)	**91.99**	84.13	**11.81**
Uni&Bi(Web,10 分割)	91.30	84.17	12.13
Uni&Bi(Web,20 分割)	91.60	83.92	12.11

4.2 結果と考察

論文コーパス 30 万文で行ったアライメント実験の結果を表 2 に示す。まず、単語間結合度を利用せずアライメント距離確率のみを計算した場合でも、ベースラインに比べ大きな改善がみられる。これは、ベースラインでは単語列に関する情報は利用されていなかっ

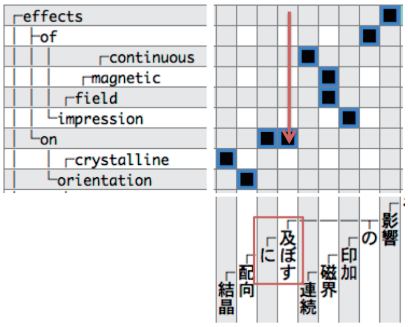

図 3: アライメントが改善された例

たためである。論文コーパスから計算した単語間結合度を利用した場合は、多くの場合でさらなる改善がみられ、単語間結合度を利用しない場合に比べ AER が最大 0.29 改善した。一方、単語間結合度を Web コーパスから計算した場合、論文コーパスに比べるとアライメントの精度は改善しなかった。これは論文コーパスに特有の専門的な表現に対応できなかったためと考えられる。

図 3 に提案手法（Uni&Bi(JST,20 分割)）でアライメントが改善された例を示す。“及ぼす” の対応先が “effects” から “on” に移動し、“に 及ぼす ↔ on” という正しい対応が得られている。

逆に提案手法でもアライメントを改善できなかった例としては、構文解析の誤りにより複単語表現を構成する単語が依存構造木で不連続になっているものがあった。このケースについては、構文解析時にも単語間結合度を導入し、結合度の強い単語同士は依存構造木で連続になるようにできれば改善が期待できる。

5 翻訳実験

提案手法で論文コーパス 99.4 万文をアライメントしたデータを用いて翻訳実験も行った。翻訳に用いたシステムは中澤らの EBMT システム [3] である。テストデータには日英論文抄録コーパスのうち、アライメントには使用していない 2000 文を用いた。翻訳の評価には自動評価手法の BLEU を用いた。

実験の結果を表 3 に示す。AER の改善とは対照的に、翻訳精度の改善はみられなかった。一般的に、アライメントが改善されたからといって直ちに翻訳精度が改善するとは限らないと言われている。翻訳精度を向上させるためには、翻訳においても複単語表現を適切に扱う枠組みを検討する必要がある。

表 3: 翻訳実験の結果（BLEU スコア）

	AER	英 → 日	日 → 英
ベースライン	12.62	**22.52**	**17.60**
Uni&Bi(JST,20 分割)	11.80	22.20	17.51

6 おわりに

本研究では、複単語表現のアライメント精度を改善するため、単語間結合度が強い単語対ほど近接するフレーズに対応されやすくなるようなモデルを提案した。実験の結果、ベースラインに比べアライメントの精度は向上し、特に複単語機能表現に関して改善が見られた。

今後は、同モデルを日中など他の言語対に適用して効果を調べる。また、翻訳において複単語表現を適切に扱う枠組みの検討を進める予定である。

参考文献

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *HLT-NAACL2003: Main Proceedings*, pp. 127–133, 2003.

[2] Zhanyi Liu, Haifeng Wang, Hua Wa, and Sheng Li. Improving statistical machine translation with monolingual collocation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 825–833, 2010.

[3] Toshiaki Nakazawa and Sadao Kurohashi. Alignment by bilingual generation and monolingual derivation. In *Proceedings of COLING 2012*, pp. 1963–1978, 2012.

[4] Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pp. 47–54, 2009.

[5] Masao Uchiyama and Hitoshi Isahara. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 72–79, 2003.