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

Improvement in Alignment of Multi-word Expressions Based on Strength of Word Connection

京都大学大学院情報学研究科 中澤　敏明

PROFILE 2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

nakazawa@nlp.ist.i.kyoto-u.ac.jp TEL 075-753-5346

川崎重工業株式会社 塩田　嶺明

PROFILE 2014 年京都大学大学院情報学研究科知能情報学専攻修士課程修了。修士（情報学）。現在は川崎重工業株式会社に勤務。

京都大学大学院情報学研究科教授 黒橋　禎夫

PROFILE 1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006 年 4 月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事

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品質を確保する必要がある。英仏などの似通った言語間では既に高いアライメント精度が実現されているが、日英など言語構造の大きく異なる言語間では、これらと比較すると十分な精度を達成できているとは言えない。

アライメントが困難となる原因の一つに、複数の単語で一つの意味を表す複単語表現がある。複単語表現は言語ごとに様々な種類があり、機能表現や慣用句、固有名詞などがその例である。複単語表現の一部には、構成語単体では全体の意味と結びつかない性質をもつものがある。この性質は、特に“かもしれない”や“in order to”などの機能表現に顕著にみられる。また、このような性質を持つ表現は、単語単体では相手言語側に明確に対応する語をもたないことが多い。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、アライメントの難しさが顕著になっている。

近年主流となっているフレーズベースの統計的機械翻訳システム[1]では、単語アライメントの結果に対してヒューリスティクスを適用することで複数の単語からなるフレーズを獲得している。しかし、ここでは隣接する単語を機械的に連結してフレーズを作っているだけである。つまり意味的なまとまりのある単語の集合と、単なる単語 n-gram は明示的に区別されていないため、複単語表現の問題に十分に対処できているとは言い難い。

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を加え、複単語表現のアライメントの改善を目指す。具体的には、隣接する二単語間の結合度を定義し、結合の強い単語同士は相手言語において近接するフレーズに対応されやすいようなアライメントモデルを構築する。実験の結果、ベースラインに比べアライメント精度が向上することが確認できた。

2 関連研究

複単語表現の機械翻訳での扱いに関する先行研究では、複単語表現を事前に獲得して利用する方法が多く見

られる。Lambert ら [2] は、双方向のアライメントの非対称性に着目して複単語表現の候補とそのスコアを獲得し、スコアが閾値以上のフレーズをアライメントや翻訳において一単語として扱っている。Ren ら [3] 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利用して SMT システムを補強している。Liu ら [4] は、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う。そこから同じ単語同士の対応を取り除くと共起しやすい単語同士の対応が得られる。これを用いて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 アライメントモデル

本研究では中澤ら [5] のアライメントモデルをベースラインとし、隣接する二単語間の結合度に基づき、つながりの強い単語同士が相手言語において近接するフレーズに対応されるようなアライメントモデルを構築した。

3.1 ベースラインシステム

中澤らのアライメントモデルでは、言語構造が異なる言語間でのアライメントを改善するため、依存構造木上でフレーズの依存関係を考慮することで語順の違いを吸収し、単言語での単語の派生モデルを導入することで相手言語側に明確に対応する語を持たない孤立機能語に対応している。アライメントは以下のステップで行われる。まずは既存の単語アライメントツールによって単語レベルの対応を推定し、ヒューリスティックなルールを用いて依存構造木上で連続なフレーズの対応を獲得する。これを出発点とし、サンプリングを複数回行いながらアライメントを修正していく。

このモデルでも、フレーズの言語的なまとまりとしての適切さは考慮されていない。【図 1】は、複単語表現に関するアライメント誤りの例である。ここでは、“に及ぼす”がかたまりとして扱われておらず、“及ぼす”が誤って“effects”と対応してしまっている。

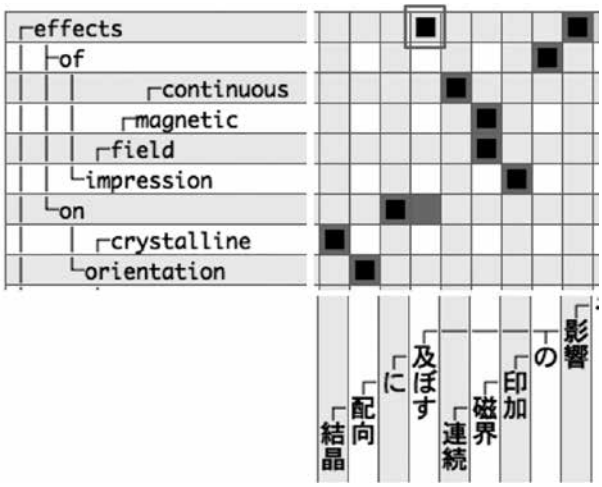

図 1　アライメント誤り例

また、中澤らは依存構造解析において意味主辞に基づく依存構造木を利用している。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という特長がある。一方、例えば【図 4】の“なければ なら ない”のように、動詞の子になる単語が増えるため、これらの単語を適切に扱う必要がある。

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を事前に獲得して利用するのではなく、隣接する単語間の結合度を用いて複単語表現に対応する。こうすることで、扱う表現を限定することなく様々な表現に柔軟に対応できる上に、他の言語対にも容易に拡張することができる。

まず、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と対応しやすいと仮定する。この仮定に基づき、隣接する二単語間の結合度を定義し、結合度に対して相手言語側での二単語の対応先の距離を次のように確率化する。

$$P(\delta|\mathrm{r(bigram)})$$

なお、r(bigram) はある二単語の単語間結合度であり、δ は対応先の距離である。以下、対応先の距離をアライメント距離、上記確率をアライメント距離確率と呼ぶことにする。アライメント距離は、相手言語側で対応するフレーズの単語列上の最短距離とする。ただし、二単語が同じ対応に含まれる場合はアライメント距離を 0 とする。また、注目する二単語に NULL 対応が含まれ

る場合は、1 単語目が NULL 対応、2 単語目が NULL 対応、両方 NULL 対応の３パターンに分けて扱う。なお、ここで隣接する二単語とは、単語列で連続かつ依存構造木上で連続もしくは兄弟の関係になっている二単語を指すこととする。

アライメント距離確率は、前述の仮定から、単語間結合度が強い場合はアライメント距離が小さいほど確率が高く、アライメント距離が大きいと確率は低いという【図２】のような分布になると考えられる。一方、単語間結合度が弱い場合はこの傾向は弱まり、【図３】のような分布になると考えられる。

【図４】に提案モデルの概要を示す。この図では日英の対訳文が依存構造木で表されており、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。例えば、“なければ なら”の二単語に着目する。この二単語はいずれも“must”と対応している。この二単語は“なければ なら ない”という複単語表現の一部であり、結合は強いため、同じフレーズと対応する確率は高いと考えられる。一方、“は 回避”の二単語のアライメント距離は 6 であるが、この二単語は結合が弱いため、確率はそれほど低くならないと考えられる。

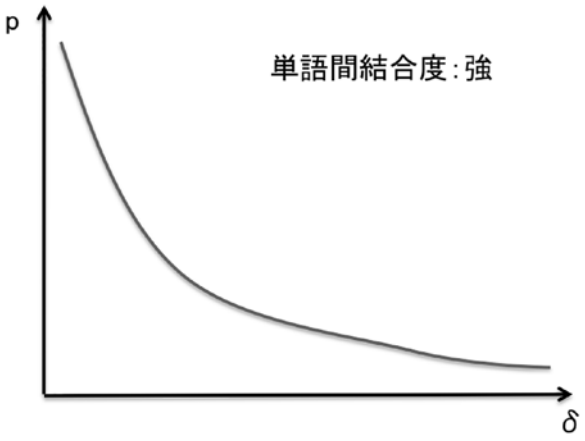

図２　単語間結合度が強い場合の確率分布

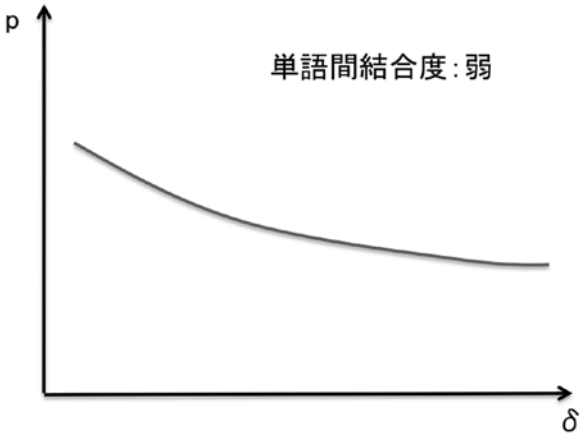

図３　単語間結合度が弱い場合の確率分布

4　単語間結合度

隣接する単語間の結合度のスコアとしては、以下に説明する３種類のものを用いた。一つ目は、Bigram 確率（Bigram）である。

$$P\left(w_n|w_{n-1}\right)$$

２つ目は順方向 Bigram 確率と逆方向 Bigram 確率の幾何平均をとったもの（Bidirect）である。こうすることで、より共起しやすい二単語のスコアを高くすることができる。

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

３つ目は、両方向の Bigram 確率に加え、両単語の出現確率も組み合わせたもの（Uni&Bi）である。これにより、コーパス中に高頻度で出現する機能表現に対応する。なお、実験では Bigram 確率に重みをおくため $\alpha=1/10$ とした。

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

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 [6]（約 100 万文）から計

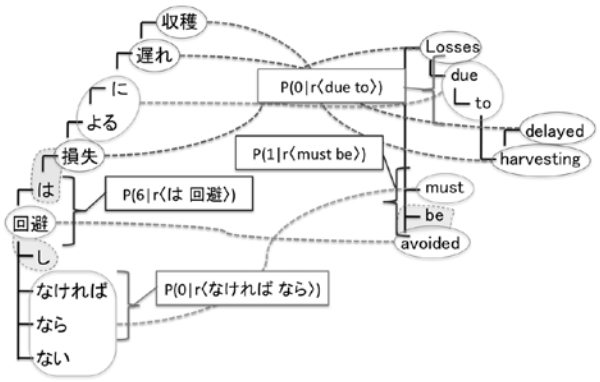

図４　提案モデル概要

算した。【表1】に、単語間結合度のスコアの例を示す。このうち、複単語表現を構成する“なければ なら”や“due to”などはいずれの方法でも高いスコアとなっている。一方、“に よる”は Bigram 確率のスコアは低いが、逆方向 Bigram 確率や単語の出現確率も考慮すると高いスコアとなる。また、つながりの弱い“は 回避”や“be avoided”はいずれの方法でも低いスコアとなっている。

表1　単語間結合度のスコア例

	Bigram	Bidirect	Uni&Bi
なければ なら	0.845376	0.602368	0.097682
なら ない	0.591857	0.197442	0.042676
に よる	0.068922	0.261949	0.103429
は 回避	0.000044	0.000936	0.000240
due to	0.989265	0.191227	0.059808
be avoided	0.003486	0.045504	0.008718

アライメント距離確率を計算する際、単語間結合度の値そのものに対して行うとスパースになってしまう。これに対処するため、単語間結合度をいくつかの範囲に分割して、その範囲内でアライメント距離確率の計算を行っている。この際、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均等になるように分割を行っている。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。本研究では実際のデータに基づいてアライメント距離確率の計算を行ったが、確率分布を仮定してパラメータ推定を行う方法も考えられ、これについては今後の検討課題とする。

5　アライメント実験

5.1　実験設定

実験に使用したコーパスは、前節で述べた 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 万文を用いて実験を行った。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、フレーズの主辞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した。アライメントは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実験を行った。提案モデルとしては、単語間結合度を利用する方法と、単語間結合度は利用せず、全ての二単語間で同じアライメント距離確率 P(δ) を用いる方法を採用した。単語間結合度を利用する場合は、範囲の分割数を 20 とした。

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、次の式で計算される Precision、Recall、Alignment Error Rate (AER) を用いた。

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

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

ただし、A はシステムの出力（【図5】の■または□）、S は必要な正解（【図5】の濃い青のマス）、P は日本語の接尾辞や英語の冠詞などのように、あっても誤りではない正解（【図5】の薄い青のマス）である。AER はアライメントの総合的な精度の良さを表す指標であり、数値が小さいほど精度が良いことを意味する。

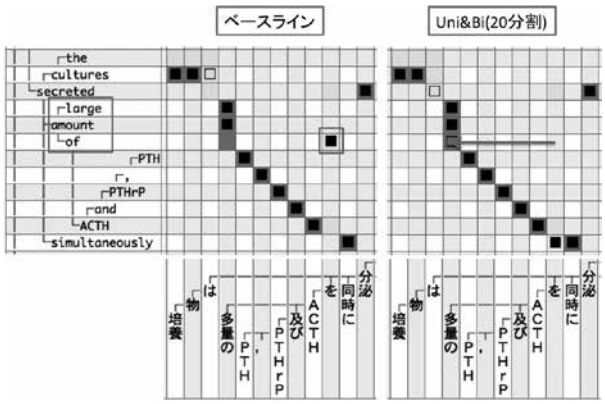

図5　アライメントが改善された例

5.2　結果と考察

実験結果を【表2】に示す。まず、単語間結合度を利用せずにアライメント距離確率のみを計算する方法でも、ベースラインに比べ AER が 0.6 ほど改善している。これは、ベースラインシステムにおいて単語列上の情報がほとんど利用されていなかったためである。単語間結合度を利用した場合はさらに改善がみられ、単語間結合

表２　アライメント実験の結果

	Prec.	Rec.	AER
ベースライン	91.40	82.67	12.74
単語間結合度なし	91.47	84.07	12.10
Bigram	91.85	83.89	12.00
Bidirect	91.91	84.11	11.86
Uni&Bi	**91.99**	**84.13**	**11.81**

度を利用しない場合に比べ最大で AER が 0.3 ほど改善している。全体的に Precision が上昇する傾向がみられ、結合の強い単語同士がまとまって対応づけられたことによりアライメント誤りが抑制できていると考えられる。

【図５】に、提案手法によってアライメントが改善された例を示す。ベースラインでは、“large amount of ⇔ 多量の” の対応のうち、“of” のみ対応先が誤っている。提案手法ではこれらの対応先の距離が遠いため “of ⇔ を” の対応が外れ、“of” が正しく “多量の” に対応付けられている。

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上で不連続になっているケースがあった。この問題に関しては、構文解析時にも単語間の結合を考慮し、結合の強い単語同士が依存構造木上で連続になるようにすれば改善が期待できる。また、アライメント結果を利用して構文解析を改善する方法も考えられる。他のアライメント誤りの例としては、孤立機能語が他の大きな対応に吸収されてしまうような副作用もみられた。この問題については、単語の品詞など文法情報を利用することで改善できると考えられる。

6　おわりに

本研究では、複単語表現のアライメントを改善するため、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精度は改善し、特に複単語表現を構成する単語が誤って遠くの単語と対応するアライメント誤りを抑制する効果を確認した。

今後は、提案モデルを日中など他の言語対に適用してアライメントを行い、効果を調べる予定である。また、翻訳精度を改善する方法の検討も今後の課題である。具体的には、翻訳において複単語表現を過不足なくカバーする適切なフレーズが選択されるような枠組みを検討する必要がある。

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] Patrik Lambert and Rafael Banchs. Grouping Multi-word Expressions According to Part-Of-Speech in Statistical Machine Translation. In Proceedings of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context. Trento, Italy. pp. 9–16, 2006.

[3] Zhixiang Ren, Yajuan L u, 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.

[4] 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.

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

[6] 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.