意味主辞に基づく依存構造木を利用した対訳文の句アライメント

塩田 嶺明　　中澤 敏明　　黒橋 禎夫
京都大学工学部　京都大学大学院情報学研究科
{shioda, nakazawa}@nlp.ist.i.kyoto-u.ac.jp kuro@i.kyoto-u.ac.jp

1　はじめに

英語と日本語のように、語順や言語構造が大きく異なる言語間で対訳文アライメントを行う際には、依存構造解析 [5] の情報を取り入れることで精度を向上することができる。依存構造木におけるフレーズ内の主辞の定義のしかたには、統語主辞（Syntactic Head）と意味主辞（Semantic Head）の２つがある。統語主辞は、フレーズ内で文法的に重要な要素を主辞として定義する一方、意味主辞はそれを意味的に重要な要素と定義する。図 1 に統語主辞に基づく依存構造木、図 2 に意味主辞に基づく依存構造木の例を示す。同じ文でも、英語文は統語主辞が “is”、意味主辞が “studied” で、日本語文は統語主辞が “いる”、意味主辞が “研究” であり、それぞれ異なっている。統語主辞と意味主辞は同じ語である場合もある。

既存の手法では、統語主辞に基づく依存構造木が用いられてきた [5] 。しかしこの場合、膠着語である日本語は内容語に後続する機能語が主辞となる一方、英語では内容語自体が主辞となるケースや、逆に日本語は内容語が主辞となる一方、英語は対応する内容語に前置される助動詞が主辞となるケースなど、語の依存関係の不一致がしばしば起こる。図 3 に、語の依存関係の不一致のためにアライメントが誤った例を示す。なお、この図では■がシステムの出力を表し、濃いグレーと薄いグレーが正解を表す。英語の主辞 “are” に対し、日本語の主辞は “示した” である。“示した” に対応する “presented” は子になっており、結果としてその対応を獲得できていない。同様に、“described”、“述べ” の対応も獲得できていない。これらは主辞となる語を言語間で一致させることにより改善が期待できる。

そこで本論文では、各言語の単語の依存関係をより近づけるために、従来の統語主辞に変わって、意味主辞に基づく依存構造木を利用したアライメントを行うことを提案する。アライメント実験の結果、統語主辞を用いた場合に比べてアライメント精度を向上することができた。

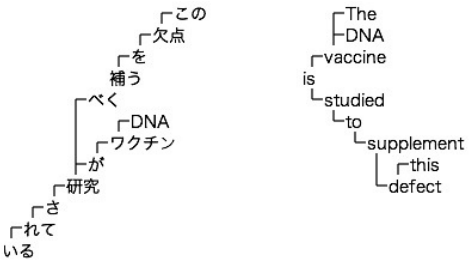

図 1: 統語主辞に基づく依存構造木

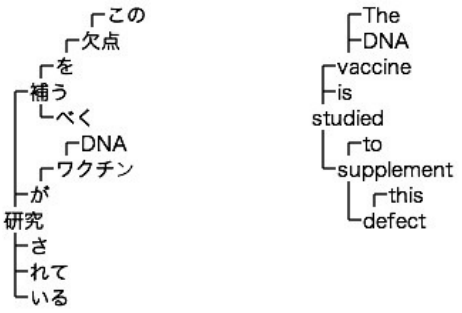

図 2: 意味主辞に基づく依存構造木

2　関連研究

依存構造木の主辞情報を用いた翻訳の研究として、磯崎ら [1] の Head Finalization、Hong ら [2] の研究などがある。磯崎らは、英語（SVO 型）、日本語（SOV 型）の語順変化に着目し、英語の依存構造木で統語主辞がフレーズ内の子に先行する場合、フレーズ内で語順を交換することで日本語に似た語順が得られるという手法を提案した。Hong らは、英語・韓国語の機械翻訳で同様の手法をとるが、こちらは英語文の構文解析の際意味主辞に基づく依存構造木を生成して用いている。

また、文の構造情報を取り入れることでアライメントの精度向上を目指した研究として、Quirk ら [3] 、

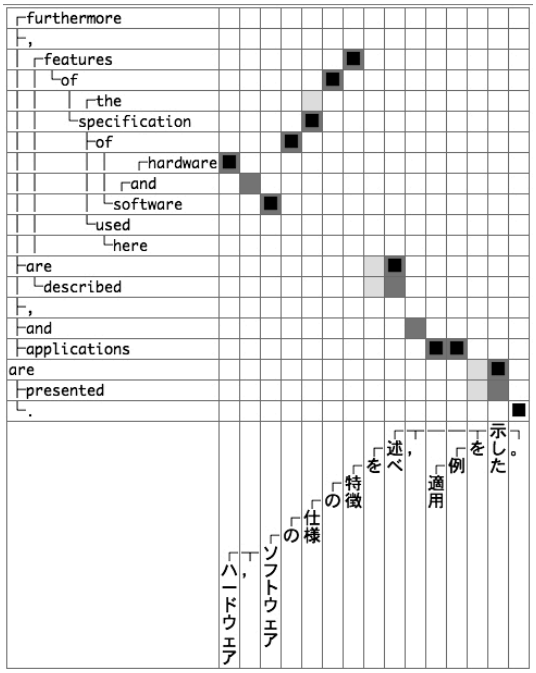

図 3: アライメントの誤り例

Cherry ら [4] 、中澤ら [5] などがある。Quirk らは単語列アライメントを行った後に文の構造情報を統合しているが、そもそも単語列アライメントの精度が高くないため有効とは言いがたい。Cherry らは依存構造解析の結果をアライメントに用いているが、単語単位の手法で、1 対 1 対応に限るという制約がある。中澤らは依存構造木上で連続な単語に対応を拡大し、句の対応も可能とした。これらの研究では、統語主辞に基づく依存構造木が用いられている。

3 意味主辞に基づくアライメント

3.1 ベースライン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とした。このモデルでは、依存構造木上でサンプリングによる統計的句アライメントを行っており、依存構造木を用いることで言語間の構造の違いを吸収し、また多対多対応を獲得することができる。

3.2 意味主辞に基づく依存構造木の生成

意味主辞に基づく依存構造木を対訳文から生成する手順について述べる。まず日本語文については、JUMAN を用いて形態素解析し、その結果を KNP を用いて意味主辞に基づく依存構造に変換する。意味主辞に基づく依存構造木では、文末の基本句内の内容語を主辞とし、後続する機能語はすべてこれに係る子とみなす。英語文については、Charniak の nlparser を用いて句構造に変換し、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する。変換ルールを意味主辞に基づいて定義し、目的の依存構造木を生成する。表 1 に変換規則の例を示す。

表 1: 英語文の単語依存構造変換ルールの例

親	子	統語主辞	意味主辞
VP	AUX,RB,VP	AUX	VP
VP	MD,ADVP,VP	MD	VP
VP	VBD,VP	VBD	VP
VP	AUX,ADJP	AUX	ADJP
VP	TO,VP	TO	VP
SQ	MD,NP,VP	MD	VP

意味主辞に基づく依存構造木では、1 つの内容語に複数の機能語が係るなど、統語主辞に基づく依存構造木にはない形が現れることがある。従来のアライメントモデルでは、部分木になるような句の対応しか許していなかったが、この構造の変化に対応するため、依存構造木上で不連続であっても、単語列として連続で、かつ係り先が同じ語であれば（兄弟語）対応を拡大できることにした。図 2 の “さ”・“れて”・“いる” などは兄弟語の例である。

4 実験

4.1 実験設定

提案手法の有効性を検証するため、アライメント実験を行った。使用したコーパスは、内山・井佐原らの方法により作成した JST 日英抄録 [6] (約 100 万対訳文) である。アライメントの評価には、人手で正解を与えた 479 対訳文を使用した。実験の条件は以下の 3 パターンとした。

- 統語主辞
- 意味主辞
- 意味主辞+兄弟語の対応あり

評価には以下の式で表される Precision、Recall、Alignment Error Rate(AER) を用いた。AER はアライメントの総合的な精度を表す指標で、数値が小さい

表 2: アライメント実験の結果

主辞	Precision	Recall	AER
統語	**89.67**	62.98	25.69
意味	88.82	65.77	23.94
意味 (+兄弟語)	88.46	**67.07**	**23.31**

ほど精度が良い。なお、A はシステムの出力（図 3 の■）、S は必要な正解（濃いグレー）、P はあっても誤りではない正解（薄いグレー）である。

$$Precision = \frac{|A \cap P|}{|A|}$$

$$Recall = \frac{|A \cap S|}{|S|}$$

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

4.2 結果と考察

表 2 にアライメント実験の結果を示す。提案手法は統語主辞ベースと比較すると Precision は低下しているが、Recall が大きく向上しており、AER は兄弟語なしの場合 1.7 ポイント、兄弟語ありの場合で 2.4 ポイント改善している。Recall が向上したのは、両言語間で単語の依存関係が近くなったことにより、統語主辞ではとれていなかった単語の対応がとれたことが大きな要因である。Recall の向上は、用例ベース機械翻訳において利用できる翻訳知識を増やす意味において重要である。逆に提案手法において Precision が統語主辞に比べて低いのは、依存構造木上で位置の近い誤った語を対応づけてしまったケースが影響していると考えられる。

図 4 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。図 3 は同じ文を統語主辞に基づく依存構造木を用いてアライメントした結果である。比較してみると、従来は NULL 対応だった “described”、“presented” がそれぞれ “述べ”、“示した” と正しく対応している。依存構造木における位置が一致したことが効果的だったと言える。

図 5 に、兄弟語の対応について示す。統語主辞の場合は “may be”、“ことがある” というまとまった対応がとれているが、意味主辞ではこれがとれていない。“役立つ” に続く機能語群が木になっていないためである。兄弟語の対応を許したところ、まとまった対応がとれている。意味主辞に基づく依存構造木を用いる際

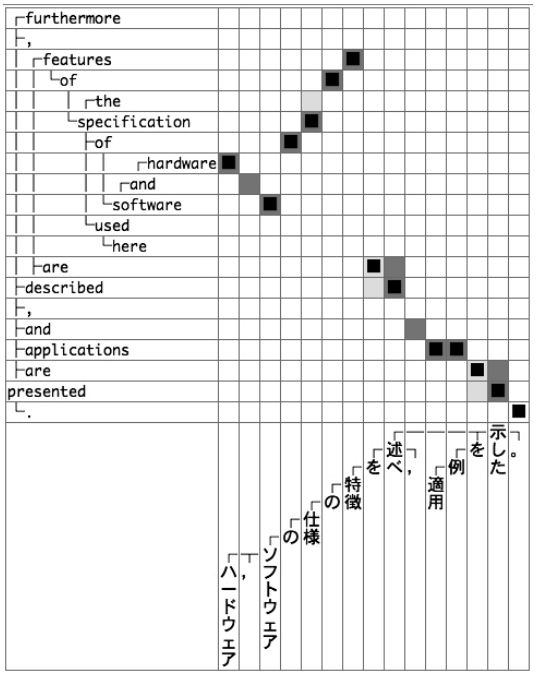

図 4: アライメントの改善例

には、こうした点に留意する必要がある。表 2 において意味+兄弟語の Recall が高くなっていることもこれを裏付けている。なお、兄弟語ありの場合 Precision が兄弟語なしの場合に比べて下がっているが、この原因は統語主辞と同じ初期アライメントのミス、もしくは誤って冠詞やカンマを併合したことによるものが多い。後者は簡単なルールを作って除外することができる。

図 6 に、提案手法でアライメントが誤った例を示す。これは、パースエラーと木構造の変化に起因すると考えられる。まず、副詞の “より” が誤って動詞と判定され、後続する “よく” が子になっている。“よく” と “接近” の木構造上での距離が遠くなったため、“よく” と対応すべき “better” が誤って “うる” と対応してしまっている。

この他、同じような意味の単語が複数ある場合に選択を誤るケースも目立つ。これは統語主辞の場合にもみられるため、主辞の定義によらない問題と思われ、別に対策を講じる必要がある。

5 結論

本研究では依存構造解析の情報をアライメントに組み込む際、従来用いられていた統語主辞ではなく、意味主辞に基づく依存構造木を利用することを提案した。この手法では単語の依存関係を言語間でより近づけることができ、単語の対応をとりやすくなる。アライメ

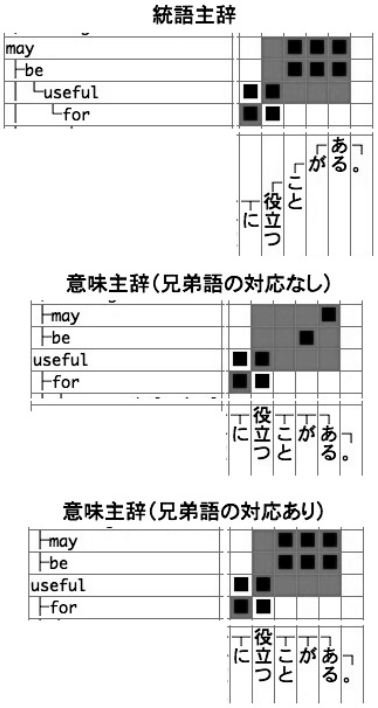

図 5: 木構造の変化に伴う対応の拡張

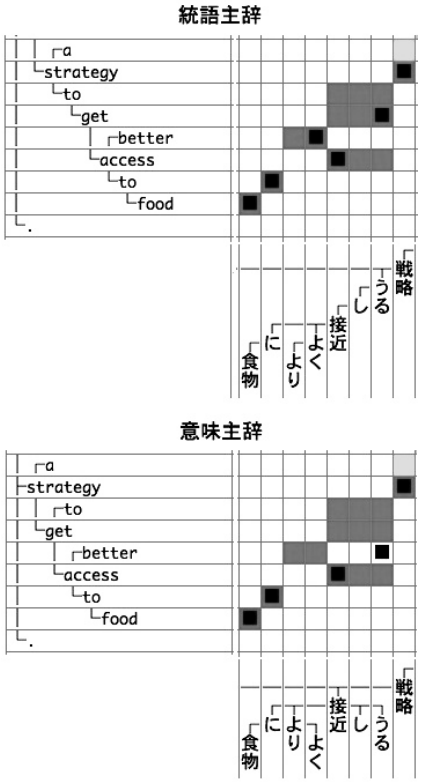

図 6: アライメントの誤り例

ント実験の結果、提案手法で精度が向上できることが確認できた。特に、統語主辞に基づく依存構造木では語の依存関係の不一致により誤っていた対応が正しくとれるようになっており、提案手法の有効性を示すことができた。

今後は、この結果を用いて翻訳実験を行い、翻訳においても意味主辞が有効であることを示す予定である。

参考文献

[1] Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh, Head Finalization: A Simple Reordering Rule for SOV Languages, *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR,* pp. 244–251, 2010.

[2] Gumwon Hong, Sweng-Wook Lee, and Hae-Chang Rim, Bridging morpho-syntactic gap between source and target sentences for English-Korean statistical machine translation, *Proceedings of the ACL-IJCNLP Conference Short Papers,* pp.233–236, 2009.

[3] Chris Quirk, Arul Menezes, and Colin Cherry, Dependency treelet translation: Syntactically informed phrasal SMT, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* pp. 271–279, 2005.

[4] Colin Cherry and Dekang Lin, A probability model to improve word alignment, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* pp. 88–95, 2003.

[5] Toshiaki Nakazawa and Sadao Kurohashi, Bayesian Subtree Alignment Model based on Dependency Trees, *Proceedings of the 5th International Joint Conference on Natural Language Processing,* pp. 794–802, 2011.

[6] Masao Utiyama and Hitoshi Isahara, A Japanese-English patent pararell corpus, *MT summit XI,* pp. 475–482, 2007.