多様な情報源と整合性尺度に基づく高精度対訳文アラインメントに関する研究

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。

kuro@i.kyoto-u.ac.jp　075-753-5344

京都大学大学院情報学研究科
中澤　敏明

PROFILE 2006年東京大学大学院情報理工学系研究科電子情報学専攻修士課程修了。現在京都大学大学院情報学研究科博士後期課程在学。機械翻訳の研究に従事。

nakazawa@nlp.kuee.kyoto-u.ac.jp

1 はじめに

我々は構造的言語処理の高度化と機械翻訳の高度化を目指し、用例ベース翻訳の研究を行っている。用例翻訳では言語リソースを可能な限り利用し、できるだけ大きな翻訳例を用いることで文脈を安定させ、翻訳を適切にする。語や小さな語列ではなく、できるだけ大きな翻訳用例を利用しようとすれば、語列としては不連続であっても構造的につながっている用例を扱える方がよく、（必然ではないにしても）構文情報を用いることが自然である。

しかしながら、現在の機械翻訳研究の主流は統計翻訳である。統計翻訳は、パーサ、対訳辞書などの言語リソースがない場合に自然なアプローチであり、一方、言語リソースがある場合には、それらと対訳コーパスを最大限に利用してどこまで機械翻訳を高度化できるかという問題に挑戦することも自然である。我々は後者の問題設定を選び、用例翻訳研究を進めている。

また統計翻訳は英語とヨーロッパ言語などのように、言語構造の似た言語対には有効に働くが、日英などのように構造が大きく異なる言語ついには不利であると言われている。用例ベース翻訳では深い言語処理技術を用いるため、言語構造の違いを柔軟に吸収し、精度よい翻訳が可能である。

本稿では、我々が開発した用例翻訳システムのアライメントモジュールについて説明を行う。また、特許文の自動翻訳、翻訳支援の可能性の検討の第一歩として、本手法を特許日英対訳文に適用した結果を示す。

2 対訳文のアラインメント

翻訳は、対訳文の用例化（対訳文アラインメント）と、入力文とマッチする用例の検索とその組み合わせからなる。今回は特に対訳文アラインメントについて説明する。

対訳文のアラインメントの例を【図1】に示す。【図1】では、木構造のルートノード（文全体のヘッドノード）は最も左側に、各文節は上から下に順番に並んでいる。また下線のついている各語は、対訳辞書によって対応が得られたものである。枠で囲まれた各部分、およびそれらの組み合わせが用例としてデータベースに登録される。

対訳文アラインメントにおいて重要な点は2つある。1つは対訳文中の対応候補を可能な限り多く探索し、粒度の細かいアラインメントを行なうことである。翻訳での利用を考えると、サイズの大きな用例しか学習できて

いなければ、入力文と一致する用例が見つからない可能性が高くなってしまうからである。

多くの対応候補を探索していくと、曖昧性のある対応候補や、文脈上誤っている対応候補も増えてくる。そこで2つめの重要な点が、見つかった対応候補の中から、適切な対応のみを選択することである。これが正確に行なえないと、誤ったアラインメントをしてしまい、結果として正しくない用例が学習されてしまう。用例の精度は翻訳の精度に直結するため、対応候補の選択は重要なステップである。

以下、対訳文アラインメントの各処理を説明する。

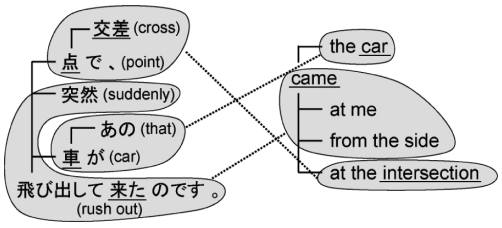

図1　アラインメントの例

2.1　日本語文と英語文の依存構造解析

日本語文の解析には、日本語の形態素解析システムJUMAN、依存構造解析システムKNPを用いる。これらは日本語文の構造を非常に高精度に解析することができ、新聞ドメインでは、形態素解析が99%、構文解析が90%の精度である。またこれらのシステムは、新聞以外のドメインの文に対しても、十分な精度での解析が可能である。日本語の依存構造の単位（ノード）は、各自立語が1ノードとなるもので、助詞、接辞、助動詞などは自立語のノードにまとめる。

英文については、Charniakパーサを用いて句構造に変換し、そこからheadを定義するルールによって依存構造に変換する。英語の依存構造の単位（ノード）は、日本語と同じく、各自立語が1ノードとなるもので、前置詞や助動詞は自立語のノードにまとめた。

2.2　語（列）の対応候補探索

日英間の語、語列の対応候補探索には、対訳辞書、Transliteration、数字のマッチングなどいくつかの手がかりを利用する。

対訳辞書は、日本語側の各形態素や連続した複数の形態素と、英語側の各単語や連続した複数の単語との組み合わせを辞書中から探し、見つかったものを対応とする。

さらに国語辞書から語の同義関係や上位下位関係を抽出し、これを用いて日本語のマッチングを柔軟に行なうことにより、対訳辞書の拡充も行なう。たとえば対訳辞書中に「さしあたり⇔for the time being」というエントリーはあるが、「当面」という語に関するエントリーがない場合でも、国語辞書に「さしあたり＝当面」という同義情報があるため、「当面⇔for the time being」という対応をとることができる。

Transliterationは、固有名詞、すなわち辞書によって人名・地名（の可能性）となるものとカタカナ語（未知語が多い）に対して用いる。これらの語に対応する英語綴りの候補を自動的に生成し、それと英単語列の編集距離に基づく類似度を定義し、それが閾値以上の英単語列があれば対応をつける。

たとえば次のような語がTransliterationによって対応付けられる

新宿→ Shinjuku ⇔ Shinjuku （similarity：1.0）
ローズワイン→ rosuwain
⇔ rose wine （similarity：0.78）

これらの語は辞書で対応付くことは極めて少ないが、この方法によって非常に高精度に対応つけることができる。

数字のマッチングは、それぞれの言語において異なる数字表現を算用数字に汎化することにより、対応候補を得る。例えば日本語の「二百六十万」と英語の"2.6 million"は共に同じ数字"2600000"を表しているため、それぞれ汎化することにより対応候補とすることができる。

2.3　適切な対応候補の選択

前章で得られた対応候補の中には、曖昧性を持つ候補や、曖昧ではないが文脈上不適切な候補が含まれることがある。

例えば【図2】において、日本語の「保険」と英語の"insurance"はそれぞれ2度ずつ出現しており、組み合わせで4つの対応候補が得られることになり、曖昧性が生じる。さらに「申し立て」の訳語として"file"と"claim"の2つがみつかり、ここでも曖昧性が生じる。このため、見つかった対応候補の中から適切な候補のみを選び出す基準が必要となる。これについては3章で詳しく述べる。

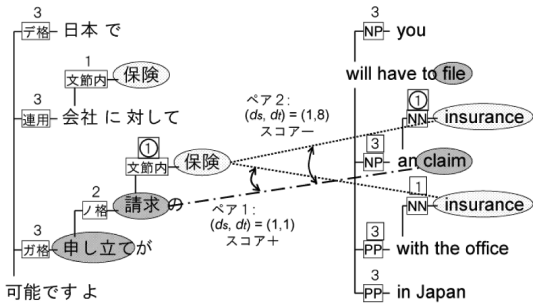

図2　曖昧性解消の例

2.4　未対応部分の推定

ここまでの処理により対訳文間にいくつかの対応が見つかったが、いくつかのノードが対応付けられずに残る場合がある。これらのノードは簡単なルールにより他の対応に併合する。

まず日本語、英語ともに名詞句内で未対応部分があれば名詞句内の他の対応に併合し、それ以外の未対応ノードはすべて親ノードの対応に併合する。ただし、節の区切りなどの大きな区切りを越えての併合は行わない。

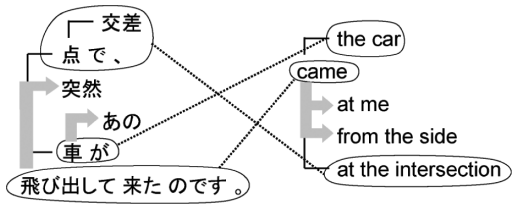

図3　対応の併合の例

【図3】に併合の例を示す。

ここまでで構築された対応を基本対応と呼ぶことにする。

2.5　用例データベースの構築

基本対応がえられたら、各基本対応と、日英両側で連続である（依存構造のどこかで親子関係にある）基本対応の組み合わせすべてを用例として登録する。

【図1】の例からは、それぞれの基本対応と、それを組み合わせた（日本語側で）「交差点で、突然飛び出して来たのです」や「突然あの車が飛び出して来たのです」などを用例として登録する。

3　整合性尺度に基づく構造的句アラインメント

対訳文全体として整合的なアラインメントを行うために、任意の一組の対応に対して整合性スコアを定義する。最も整合的なアラインメントは整合性スコアの平均を最大とするような対応候補の組み合わせとして得られる。

$$\arg\max_{alignment} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{\text{整合性スコア}(a_i, a_j)}{n(n\text{-}1)} \qquad (1)$$

整合性スコアは二つ一組の対応候補に対して計算され、対応候補ペアの関係が適切ならばプラス、そうでなければマイナスのスコアとなる。対応候補ペアの関係は、原言語・目的言語のそれぞれの木構造上の距離の関係で表される。木構造上の距離とは、あるノードから別のノードまでたどる際に、どれだけの枝を通るかということである。このとき、各枝は係り受けの強さによって重み付けされている（係り受け関係が強いほど枝の距離は小さく、関係が弱いほど距離は大きい）。

例えば【図2】において、ペア1は両言語において距離が小さく、適切な関係と判断できるので、プラスのスコアを与えるが、ペア2は一方では距離が小さく、他方は大きいので、不適切であると判断し、マイナスのスコ

アを与える。依存構造木においては、このようなことは起こらないはずだからである。

整合性スコアを用いて、あらゆるアラインメント候補の取捨選択のパターンの中から、（1）式により最適なアラインメントを得ることができる。

4 特許日英対訳文のアライメント結果

本手法を、AAMTから提供された特許日英対訳文に適用した。結果の具体例を【図4】および【図5】に示す。

特許文では、文が比較的長いこと、カタカナを含む専門用語が多いことなどの問題があるが、構文解析による構造の把握、transliterationによるカタカナ語などの扱い、整合性の観点からの対応の曖昧性解消などはいずれも有効に機能している。特許文の次世代高度翻訳支援、さらに自動翻訳の可能性を十分に示す結果であると考えられる。

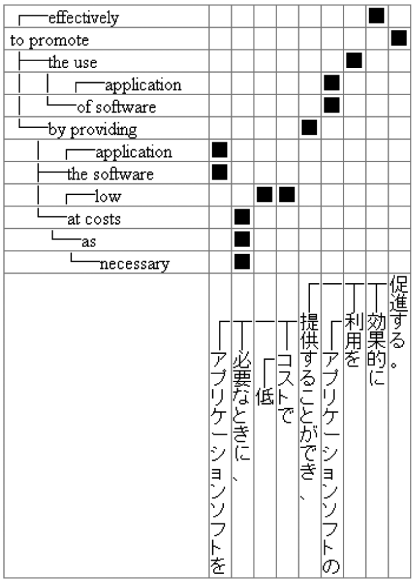

図4　アラインメント結果1

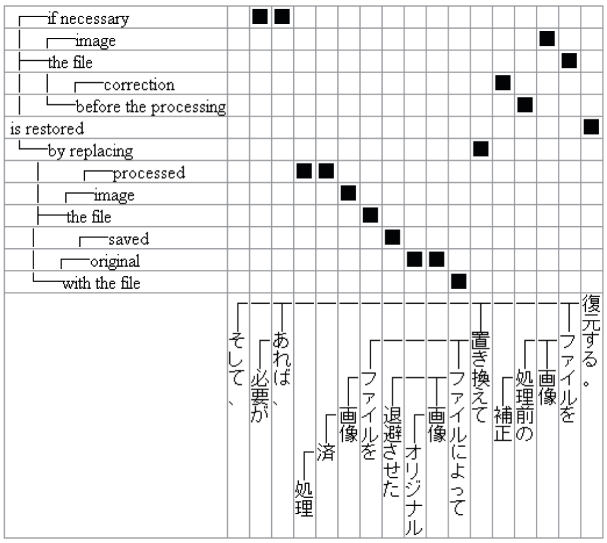

図5　アラインメント結果2

5 まとめと今後の課題

本稿では、我々が開発した用例翻訳システムのアライメントモジュールについて説明を行った。本手法を特許日英対訳文に適用したところ、特許文の次世代高度翻訳支援、さらに自動翻訳の可能性を十分に示す結果がえられた。

今後の課題として、アライメント自動評価データを構築し、対訳特許文の特性をより詳細に検討すること、さらに特許文翻訳の可能性について検討を行うことが考えられる。

参考文献

[1] 中澤敏明, Yu Kun, 黒橋禎夫：“整合性尺度を用いた構造的対訳文アラインメント”言語処理学会 第13回年次大会 pp.368-371 （2007.3）.

[2] Toshiaki Nakazawa, Kun Yu, Daisuke Kawahara, Sadao Kurohashi：Example-based Machine Translation based on Deeper NLP In Proceedings of International Workshop on Spoken Language Translation（IWSLT'06）, pp.64-70, Kyoto （2006.11）.