対訳コーパスを用いた省略可能情報に基づく格フレームの自動獲得

三上 優　越前谷 博　桃内 佳雄

北海学園大学大学院工学研究科

1 はじめに

格フレームは機械翻訳において，意味的な情報が付与された重要な言語知識の一つである．格フレームの自動獲得の研究としては対訳コーパスを利用する手法[1]や Web 上の大規模なテキストを利用する手法[2]が提案されている．しかし，これらの手法は，構文解析ツールに強く依存するため，多言語への適用が困難であることが問題となる．本稿では，様々な言語の格フレームの自動獲得を目的とした新たな手法を提案する．本手法は，構文解析ツールが豊富に存在し，かつ，最も広く利用されている英語を中間言語として利用することで，構文解析ツールが十分には得られない言語においても格フレームを自動獲得する．例えば，日本語とスペイン語を構文解析ツールが不十分な言語と位置付けた場合，日本語-英語-スペイン語間対訳コーパスを用いて，日本語-スペイン語間の格フレームを自動獲得する．また，本手法では，名詞句を効率よく抽出するために，文法知識として名詞句ルールを自動獲得する．

2 処理過程

2.1 高頻度の名詞句の抽出

本手法では，始めに対訳コーパスから出現頻度が２以上の名詞句を抽出する．構文解析ツールの豊富な英語を中間言語に用いることで，英文については構文解析ツールより名詞句を決定できる．そして，その英語の名詞句に対応する部分を英語以外の言語の文から抽出する．

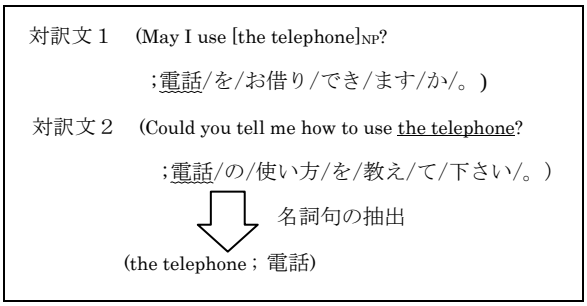


図１　高頻度の名詞句の抽出例

図１に高頻度の名詞句の抽出例を示す．図１では対訳文１に対して構文解析ツール用いることで，”the telephone”が名詞句となる．次いで，”the telephone”に対応する名詞句を日本文から決定するために，”the telephone”が出現する対訳文２を対訳コーパスから選択する．そして，日本文間の共通部分[3]を決定する．図１においては”電話”が共通部分となる．そして，”the telephone”と”電話”との間の部分間類似度[4]を求め，閾値以上であれば，”電話”を対応部分とする．また，共通部分が複数存在する場合には，”the telephone”との部分間類似度が最も高く，かつ閾値以上のものを対応部分とする．図１においては共通部分が”電話”のみで，かつ部分間類似度が閾値を上回っているため名詞句として(the telephone；電話)が抽出される．このような処理により，英文以外の言語文においても構文解析ツールに強く依存することなく名詞句を決定することが可能となる．

2.2 名詞句ルールの自動獲得

本手法では，低頻度の名詞句を抽出するために名詞句ルールを自動獲得する．図２に名詞句ルール獲得の具体例を示す．

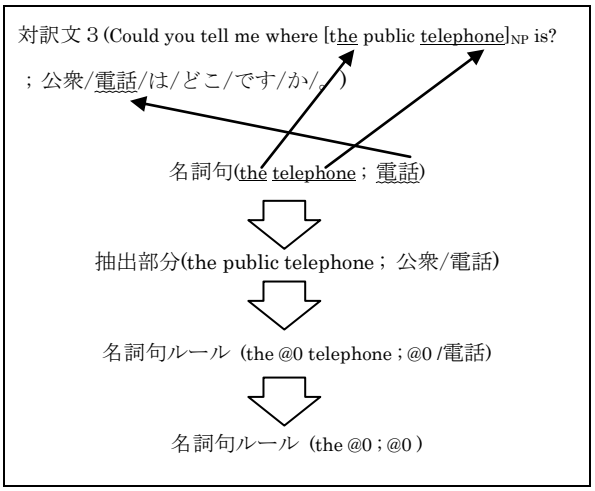


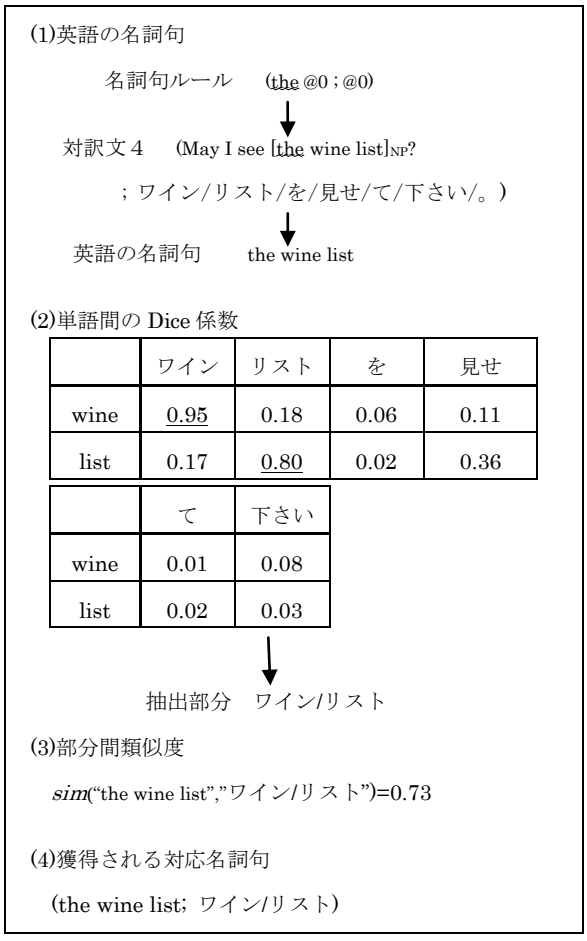
図２ 名詞句ルールの獲得の具体例

始めに，2.1 より抽出した名詞句を含む対訳文を対訳コーパスより選択する．図２では”the”と”telephone”, ”電話”を含む対訳文３が選択される．

次いで, 英文の名詞句”the public telephone”の先頭単語”the”と末尾の単語”telephone”が名詞句(the telephone； 電話)の英語部分と一致するため，”the public telephone”が抽出対象となる．その際, ”the”と”telephone”に挟まれた”public”を省略可能な部分[4]と位置付ける．日本文においては，一致する部分が”電話”のみであり，抽出範囲が決定できないため bigram 確率と部分間類似度に基づき”the public telephone”に対応する部分を決定する．その結果，図２では”公衆/電話”が抽出される．この場合は，省略可能な部分は”電話”以外の”公衆”となる．そして，省略可能な部分”public”と”公衆”を変数に置き換えることにより一般化された(the @0 telephone；@0/電話)を名詞句ルールとして獲得する．この名詞句ルールに対して更なる一般化を行うことで，より一般化された名詞句ルール(the @0 ； @0)が獲得される．

2.3 名詞句ルールに基づく低頻度の名詞句の抽出

2.2 より獲得された名詞句ルールを用いることで，対訳コーパス中の低頻度の名詞句を効率よく抽出することが可能となる．図３に名詞句ルールに基づく名詞句の抽出例を示す．



	ワイン	リスト	を	見せ
wine	0.95	0.18	0.06	0.11
list	0.17	0.80	0.02	0.36

	て	下さい
wine	0.01	0.08
list	0.02	0.03

抽出部分 ワイン/リスト

図３ 名詞句ルールに基づく名詞句の抽出例

図３では，構文解析ツールにより”the wine list”が名詞句として決定され，先頭単語の”the”が名詞句ルールと一致するため，対訳文４を用いた．次いで，この”the wine list”に対応する部分を日本文より抽出する．名詞句ルール(the @0；@0)の日本語部分は”@0”であるため，日本文全体が抽出対象となる．

そこで，名詞句”the wine list”の”the”を除く”wine”と”list”と最も高い類似度を持つ単語を日本文より決定する．その際，類似度の計算方法には Dice 係数を用いる．図３より，”wine”においては”ワイン”，”list”においては”リスト”が最も高い類似度を示した．

したがって，日本文からは”ワイン”から”リスト”までの”ワイン/リスト”が”the wine list”に対応する部分として抽出される．この”the wine list”と”ワイン/リスト”間の部分間類似度を求め，閾値以上であれば，名詞句として抽出する．

名詞句ルールを用いた名詞句の抽出は共に冠詞を持つ言語においては特に有効となる．例えば，英語-スペイン語間において，名詞句ルールとして(a @0 ; un @0)が獲得できていれば，スペイン文から冠詞”un”を含む正しい名詞句を抽出することが可能となる．

2.4 格フレームの獲得

本手法では，名詞句を効率よく抽出することで格フレームを自動獲得する．図４に格フレーム獲得の具体例を示す．

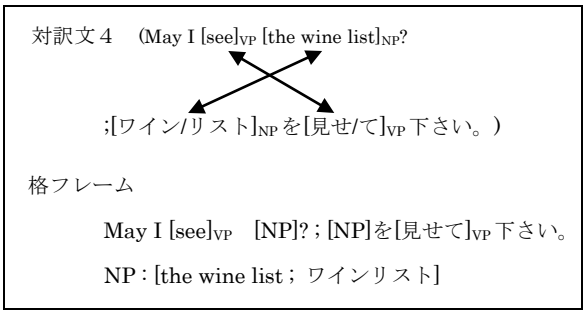


図４ 格フレームの獲得例

図４では，2.3 より名詞句”ワイン/リスト”が抽出された対訳文４において，2.3 と同様の処理により，動詞句”see”に対する部分”見せ/て”を抽出することで，格フレームを自動獲得する．

3 性能評価実験

3.1 実験結果

本稿では，低頻度の名詞句の抽出に有効となる名詞句ルールの自動獲得についての性能評価実験を行った．本実験では，旅行用会話文[5]の日本語-英語-スペイン語の対訳文 1,042 組から名詞句ルールの自動獲得を行った．実験結果を表１に示す．表１より，英語-日本語間においては名詞句ルールの精度は 87.5%，英語-スペイン語においては 80.0%となった．

表１ 実験結果

英語-日本語間名詞句ルール	
正しい名詞句ルール	誤った名詞句ルール
7	1
精度 87.5%(7/8)	
英語-スペイン語間名詞句ルール	
正しい名詞句ルール	誤った名詞句ルール
4	1
精度 80.0%(4/5)	

3.2 考察

表１より，本手法では高い精度で名詞句ルールを自動獲得できることが明らかとなった．3.1 の実験より獲得された名詞句ルールの具体例を表２に示す．

表２の中の正しい名詞句ルール(the @0 ; @0)，(a @0 ; @0)，そして，(a @0 ; un @0)は，より一般化された名詞句ルールである．また，誤った名詞句ルールにおいては，全て変数の位置が誤っていた．

また，英語–スペイン語間の名詞句ルールについては，一般化された名詞句ルールが(a @0 ; un @0) のみであり，不十分である．スペイン語においては，英語よりも冠詞の数が多く，英語の“the” に対して，”la” や，”de” など，その数は１対多の関係にあるため，英語–日本語間に比べ，名詞句ルールの獲得が困難なことが原因と考えられる．

言語間の名詞句の決定においては，名詞句ルールは重要となることから，多くの名詞句ルールを獲得する必要がある．

4 まとめ

本稿では英語を中間言語とした，多言語の格フレームの自動獲得のための手法を提案した．そして，名詞句を決定するための有効な言語知識である名詞句ルールの獲得についての性能評価実験を行った．その結果，高い精度で名詞句ルールを獲得できることを確認した．

今後は，より多くの名詞句ルールを獲得するための改良を行うと共に，格フレームの獲得のための実験を行う予定である．

表２　獲得された名詞句ルールの具体例

英語-日本語	
正しい名詞句	誤った名詞句
(a @0 room ; @0/部屋) (a @0 restaurant ; @0/レストラン) (a @0 tour ; @0/ツアー) (the @0 telephone ; @0/電話) (the @0 bus; @0/バス) (the @0 ; @0) (a @0 ; @0)	(a @0 hotel; ホテル/@0)
英語-スペイン語	
正しい名詞句ルール	誤った名詞句ルール
(the @0 city ; @0 la ciudad) (a @0 restaurant ; un restaurant @0) (a @0 car ; un coche @0) (a @0 ; un @0)	(a @0 car ; @0 un coche)

謝辞

本研究の一部は，北海学園大学 ハイテク・リサーチ・センター研究費(私立大学戦略的研究基盤形成支援事業)の補助によって行われている．

参考文献

[1]宇津呂 武仁，松本 裕治，長尾 眞，”二言語対訳コーパスからの動詞の格フレーム獲得”情報処理学会論文誌，vol.34，No.5，pp.913-924,May. 1993.

[2]河原 大輔，黒橋 禎夫，”高性能計算機環境を用いたWebからの大規模格フレーム構築”情報処理学会研究報告(2006-NL-171 (12))， pp.67-73，2006.

[3]荒木 健治,” 自然言語処理ことはじめ―言葉を覚え会話のできるコンピュータ”， 森北出版，東京，2004.

[4]寺島 涼，越前谷 博，荒木 健治，”学習型機械翻訳手法における省略可能性を用いた翻訳ルールの自動獲得とその有効性” 情報処理学会研究報告(2008-NL-183 (19))， pp.127-134，2008.

[5]ひとり歩きの４カ国語会話 ヨーロッパ編３ 自由自在 英語・フランス語・イタリア語・スペイン語，JTBパブリッシング，東京，2006.