翻訳用例の柔軟な結合に関する考察

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.kuee.kyoto-u.ac.jp　TEL 075-753-5346

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

現在の機械翻訳手法には、大きく分けてルールベース翻訳（RBMT）、統計翻訳（SMT）、用例ベース翻訳（EBMT）の３つがあり、世界的に最も多く研究されている手法は SMT である。SMT は欧米の言語間など、語順などの言語体系が似た言語対では、すでに実用レベルのシステムが構築されているが、日本語と英語などの言語構造の大きく異なる言語対に対しては、未だに十分な精度を達成していない。このような言語対を扱う際には、言語構造の違いを克服し、高精度な翻訳を実現するために、各言語の構造解析技術を積極的に利用した翻訳手法を適用する必要があると考えられる。

このような構造的言語処理を志向する翻訳手法の一つとして、単語依存構造に基づく用例ベース機械翻訳システムがある [1]。言語間で大きな語順の違いがあっても単語の依存関係は保存されていることが多く、依存構造解析を行うことによって、単純な単語列では扱えないような柔軟な翻訳を行えることが期待される。そこでは以下の手順により翻訳を行う。まず入力文を依存構造解析し、依存構造木に変換する。次に全ての部分木について、翻訳に利用可能な用例をあらかじめ構築されている用例データベース（翻訳知識）から検索する。最後に、様々な尺度（用例の大きさや翻訳確率など）を用いて、検索された全ての用例の中から翻訳として最適な用例の組み合わせを選択し、それらを入力文のヘッドをカバーするものから順に張り合わせて目的言語側の依存構造木を構築することにより翻訳文を生成する。

2 用例の結合

【図 1】に日英用例ベース機械翻訳システムの概要を示す。【図 1】では入力文の翻訳に３つの用例（太線で囲まれた部分）を使っている。用例同士を結合する際には、糊代となるノードを手がかりとして利用する。糊代とは、対訳文のうち実際に翻訳で利用される部分（ボディと呼ぶ。図では太い線で囲まれた部分）に、両言語において木構造上連続な単語のことである。糊代の部分に他の用例を張り付けることにより用例を結合することができる。図では一つ目の用例の“水素”と“石油”の部分が糊代となっており、“水素”の部分を“ウイスキー”に、“石油”の部分を“オオムギ”にそれぞれ置き換えることにより入力依存構造木を再現することが出来る。ここで“水素”に対応する英語は“hydrogen”であるということがあらかじめ知識として獲得されているので、“ウイスキー”の対応する英語“whisky”はこの“hydrogen”

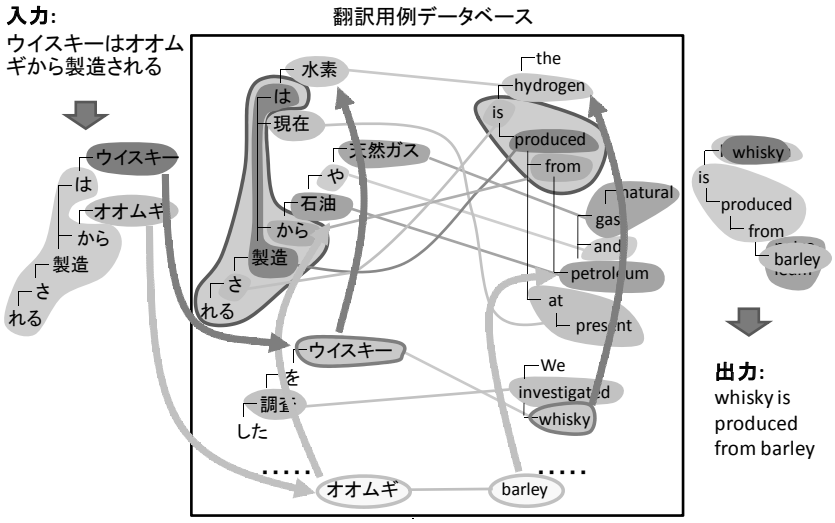

図 1　用例ベース機械翻訳システムの概要

の部分に貼り付ければよいことがわかる。同様に“オオムギ”の対応する英語“barley”は“石油”の対応する英語“petroleum”の部分に貼り付ければよい。このようにして、目的言語側の依存構造木および最終的な出力文が生成される。

言語ごとに単語の依存関係は保存されていることが多いが、そうでないケースも存在する。このような場合、上で述べた用例同士の結合が単純には行えない。本稿では用例同士の結合が単純には行えない場合についての考察を行う。

3 用例の柔軟な結合

本章では、上で述べた基本となる用例の結合では扱えない場合について検討する。

3.1 目的言語側の糊代がボディと不連続

【図 2】のような対訳文から学習された翻訳知識を翻訳で用いることを考える（ここでは助詞の細かい扱いについては省略する）。【図 3】のように“He wants to swim.”という文を翻訳することを考える。翻訳知識として“wants to ⇔ たい”が利用でき、さらに“eat”の部分が“swim”を翻訳するための糊代として機能する

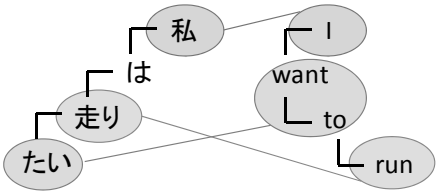

図 2　対訳例

ことがわかるが、“He”を翻訳するための糊代として機能するであろう“I”は、その対応先である“私”が目的言語側で“たい”と木として不連続な位置にあるため、単純には糊代として利用できない。糊代情報を利用せずに“He”を翻訳した場合、その訳となるであろう“彼は”を目的言語側でどこに貼り付ければよいかの情報がなくなってしまい、正確な翻訳が行えなくなってしまう。

このような場合にはボディと連続な糊代から順に他の用例に置き換えることにより、目的言語側で不連続な糊代であっても適切に利用することができる。【図 3】の場合では“eat”を先に“swim”の用例で置き換え、次に“I”を“He”の用例で置き換えることにより、適切な訳“彼

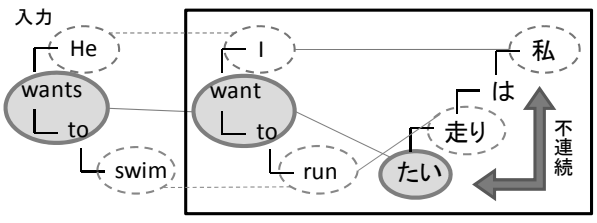

図 3　目的言語側の糊代が不連続な例

は泳ぎたい”を生成することができる。

3.2　原言語側の糊代がボディと不連続

次に同じ用例を用いて、【図 4】のように“彼は泳ぎたい”という文を翻訳することを考える。これは 3.1 節の原言語と目的言語を入れ替えたものである。今までの例では用例の原言語側において、全ての糊代はボディと連続であったが、【図 4】の場合は利用したい糊代が原言語側で不連続だが、目的言語側で連続となっている。このような場合には糊代として利用可能な部分のさらに外側についても、糊代として利用可能な部分があるかどうかをチェックすることが必要となる。糊代の先の糊代が利用可能であることがわかれば、用例の貼り付けは通常と全く同じように行うことができる。

以上の 2 パターンはモダリティを含む翻訳では頻繁に起こるため、これらを正確に扱うことは非常に重要である。

3.3　糊代内の依存関係の逆転

【図 5】のような場合を考える。入力の“駅”を翻訳するための糊代は用例の“2”の部分となるはずであるが、この対応する英語“2nd”はボディと不連続な位置にある。これは複合名詞内の依存関係が逆転しているためであるが、複合名詞全体を一つの糊代として拡大解釈すれば、通常の糊代と全く同様に扱うことができる。

ただし、いつでも拡大解釈してよいというわけではなく、以下の条件を満たすもののみ許容する。

条件：糊代とする部分（図の“2”の部分）の対応先（“2nd”）から、ボディ（“went to”）までのパス上に存在する全ての単語（“terminal”）について、それぞれの対応先（“ターミナル”）が元の糊代よりも下位にある

ここで「下位」とは、ある単語から見て葉の方向という意味である。この条件を満たすものは直接の係り受け関係になくとも、複数単語をひとかたまりとして見ることにより、糊代として利用することができる。

3.4　ギャップを含む用例を使う

【図 6】のような対訳文を用いて翻訳を行うことを考える。英語のヘッドとなる“has”は日本語側には明確な対応先がないが、翻訳での利用を考えると各言語のヘッド同士は対応関係にあるほうが望ましい。そうでないと“象は鼻が長い”を翻訳した際に出力から“has”がなくなってしまうことになり翻訳誤りを起こす。そこで“has”は“長い”の対応先である“long”にマージする。すると、“long hair”は単語列としては連続であるが、木構造上では不連続となってしまう。このような言語間の構造のずれは、特に日本語と英語のような大きく異なる言語間ではしばしば起こりうるものである。

このようなギャップを含む用例を用い、ギャップの部分を糊代として他の用例を貼り付けることにより、言語間の構造のずれが吸収できる。ただし、ギャップを含

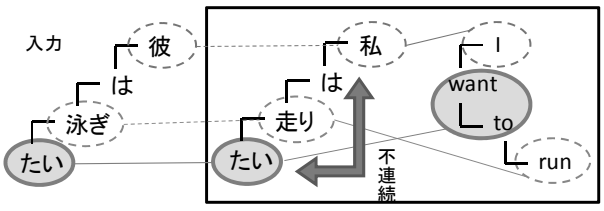

図 4　原言語側の糊代が不連続な例

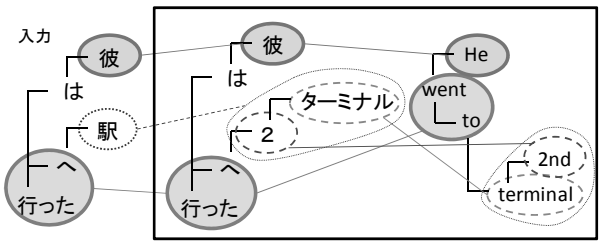

図 5　糊代内での依存関係の逆転

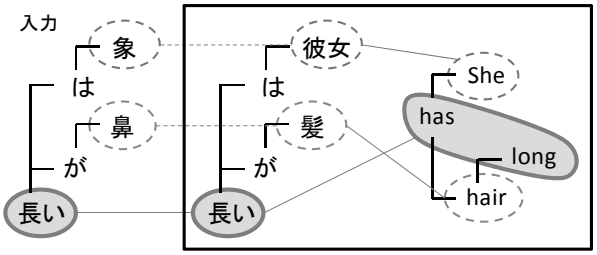

図 6　ギャップを含む用例

む用例を翻訳で利用する際には、ギャップの部分が糊代であり、必ず他の用例で置き換えられることが必須である。

4 まとめと今後の課題

本稿では言語間の構造の違いを柔軟に吸収するために、用例の張り合わせの際の手がかりとなる糊代をより柔軟に利用することを検討した。今後は提案した方法が実際に翻訳に有効であることを定量的に評価する必要がある。

[参考文献]

[1] Toshiaki Nakazawa and Sadao Kurohashi
Fully Syntactic EBMT System of KYOTO Team in NTCIR-8
In Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-8), pp.403-410, Tokyo, Japan (2010.6).