分野判定トリガー情報のフィードバックによるWeb翻訳

5Y−5

羽鳥洋美　神山淑朗
日本アイ・ビー・エム株式会社

1. はじめに

機械翻訳において、翻訳時に自動的に適切な分野辞書を選択するということは、訳質を向上させる上での重要な技術のひとつになっている。パターンベース翻訳システムPalmTree[1,2,3]では、辞書の自動切り換え機能[4]を実装し、また、これに最適な構成を持つ翻訳辞書[5]を構築した。しかし、これは一文を対象とした局所的な分野判定であるために、文章全体の中には適切な分野辞書が使用されずに翻訳されてしまう文も存在する。本論文では、翻訳対象をWebコンテンツに限定し、Webコンテンツの特徴を利用して、さらに効果的に適切な分野辞書を選択する方法を報告する。

2. 辞書の自動切り換えの問題点

辞書の自動切り換えとは、翻訳時にある分野辞書のトリガーとなる複合語が使用されると、その分野の話題であると判定し、その分野辞書の優先順位を上げるという方法である。これにより、どのような翻訳対象についても、話題の転換に柔軟に対応したより精度の高い翻訳結果が得られるようになった。しかし、トリガーとなる複合語が使用される以前の文章に対しては、適切な分野辞書を使用することができない。例えば、Javaという単語の訳語が、基本辞書には「ジャワ島」、コンピュータ辞書には「ジャバ」、さらに「programming language」がコンピュータ辞書のトリガーとなる複合語として登録されている状態で以下の文章を翻訳すると、第一文の「Java」に誤訳が生じる。

(原文) What is Java? Java is a programming language.

(訳文) ジャワ島は何ですか？ジャバはプログラミング言語です。

さらに、適切な分野辞書を選択した後、10文翻訳する間にその分野辞書のトリガーとなる複合語が使用されない場合には、優先順位を元に戻す仕組みなので、それ以降の文章に対して適切な分野辞書を使用することができないという問題点がある。

3. Webコンテンツ

3.1 Webコンテンツの特徴

翻訳対象をWebコンテンツに限定し、その特徴を考慮すると、翻訳時に得られる分野判定情報をフィードバックすることにより、辞書の自動切り換えの問題点が解決され、より精度の高い翻訳結果が得られることが期待できる。ここで考えたWebコンテンツの特徴とは、以下の2点である。

(1) URLが同じであれば、前回アクセスしたときと同一の話題か、更新されていたとしても同一分野の話題である可能性が高い。

(2) 一つのページは一つの分野の話題である場合が多い。

(1)の特徴から、翻訳時に判定された分野名をURLとともに記録すれば、次回このURLを翻訳するときに、最初から適切な分野辞書を使用することができると考えられる。また(2)の特徴から、特定の分野辞書を使用して、ページの最初から最後までを翻訳してもよいと考えられる。

3.2 Webコンテンツの分析

さらに現実のWebコンテンツを調査したところ、以下の4つのタイプに分類できることが分かった。

(i)**分野なし型**：特定の分野の話題がないもの。（例）検索サイト等

(ii)**多重分野型**：複数の分野の話題が同一ページ内に混在しているもの。（例）ニュースのヘッドライン等

(iii)**分野変化型**：同じURLであっても、コンテンツの更新とともに話題の分野までも変わってしまうもの。（例）ニュースの記事等

(iv)**特定分野型**：特定の分野の話題であるもの。（例）コンピュータ情報のサイト、スポーツニュース等

これらのうち(iv)については、翻訳時に得られる分野判定情報をフィードバックする本手法が有効であると考えられる。しかし、(i)〜(iii)については、従来の辞書の自動切り換えを適用する方がよい。

Web translation by feeding back information for judging category
Hiromi Hatori, Yoshiroh Kamiyama
IBM Japan, Ltd. (e-mail: hsaitoh@yamato.ibm.com)
1623-14, Shimo-tsuruma, Yamato-shi, Kanagawa-ken

4. 分野判定情報のフィードバック

翻訳時に得られる分野判定情報を、先に述べた「(iv)特定分野型」だけにフィードバックするため、以下の手順を考え、実装した。

(1) Webコンテンツの翻訳時に以下の情報を収集する。
A. 翻訳した文の総数
B. 分野辞書ごとの使用されたトリガーとなる複合語の数

(2) (1)で収集した情報から、分野辞書ごとに相対頻度(B/A)を求め、その値とある閾値をもとに、Webコンテンツの分野を推定する。

(3) (1)(2)の結果をURLをキーとしてファイルに保存する。ただし、同じURLに対して既存のデータがあれば上書きせずに履歴を残す。(最大N回まで)

(4) 次回の翻訳時には、与えられたURLの推定分野が存在するかどうかを調べ、存在すればその分野の辞書を最優先で使用して翻訳を開始する。翻訳が始まったら上記(1)(2)(3)の処理を行う。

手順(2)では、ある分野辞書のトリガーとなる複合語が最も多く使用されたとしても、相対頻度がある閾値を超えなければ分野不明と推定される。これにより、「(i)分野なし型」のコンテンツには、分野判定情報はフィードバックされない。また、複数の分野辞書の相対頻度が閾値を超える場合にも、分野不明と推定される。これにより、「(ii)多重分野型」のコンテンツにも、分野判定情報はフィードバックされない。さらに、「(iii)分野変化型」のコンテンツでは、手順(3)において話題が変化する前後の情報が相殺されてしまうため、結果的に分野判定情報はフィードバックされない。このようにして、最終的に「(iv) 特定分野型」だけに分野判定情報がフィードバックされることになる。さらに、手順(1)で収集する情報は、翻訳の副産物として得られるものであるため、この手法を用いることにより、翻訳速度を犠牲にすることなく、より精度の高い翻訳結果を得ることができる。

5. 評価

代表的なWebサイトから20ページずつ選んで本手法の評価を行った。【表1】は各サイトのページがどの分野に分類されたかを示す。上段の数値はその列の分野に推定されたページ数、下段は使用されたトリガーとなる複合語の数の合計を表す。

【表1】より「(i) 分野なし型」に属するyahooのような検索サイトや「(ii)多重分野型」または「(iii)分野変化型」に属するcnnのようなニュースサイトは、期待どおりに分野不明と推定されている。一方、「(iv)特定分野型」に属するサイトであるmicrosoftはコンピュータ分野、golftodayはスポーツ分野と、やはり期待どおりの分野に推定されている。この結果から、先に述べた手順を用いて分野判定情報をフィードバックする本手法により、翻訳時に適切な分野辞書が選択できるようになるといえる。

【表1】：推定された分野と複合語のヒット件数

	art	com	ent	hom	pae	sci	spo	?
Yahoo	1	0	1	0	0	0	0	18
	8	51	20	0	14	0	1	-
Cnn	0	0	0	0	5	0	2	13
	3	84	57	3	191	0	199	-
Microsoft	0	20	0	0	0	0	0	0
	0	1264	0	61	52	0	4	-
golftoday	0	0	0	0	0	0	19	1
	7	58	77	0	4	0	1734	-

art：アート、com：コンピュータ、ent：エンターテイメント、hom：家庭、pae：政治経済、sci：科学、spo：スポーツ、?：分野不明

6. 考察

本手法により蓄えられた分野判定情報を調べると、あるサイトがどのような分野のコンテンツを持っているかが分かる。例えば、あるサイトがコンピュータ分野に属するコンテンツばかりを持っているということが分かれば、そのサイト内の別のページに対してもコンピュータ辞書が適切であると考えることができる。このように、蓄えられた分野判定情報は、ページ単位だけでなく、サイト単位の翻訳にも適用することができる。また、これを分野判定機能を持たない翻訳システムへフィードバックすることも可能であろう。さらに、より多くの分野判定情報を蓄えられるという点からは、サーバー翻訳システムへの適用がより効果的であると考えられる。今後の課題としては、翻訳速度を考慮しつつ、翻訳開始前に情報を収集して適切な分野辞書を選択する「2パス翻訳」を併用して、より精度の高い翻訳結果を得られる手法を検討したい。

参考文献

[1] Takeda, K., "Pattern-Based Context-Free Grammar for Machine Translation," Proc. of 34th ACL, pp. 144-151, 1996
[2] Takeda, K., "Pattern-Based Machine Translation," Proc. of 16th Coling, Vol. 2, pp. 1155-1158, 1996
[3] 渡辺, 武田, "パターンベース翻訳システム：PalmTree", 情報処理学会第55回全国大会, 1997
[4] 宮平, 神山, 羽鳥, "パターンベース翻訳システム PalmTree の訳語選択", 情報処理学会第59回全国大会, 1999
[5] 羽鳥, 宮平, "辞書の自動切り換え機能を考慮した翻訳辞書", 情報処理学会第61回全国大会, 2000