大規模言語知識を用いた既存の語彙翻訳規則からの類似語彙規則半自動生成

E-55

定政 邦彦, 土井 伸一, 奥村 明俊
日本電気株式会社 マルチメディア研究所
k-sadamasa@az.jp.nec.com, s-doi@ah.jp.nec.com, a-okumura@bx.jp.nec.com

1 背景

機械翻訳の精度向上には、語彙ごとの訳し分けが必要である。訳し分けの知識を人手で作成するのは高コストであるため、従来、翻訳事例から機械的に語彙ごとの訳し方を表した語彙翻訳規則を獲得する用例ベース翻訳[1]や、シソーラスにより語彙翻訳規則の適用可能範囲を広げる手法[2]が提案されている。しかしながら、入力に意味的に近い語彙規則を翻訳に利用する手法では、事例数が少ない場合に、語彙規則の適用範囲を過度に広げてしまう可能性があった。本手法では、語彙規則の一般化を4種類の基準で抑制することで、語彙規則の適用範囲を適度に広げる手法を提案する。本稿では、そのうちの1基準について詳細に報告する。

2 方法

本手法では、語彙規則中の名詞部分を、意味が類似する別の名詞に置き換えることで類似規則を生成する(図1)。

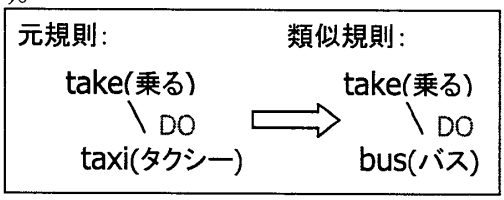


図1 類似規則の生成

本手法は、以下の4つのフェーズで構成される。

1. 規則中の置き換えを行う原言語名詞に関し、意味的に対応する原言語シソーラス上のノードを特定する。その際、規則中で与えられた訳語の意味に対応するノードを選択することで、類似規則の生成誤りを低減する。
2. 特定されたノードの周囲に位置する原言語単語を、置き換えを行う原言語名詞の類義語とする。類義語として認める範囲は、その中に含まれる単語の意味が散らばり過ぎない程度に狭める。これにより、抽象度の高い名詞が類義語として選択されるのを防ぐ。
3. 置き換えた原言語単語の訳語を選択する。選択基準には、(1)置き換え元の原言語単語との意味の近さ、(2)その訳語を選択した結果生成される訳出表現の頻度、(3)デフォルト訳語の優先、を用いる。
4. 大規模コーパスを用いて稀な表現に対する規則を除去し、また、同一の規則から生成された類似規則を比較することで、特異な規則を除去する。

本稿では、フェーズ1に関して詳細に述べる。シソーラスで既存規則の適用範囲を広げる際の問題点としては、元となる規則に多義語が含まれる場合、誤った語義に基づいて類似規則を生成する可能性があることが挙げられる。例えば、[keep,mouse]→[ネズミ,を,飼う]という規則があった場合、mouseには、パソコン周辺機器と動物の意味があり、前者の意味を元に類似規則を生成すると、[keep,keyboard]→[キーボード,を,飼う]という誤った規則が生成されてしまう。フェーズ1ではこの問題に対処する。以下、フェーズ1のアルゴリズムを述べる。

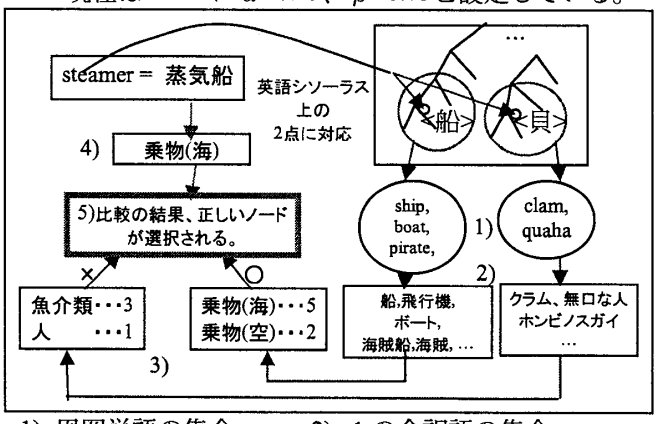
n個の意味を持つ原言語単語W_Sは、原言語シソーラス上でn箇所のノードにマップされる。その中から規則中でW_Sに与えられた一訳語W_Tの意味に対応したノードを特定する。W_Sの各意味につき以下のステップを計算する。

1) 対応する原言語シソーラス上のノードN_Sについて、その周囲(親・兄弟・子供)の原言語単語の集合S_Sを求める。W_S自身は含めない。
2) S_S中の全単語の訳語による集合S_Tを求める。
3) S_T中の全単語について、目的言語シソーラスで属するノードを求め、それらノードの集合Aを求める。
4) 目的言語シソーラスで訳語W_Tが属するノードの集合B(W_Tが多義語ならBは複数要素を持つ)を求める。
5) A中のノードNの頻度をfreq(N)、頻度最大のノードをNmax、S_Tに含まれる訳語数をmとしたとき、以下の式によりAとBの共通度Scoreを求める。最もScoreの高いノードN_Sが求める意味である。

$$Score = \frac{1}{m} freq(N\max)^{\alpha} \cdot (bonus + \sum_{N \subset A, N \subset B} freq(N))$$

$$bonus = \beta (if\ N\max \subset A) \quad or \quad 0 (otherwise)$$

現在はad hocにα=0.25、β=0.75と設定している。



1) 周囲単語の集合　2) 1の全訳語の集合
3) 2の属するノードの集合
4) 訳語の属するノード　5) 3，4の比較で意味特定

図2 意味特定の具体例

英日翻訳における具体例を図2で示す。図2中の番号は上記アルゴリズムのステップ番号に対応している。英単語steamerは<船>に類する意味と<貝>に類する意味を持ち、即ち英語シソーラス上で２ノードにマップされる。ここで訳語として“蒸気船“が与えられた場合、対応する<船>の意味のノードが特定される。

3　実験・考察

実験では、英日翻訳を扱い、原言語シソーラスとしてWordNet[3]、目的言語シソーラスとしてEDR概念辞書[4]、大規模コーパスとして8GBのWebコーパスを用いた。50000の既存規則に本手法を適用した結果、18392の類似規則が生成された。

まず、類似規則の有無による翻訳差分を評価し、有効な類似規則の割合を調べた。評価には1342の類似規則を用い、英文200万文に対する翻訳差分を取り、翻訳差分を生じた308規則を、有効、大差無、悪影響有の３値で評価した (表1)。結果、78.5%の類似規則が翻訳改善に有効であった。これは、同様の規則を人手で作成する場合と比較して、より効率的に規則を作成することができる精度である。以下に改善例を示す。

元規則：　[(人),have,frame]→[(人),が,体格,を,する]
生成規則：[(人),have,body]→[(人),が,体,を,する]
翻訳差分：He has a pliant body.
　前：彼は柔軟なボディを持っています。
　後：彼は柔軟な体をしています。

また、悪影響を与える規則を分析した。最大の原因は、元となる規則より、規則を適用するための条件を狭めねばならない類似規則が存在することであった。

[new,road]→[新道] ⇒ [new,way]→[新道]
× new way of running → 「走ることの新道」

この問題に対処するには、既存の規則(way of ～ing)との優先度を適切に調整する必要がある。

評価規則数	1342(／18392)
翻訳差分を生じた規則数	308(／1342)
翻訳結果up	242(／308), 78.5%
翻訳結果even	13(／308), 4.2%
翻訳結果down	53(／308), 17.2%

表1　翻訳差分の評価結果

次に、フェーズ1の意味特定の効果を評価した。50000の既存規則から類似規則を生成する際に、意味特定は12766回行われ、第一語義以外が選択された回数は5441回であった。そのうちの60種の語義選択に関し、決定的に第一語義を選択する場合と比較して正しいか、同等か、誤りか、の３値に分類した(表2)。結果、第一語義の選択時 ((25+8)/60=55%)に比べ、フェーズ1実行時((25+27)/60=83%)に、より正確に意味を特定できることが分かった。正しい意味特定による類義語の変化を表3に示す。また、誤った意味を選択する規則を分析したところ、characterに対してキャラクターという訳語が付与されている等、訳語から意味の曖昧性が解消されないというケースが最も多かった。この問題に対しては、各意味について一旦類似規則を生成し、その結果を意味特定にフィードバックするといった改良を行う必要がある。

語義選択回数	12766回
第一語義以外を選択	5441回
評価対象	60種の選択
正しい選択	27(／60)
同等の選択	25(／60)
誤った選択	8(／60)

表2 意味特定の評価結果

英単語	付与訳語	類義語(意味選択無)	類義語(意味選択有)
medicine	薬	surgery,neurology	drug,intoxicant
mine	機雷	pit,pool,well	bomb,firework
address	住所	code,ASCII,software	street address
attraction	魅力	reaction,stress,pull	appearance,excellence
ring	指輪	noisiness,voice	jem,band,necklace
field	分野	diamond,centerfield	domain,subject,realm
removal	免職	separation,tear	firing,discharge,sack

表3 意味特定による類義語の変化

4　まとめ

既存の語彙翻訳規則から、シソーラス・大規模コーパスを用いて類似語彙規則を生成する一手法を提案した。本手法では、原言語・目的言語の両シソーラスを用いて、予め規則中の語彙の意味を特定することで、誤った意味認定に基づく類似規則生成を防ぐ。英日翻訳に関し、308の生成規則を翻訳差分により評価したところ、78.5%について翻訳結果が改善した。今後は、従来法との効率的な組み合わせを検討し、精度改善を目指す。

参考文献

[1] 荒牧, 黒橋, 佐藤, 渡辺:用例ベース翻訳のためのパラレルコーパスからの対訳対発見, 研究報告「自然言語処理」No.144 – 004 (2001)
[2] シソーラスとMDL原理を用いた格フレームの一般化, 李航, 安部直樹, 自然言語処理における学習シンポジウム予稿, pp.1-8, (1994)
[3] Cognitive Science Laboratory, Princeton University WordNet, http://www.cogsci.princeton.edu/~wn/
[4] 日本電子化辞書研究所, EDR概念体系辞書, http://www.iijnet.or.jp/edr/