整合性尺度を用いた構造的対訳文アラインメント

中澤 敏明　　Yu Kun　　黒橋 禎夫
東京大学大学院情報理工学系研究科　　京都大学大学院情報学研究科
{nakazawa, kunyu}@kc.t.u-tokyo.ac.jp　　kuro@i.kyoto-u.ac.jp

1　はじめに

現在の機械翻訳の主流は、Brown ら [1] や Och ら [7] に代表されるように統計翻訳であるが、我々は用例ベース翻訳システムを研究している。これには以下の 2 つの理由がある。

1 つは、構造的な言語処理の発展を指向しているからである。機械翻訳は、言語処理研究の成果である形態素解析や構文解析などの基礎技術のアプリケーションとしてとらえることができる。つまり、基礎技術の発展がアプリケーションの精度向上につながるであろうし、逆にアプリケーション側から、基礎技術の弱点が見えてくることもある。

2 つ目は、用例ベース翻訳の問題設定が妥当である場合が少なくないからである。たとえば、マニュアルのバージョンアップの翻訳や、関連特許の翻訳などは、コーパスの規模としてはそれほど多くなく、統計翻訳がうまく働くかというと疑問が残る。しかしこのような場合、ドメインが同じならば、同じような用例が多く得られるはずであり、用例ベース翻訳の考え方が有利に働くものと考えられる。

用例ベース翻訳の研究としては、Menezes ら [6] の “Logical Form” を用いた研究や、Groves ら [3] の構文木を利用した研究などがあるが、いずれもヒューリスティックなルールに基づいた手法であり、1. 複数あるアラインメント候補から最適なものを選択する統一的な基準がなく、2. アラインメント全体としての整合性を測る尺度もない。

本稿では、二つのアラインメント間に**係り受け距離**と**距離-スコア関数**を定義し、これらを構造的アラインメント手法に組み込むことを提案する。これにより上記の 2 つの問題を解消することができる。実験結果では、ベースライン手法に比べて 3.0 ポイント以上のアラインメント精度の向上が見られた。

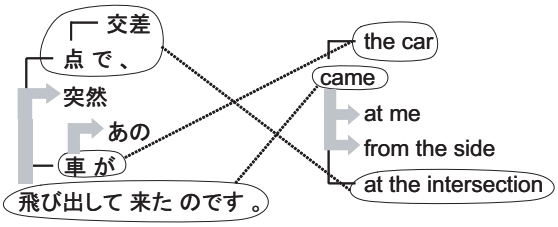

図 1: アラインメントの例

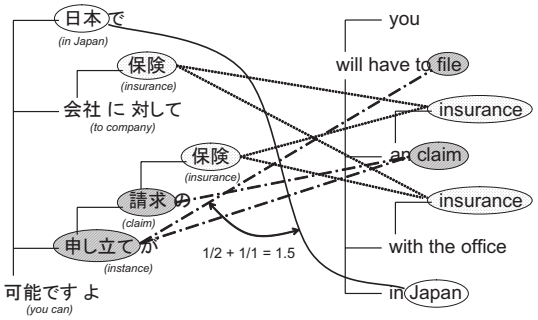

図 2: 曖昧性の例

2　構造的対訳文アラインメント

統計翻訳の基本姿勢は、対訳辞書や言語知識などは利用しないというものであるが、用例ベース翻訳ではこのような言語資源を積極的に利用する。

まず用例を得るための対訳文を、それぞれ図 1 のようなフレーズごとの依存構造を持った形に変形する。日本語文に対しては JUMAN[5] および KNP[4] を用い、英語文に対しては Charniak のパーサ [2] を用いて句構造に変換し、これにフレーズの head を定義するルールを適用することにより依存構造に変換する。

次に対訳辞書を用いて、単語ごとの対応候補を見つける。図 1 では「交差点 - intersection」「車 - car」「来た - came」の 3 つの対応候補が得られている。単語ごとの対応がわかったら、その単語を含むフレーズをひとかたまりとして、フレーズごとの対応をつける。対

応がつかなかったフレーズは、ルールによって他の対応のついているフレーズにマージする。図 1 の例では、「あの」が「車 the car」に、「突然」「at me」「from the side」が「飛び出して来たのです came」にマージされる。

以上がアラインメントの流れであるが、ここで問題となるのが、曖昧性のある対応候補や、文脈上不適切な対応候補が得られる場合である。図 2 の例では、日本語の “保険” と英語の “insurance” がそれぞれ 2 回ずつ出現しているため、組み合わせとして 4 つの対応候補が存在したり、“請求” と “申し立て” にそれぞれ “claim” という訳語があり、対応候補が衝突しているなどの曖昧性がある。このような対応候補の中から、いかに適切な対応を選ぶかが重要であり、本稿でもこの部分に注目する。

3 整合性尺度を用いた構造的対訳文アラインメント

対応候補からの適切な対応の選択を実現し、かつ全体的な整合性を測るために、任意の二つの対応候補 (a_i と a_j とする) の間に**整合性スコア**を定義する。整合性スコアを用いて、以下の式から最も整合的なアラインメントを得る。

$$\underset{alignment}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{整合性スコア}\ (a_i, a_j) \qquad (1)$$

整合性スコアは文の依存構造木上で定義される。

まず、任意の一組の対応候補 a_i(原言語の句 p_{Si} と目的言語の句 p_{Ti} との対応) と a_j (同様に p_{Sj} と p_{Tj}) に注目する。a_i と a_j の原言語側の**距離** $d_S(a_i, a_j)$ を、p_{Si} と p_{Sj} との木構造上の距離と定義する。目的言語側も同様に $d_T(a_i, a_j)$ が定義される。

この距離を用いて、整合性スコアは以下のように定義する：

$$\text{整合性スコア}\ (a_i, a_j) = f(d_S(a_i, a_j), d_T(a_i, a_j))$$

$f(d_S(a_i, a_j), d_T(a_i, a_j))$ は距離のペアに対してスコアを付与する**距離-スコア関数**である。

文全体のアラインメントの整合性は、整合性スコアを用いて式 1 のように定義される。また整合性スコアを用いて、各対応候補の整合性も以下のように計算できる：

$$score(a_i) = \sum_{j \neq i}^{n} \text{整合性スコア}\ (a_i, a_j) \qquad (2)$$

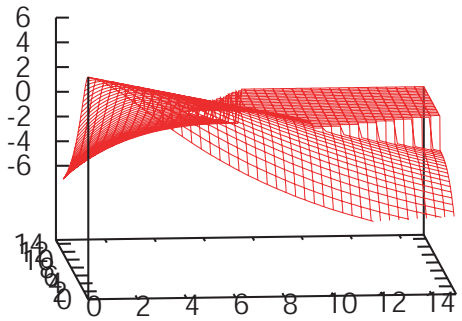

図 4: 距離-スコア関数

日本語の係り受け距離

用言:レベル C	6
用言:レベル B+/B	5
用言:レベル B-/A ト格	4
ヲ格/二格/デ格	3
ガ格/ノ格/連体	2
文節内 用言:レベル B+	1

英語の係り受け距離

S/SBAR/SA/:	5
VP/ADVP ADJP/WHADVP WHADJP	4
NP/PP/INTJ QP/PRT/PRN	3
others	2

図 5: 係り受け距離

以上の定義により、整合的なアラインメントを得る問題は、**距離**と**距離-スコア関数**をいかに実装するかという問題に置き換わった。

3.1 ベースライン手法

ベースライン手法では、曖昧でない対応候補は無条件で採用し、曖昧なものに対してのみ、距離と距離-スコア関数を用いる。

またすべての枝の距離を 1 とする。つまり、ある対応から別の対応への依存構造木上での移動距離をそのまま対応間の距離とする。

また距離-スコア関数は、$f(d_S, d_T) = 1/d_S + 1/d_T$ とする。これは、「曖昧な候補の近くにある (d_S、d_T が小さい) 曖昧でない候補は、曖昧な候補を強く支持する」という仮定に基づいている。図 2 にスコア計算の例を示す。

全ての曖昧な候補のスコアを式 2 に基づいて計算し、最も高いスコアを得たものを採用し、これと衝突する候補を棄却する。曖昧な候補がなくなるまでこの計算を繰り返す。

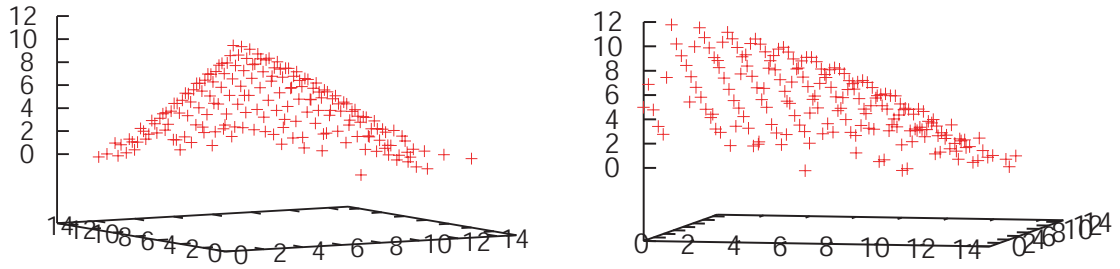

図 3: 正解データから学習された距離ペアの分布

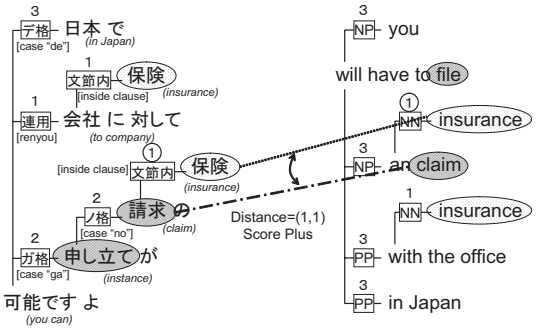

図 6: 適切な距離関係の例

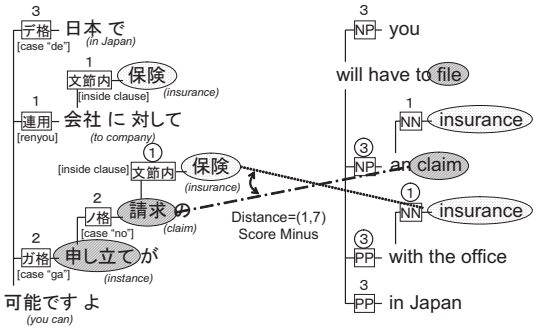

図 7: 不適切な距離関係の例

3.2 提案手法

提案手法では**距離**と**距離-スコア関数**を改善し、アラインメントの精度向上を目指す。

3.2.1 距離-スコア関数の学習

距離-スコア関数を改善する。まず毎日新聞 4 万対訳文のアラインメント正解データ [8] から、距離ペアの頻度分布を計数した。

図 3 にこの結果を示す。縦軸 (Z 軸) は頻度の log を取ったものである。二つの横軸 (X,Y 軸) は日本語、英語それぞれの距離である。3 次元のプロットを 2 次元で視覚的に捉えやすくするために、別角度からの図を 2 つ示した。分布は、距離が近いペアが最も多く、遠いペアになるに従ってゆるやかに減少し、距離の差が大きくなるに従って急激に減少している。

この結果を元に、次のような基準を考え、距離-スコア関数を人手で設定した：

- d_S・d_T とも小さければ、適切な関係である可能性が高いので、プラスの値を与える
- どちらも大きければ、二つの対応の関係の信頼性が薄いので、スコアは 0 とする
- 一方が小さく、他方が大きければ、不適切な関係である可能性が高いので、マイナスの値を与える

実際に設定した関数は図 4 に示した。

3.2.2 係り受け距離

ベースライン手法では、すべての枝の距離は 1 であるとしたが、実際には文には句や節などの単位が存在し、これらの情報を利用することが有効である。例えば、木構造上では隣接していても、異なる節に属する句同士の距離は大きいと考えられるからである。

日本語構文解析器 KNP が出力する係り受けタイプ情報と、Charniak の nlparser が出力する英語のタグ情報に対して、人手で木構造上での距離を定義した。これを**係り受け距離**と呼び、d_S および d_T の計算に利用する。節などの区切りの強さが強いものほど、距離が大きくなるようにする。図 5 にその一部を示す。また、図 6、図 7 に実際の文での適用例を示す。枝上のラベルが係り受けタイプ、ラベルの上の数字が、係り受け距離である。

図 6 の例では、注目する対応同士の距離が、日本語・英語とも 1 であり近いため、適切な関係であると判断し、プラスのスコアを付与する。一方、図 7 では、日本語の距離は 1 で近いが、英語の距離は 7 と遠い。このような対応同士は、どちらかが不適切な対応である可能性が高いため、マイナスのスコアを付与する。

またこの距離設定のもとで、前章と同様に距離ペアの頻度分布を係数し、距離- スコア関数を再設定した。

表 1: 実験結果

	適合率	再現率	F 値
ベースライン	60.26	61.68	58.79
+距離-スコア関数学習	64.35	61.58	60.81
+係り受け距離	64.93	62.64	61.91

3.3 アラインメントの整合性

最良のアラインメントは式 1 により、整合性スコアの和が最大になるように、それぞれの対応を採用・棄却していけばよい。しかし全ての場合を調べるのでは候補の数が爆発してしまうので、グリーディーに探索する。

すべての対応の候補について、式 2 で表されるスコアを計算する。スコアが最も高いものを採用し、それと衝突する候補は棄却する。同時に、このとき計算されたスコアが閾値を下回る候補があった場合、その候補は誤った対応である可能性が高いため、その場で棄却する。これを繰り返すことにより、近似的に最良のアラインメントが得られる。

4 実験結果と考察

毎日新聞対訳コーパス [8] からランダムに 500 文を選び、アラインメントを行なった。このコーパスにはアラインメントの正解データが付与されている。

評価の単位は、英語は単語単位、日本語は文字単位とした。これは、正解データも我々の出力も句単位なのだが、日本語の句の区切りが必ずしも一致しないためである。評価は、各文ごとに正解データとの適合率・再現率・F 値を算出し、その平均値を計算した。

また対訳辞書として、研究社の日英辞書 (36K 見出しから 214K 対訳を抽出) と英日辞書 (50K 見出しから 303K 対訳を抽出) を利用した。

実験結果を表 1 に示す。+距離-スコア関数学習はベースライン手法の距離-スコア関数を改善し、学習して得られたものに変えた結果で、+係り受け距離はさらに係り受け距離を利用した結果である。

距離-スコア関数を改善することにより再現率は若干下がったものの、適合率が大幅に向上し、F 値では 2.1 ポイントの精度向上が見られた。また係り受け距離を利用することにより、さらに精度が向上し、ベースラインよりも 3.1 ポイント以上の精度向上を達成した。

この結果から、我々の提案する**距離-スコア関数**と**係り受け距離**がアラインメントに効果的に働いていることがわかる。

5 結論と今後の課題

本稿では、距離-スコア関数 $f(d_S, d_T)$ と係り受け距離を利用した新しいアラインメント手法を提案した。アラインメント全体の整合性を全ての対応候補のペアのスコアの和で定義し、適切な候補の選択を可能にした。これにより、3 ポイント以上のアラインメント精度の向上を達成した。

今後の課題は、現在人手で設定している係り受け距離を自動学習で獲得することである。これには、各言語で独立に学習する方法や、言語ペアで学習する方法などいくつか考えられるが、その中から最適な方法で学習することを考えている。

参考文献

[1] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Association for Computational Linguistics*, 19(2):263–312, 1993.

[2] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan, June 2005.

[3] Declan Groves, Mary Hearne, and Andy Way. Robust sub-sentential alignment of phrase-structure trees. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1072–1078, 2004.

[4] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534, 1994.

[5] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of Japanese morphological analyzer JUMAN. In *Proceedings of The International Workshop on Sharable Natural Language*, pages 22–28, 1994.

[6] Arul Menezes and Stephen D. Richardson. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL) Workshop on Data-Driven Machine Translation*, pages 39–46, 2001.

[7] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Association for Computational Linguistics*, 29(1):19–51, 2003.

[8] Kiyotaka Uchimoto, Yujie Zhang, Kiyoshi Sudo, Masaki Murata, Satoshi Sekine, and Hitoshi Isahara. Multilingual Aligned Parallel Treebank Corpus Reflecting Contextual Information and Its Applications. In *Proceedings of the MLR2004: PostCOLING Workshop on Multilingual Linguistic Resources*, pages 63–70, 2004.