4S−5

用例をもちいた依存関係単位の翻訳

幸山　秀雄　　隅田　英一郎
(株)ＡＴＲ自動翻訳電話研究所

1　はじめに

機械翻訳システムをより使いやすく実用的なものとするためにはシステムの大規模化が必要である。大規模なシステムを構築するためには拡張性や保守性のすぐれた翻訳手法が期待される。その１つの方法として用例(原文と対訳)をもちいた翻訳 [1] [2] が考えられる。本論文では原文と対訳との部分対応データ(以下、用例データと呼ぶ)を使った文レベルの日英翻訳手法(部分変換 - 合成手法)について述べる。この手法によりデータ自身が多くの翻訳処理役割ををもつことにより、手続きは軽くなり、システムの拡張性や保守性が良くなる。また、データが後述する組合せの効果をもつため、より少ないデータで多くの文を処理できる見通しがもてる。

2　部分変換 - 合成手法

2.1　概要

ある１つの構造があったときに、それを任意の単位の部品に分解し、再び同一の構造に組み立てることを考える。多くの数の構造があったときに、それらを分解した部品を持つことによって組合せの効果により元の数より多くの構造が生成することができる。裏返せばこのことは、多くの入力構造を処理できることを意味する。

この考えを翻訳に適用するなら、日英の文の依存構造を構成する部分依存構造の対応を部品(用例データ)として持てばよい。そして入力を受ければ部品ごとに独立に変換し組み立てることにより翻訳をおこなう。組み立てられた構造は、構造としての整合性と入力構造との意味的な同一性を保持していなければならない。この制約を満たした構造だけが正しい構造と呼べる。このため、組み立てられた構造は構文制約によりチェックされる。また入力依存構造と組み立てた依存構造の意味的な同一性は、意味的に同一な部分対応構造を組み立てることによって全体の構造の意味も同一になるという前提に立つ。このように構造を部分単位に独立に変換し組み立てる手法を部分変換 - 合成手法と呼ぶことにする。なお変換処理では部品の単位が変換の単位となり、組み立ては変換された部分構造の要素の一致によりつなげることによりおこなわれる。

Translation by Dependency units of Examples
Hideo KOHYAMA, Eiichiro SUMITA
ATR Interpreting Telephony Research Laboratories

2.2　用例データ

用例データは日本語文の文節間の依存関係と、対応する英語対訳文の部分依存構造を記述したものである。英語の部分依存構造にはホーンビー [3] に基づいた構文情報が付与されている。用例データの形式の概略を示す。

用例データ := [日本語依存単位, 英語依存構造]
日本語依存単位 := [文節１ 文節２]
文節 := [番号: 自立語... 番号: 附属語...]
英語依存構造 :=
[番号: 単語１(構文情報１) 単語２(構文情報２) ...]

但し、文節２は文節１に依存し、番号は日本語依存単位内の通し番号で英語単語と対応する。単語２は単語１に依存する。

2.3　基本アルゴリズム

基本アルゴリズムを以下に示す。なお入力は解析された日本語依存構造である。

i 部分変換
日本語入力構造から２項の依存関係単位を取りだし、用例データをもちいてそれぞれ独立に変換する。

ii 合成
構造の整合性をチェックしながら、対訳の依存要素の一致により構造を組み立てる。

iii 優先制約
合成された英語依存構造が複数あれば用例データの頻度情報などにより各構造に優先順位をつける。

vi 生成
英語構文規則を用いて語順を決定する。

具体例を図１. に示す。「私は彼に登録用紙を送る」という依存構造が入力されれば、「送る - 私は」、「送る - 彼に」、「送る - 登録用紙を」という３つの依存関係単位を取り出す。次に用例データを使いそれぞれ独立に変換し、部分依存構造である send(Vt)-I(S), send(Vt)-him(Oi), send(Vt)-a_registration_form(Od) を得る。変換されたそれぞれの英語依存構造は構文としての整合性をチェックしながらその要素の一致 (send(Vt)) により合成され、１つの構造として組み立てられる。複数構造が合成される場合 (解が複数の場合) は頻度情報により、それぞれの構造の優先順位を決める。最後に構文規則により語順が決定され英語文が生成される。

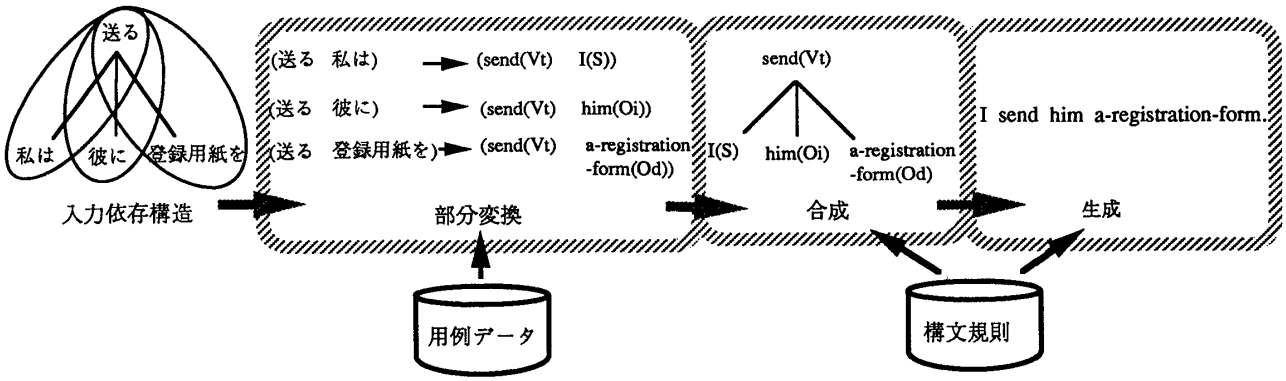

図1. 部分変換-合成手法の具体例

3 組合せの効果

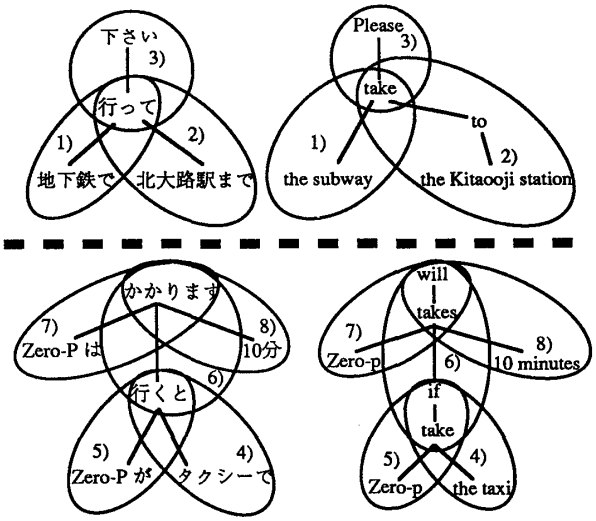

図2. AとBの依存構造

用例文の日本語文節を組み合わせることによって、用例文数以上の入力文が処理可能となる。例えば次のAとBの用例文の組合せの効果からaからeの5つの文が処理可能である (図2参照)。なおZero-P はゼロ代名詞を示し、1), 2),... は図2での部分依存構造の日英対応を示す。

A. 地下鉄で北大路駅まで行って下さい。
Please take the subway to the Kitaooji station.
B. タクシーで行くと10分かかります。
If you take the taxi, it will takes 10 minutes.

a. 1) と 3) の合成
地下鉄で行って下さい。
Please take the subway.

b. 1)、2)、5)、6)、7)、8) の合成
地下鉄で北大路駅まで行くと10分かかります。
If Zero-P take the subway to the Kitaooji station, Zero-P will takes 10 minutes.

c. 1)、5)、6)、7)、8) の合成
地下鉄で行くと10分かかります。
If Zero-P take the subway, Zero-P will takes 10 minutes.

d. 2)、3)、4) の合成
タクシーで北大路駅まで行って下さい。
Please take the taxi to the Kitaooji station.

e. 3)、4) の合成
タクシーで行って下さい。
Please take the taxi.

入力構造と合成された構造には上述aの様に、元の用例文Aの部分構造になっているものとそうでないものとがある。前者を部分木タイプ、後者を合成木タイプと呼ぶことにする。bとdは用例文AとBの合成木タイプである。またcはbの部分木タイプ、eはdの部分木タイプである。合成木タイプでは2つの用例文の中に同じ文節が1つ存在すれば2倍の入力文を処理できる可能性がある。またその合成木タイプが部分木タイプを含んでいればさらに処理可能な文数が増える。

4 おわりに

用例データの組合せにより処理可能な文の数が増えることを示した。また1つの日本語表現に対し複数の英語表現が生成される場合がある。これは部分構造の、合成タイプの組合せによりパラフレーズをおこなっているとも言える。なお今後、翻訳実験を通しどのくらいの組合せの効果が起こるかを明確にしていく予定である。

5 謝辞

本研究の機会を与えてくださったATR自動翻訳電話研究所、樽松社長、適切な助言を述べられた言語処理研究室、飯田主幹研究員、また熱心に議論に参加したデータ、言語処理研究室の諸氏に感謝します。

参考文献

[1] 隅田ほか,“用例に基づいた翻訳”, 情報処理学会第40回全国大会,1990
[2] 佐藤,“Toward Memory-based Translation”, Coling'90,1990
[3] ホーンビー,“英語の型と語法”, オックスフォード大学出版局,1981