クロスリンガル情報発信のためのキャプション翻訳処理

6V－8

井ノ上 直己　鈴木 雅実　橋本 和夫
KDD 研究所

1. はじめに

近年のインターネットの普及に伴い，日本から積極的に世界に向けて情報発信したいという希望も増えているが，日本語コンテンツを発信するには発信者自身が他言語（例えば，英語）へ翻訳したコンテンツを用意する必要があり，この翻訳に対する負担が問題となる．

そこで，この翻訳を機械翻訳で行うことを考える．しかし，現状の機械翻訳技術ではいかなる文章であっても完全に翻訳できるところまで技術レベルが到達しているとは言い難い．そのため，日本語コンテンツとして画像とそのキャプションが載っているコンテンツを対象とし，翻訳をキャプション表現に限定することで高い翻訳性能の実現を目指している．

日本語コンテンツとして日本美術データを用いて，まずキャプションの日本語表現の分析を行った．その結果，複合名詞が大多数であったことから，複合名詞に有効な翻訳手法の検討を行った．本稿では，その概要を示す．

2. システム構成

日本語以外を母国語とする人に対して日本語コンテンツを発信するシステムの構成は図1に示す通りであり，予め作成する他言語コンテンツ（ここでは，英語コンテンツ）を機械翻訳により作成する．

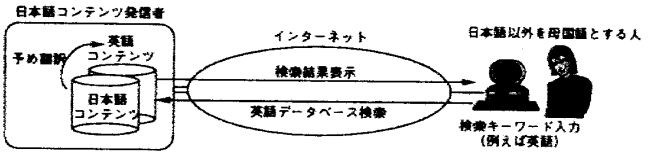

図 1: クロスリンガル情報発信システムの構成

A Caption Translation Method for Crosslingual Information Distribution
Naomi Inoue, Masami Suzuki, Kazuo Hashimoto
KDD R&D Laboratories
2-1-15 Ohara, Kamifukuoka, Saitama 356, Japan

表 1: タイトル表現種別

複合名詞	54.9%
名詞 1 語	19.7%
(複合) 名詞＋助詞＋(複合) 名詞	19.0%
その他	6.3%

3. 翻訳例の分析

日本美術データとして国際日本文化研究センターが提供している浮世絵データを用いて，翻訳パターンの分析を行った．このデータには画像の他，絵のタイトル，作者，所蔵場所などがキャプション情報として付与されており，既に英語を母国語とする翻訳者により，日本語から英語への翻訳結果が付与されている．とりわけ表現パターンにバリエーションがあり翻訳処理で問題となるのはタイトルであるため，タイトルの翻訳パターンについて分析を行った．

3.1 日本語タイトル表現の特徴

多くの場合タイトルは名詞句となっており，異なる表現パターンは比較的少なく，また，使われる単語は専門用語がほとんどであり，文の長さも一般の文と比べて長くないという特徴がある．約 1100 タイトルの表現の分析を行った結果を表 1に示す．表より，名詞 1 単語だけからなるタイトルは全体の約 20%にすぎず，約 55%が複数の名詞を連結した複合名詞であることがわかる．また，「名詞＋の＋名詞」のように名詞と名詞との間が 1 つの助詞で連結された名詞句とを併せると約 94%となり，大半が名詞と名詞とが何らかの繋がりをもって組み合わされた表現であることがわかる．

3.2 複合名詞の翻訳パターンの特徴

前節に示したようにタイトル表現の大多数は複合名詞であり，複合名詞の翻訳性能がシステム性能を大きく左右すると考えられる．表 2に，実際の翻訳例から得た複合名詞の翻訳パターンを示し，それぞれの例を従来のルールベースの翻訳方式によ

り翻訳する場合について考察する.

表 2: 複合名詞の翻訳例

語順が同じ翻訳例	
日本語	英語
仮名手本中心蔵	Kanadehon Chuushingura
北条入道時頼	Hoojoo Nyuudoo Tokiyori
恵比寿渡海	Ebisu Crossing the Sea
語順が変わる翻訳例	
日本語	英語
神功皇后	Empress Jingoo
相州袖ヶ浦	Sodegaura, Sooshuu
観瀑猿猴	Monkey Watching a Waterfall
日本語にない英単語が訳出される例	
日本語	英語
江戸梅屋敷	Plum Estate in Edo
見立源氏	Mitate of Genji
縁先官女	Court Lady on a Veranda

表2より，まず日本語と英語の単語が過不足なく対応し，しかも翻訳結果も日本語の語順と同じ場合がある．この例では単純な辞書引きで翻訳可能である．ただし,「恵比寿渡海」の「渡海」に対する英語は “cross the sea” と辞書に記述してあり，訳出する場合は構文的に正しく (この例では ing 形) 変形しなければならない.

また，日本語と英語の単語が過不足なく対応しているが，肩書きを英語では名前の先に表現するため，あるいは地名のより上位のものを後で表現するため，日本語と英語とで語順が変わる例がある．このような例では，語順が変わる場合の規則を明確にしなければならない.

さらに，日本語にはない英単語が英語で訳出される場合があり，複合名詞を構成する名詞間の意味的関係により訳出される単語が異なる．そのため，名詞間の意味関係を捕えなければならない.

4. 複合名詞のための翻訳処理手法

複合名詞の翻訳には，辞書引きだけで可能な場合，名詞間の意味的な関係に基づいて翻訳を行う場合，がある.

しかし，複合名詞は名詞を組み合わせることで無限に生成できるため，名詞間の意味的な関係を系統だてて明らかにすることは非常に困難であり，従来の翻訳方式では限界がある．そのため，名詞間の意味的な関係は明らかにする必要がなく，経験的知識として蓄積された用例の中から入力表現と最も類似する用例を求め，その用例の対訳情報を利用して翻訳する用例ベースの翻訳方式[1]が適切である.

用例ベースの翻訳方式は,「名詞＋の＋名詞」の翻訳に適用されその有効性が示されている[2]が，複合名詞は容易に単語を組み合わせることができることから，用例ベースの翻訳方式を複合名詞に適用するには入力表現と用例との類似度計算において，単語数が異なる用例に対しても類似度が計算できるように拡張する必要がある．例えば,「日本橋略図」という入力表現に対し,「日本橋一丁目略図」や「日本橋の略図」といった用例との類似度計算を行う．DP マッチングアルゴリズムは，このような類似度計算に適用可能であると考えられる.

5. まとめ

本稿では，日本美術データを対象にタイトルの翻訳手法の検討を行った．実際の日本語タイトルを分析した結果，複合名詞と「名詞＋助詞＋名詞」表現が大多数を占めることが明らかになった．複合名詞などにおいて名詞間の意味関係を求めることは困難であることから，用例に基づく翻訳方式を用いるが，単語数の異なる用例とも類似度比較が可能なように，DP マッチングアルゴリズムの適用について検討を進めている．今後は，この翻訳手法を実装して翻訳性能の評価を行う予定である.

本研究を行うにあたり，日本美術データを提供していただいた国際日本文化研究センターの山田助教授および関係諸氏に感謝いたします.

参考文献

[1] Nagao, M.: *"A Framework of a Mechanical Translation between Japanese and English by Analogy Principle"*, Artificial and Human Intelligence, Elithorn, A. and Banerji, R.(eds.), North-Holland, pp.173-180(1984)

[2] Sumita, E. and Iida, H.: *"Example-Based Transfer of Japanese Adnominal Particles into English"*, IEICE Trans. Information and Systems, vol.E75, no.4, pp.585-594(1992)