機械翻訳システムにおける確率的品詞推定とその応用

6Y−6

神山 淑朗
日本アイ・ビー・エム株式会社

1. はじめに

近年、自然言語処理の世界では、計算機の能力の飛躍的な向上を背景として膨大な言語データ(コーパス)から言語モデルを構築して応用する、確率・統計的なアプローチが盛んである。一方、我々の開発している「インターネット翻訳の王様」の英日翻訳エンジンであるPalmTree[1]は、人手による翻訳パターンの入力によってチューニングを行う方式の翻訳システムである。PalmTree は、辞書の自動切り換え機能[2,3]や実行時動的パターン生成[4]などの手法によって、翻訳精度は向上しつつあるが、いくつかの点では従来の手法では限界があることがわかっている。

今回、それらの問題に対処することを目的として、既存の翻訳エンジンに対して確率的品詞推定の手法を適用する試みを行った。また、この手法によって得られる情報は、さまざまな自然言語処理における基礎的な情報であり、翻訳以外にもいくつかの応用例を考案し、実装した。本論文では、今回実装した確率的品詞推定の手法、それを既存の翻訳エンジンへ適用した結果とその考察、およびその他の応用例について報告する。

2. 翻訳エンジンの問題点

PalmTree はその解析方式と人手によるデータ入力という性質上、次のような問題があった。

2.1. 単語のコストの問題

PalmTree の形態素辞書には、すべての単語に対して品詞ごとにコストと呼ばれる頻度情報が整数値で与えられている。しかし、コストの数値をいくつにすればよいかは勘に頼らざるを得ない上、すべての単語についてのコストのメンテナンスを行うのは多くの手間がかかる。

2.2. 長文の問題

文が長くなると品詞や構文のあいまい性が累積し、組み合わせの数が爆発的に増えるために翻訳速度・翻訳精度がともに著しく低下してしまう。

3. 確率的品詞推定の実装

前述の問題点はいずれも言語が本来持つ単語列の規則性に関する知識の不足に起因すると言える。そのため人手によって規則性の知識を補う必要があったり、単語列に規則性からくる制約を課すことができずに組み合わせを爆発させてしまうのである。そこでコーパスから得られる知識を利用して入力単語列の品詞を推定する確率的品詞推定を実装してそれらの問題に対処する試みを行った。

品詞推定(品詞タグ付け)は様々な手法が存在するが、今回は隠れマルコフモデル(HMM)を用いた[5]。品詞推定を既存の翻訳システムで利用するにあたり、コーパスとの品詞体系の違いが問題となった。コーパスの品詞体系は統計処理に適した論理的な体系となっているが、PalmTree の品詞体系は自身の翻訳メカニズムに特化したデザインになっているためである。例えば、PalmTree では"to"の品詞は常に前置詞である。前置詞は本来後ろに名詞(句)が続くものであるが、"to"は不定詞を導くので後ろに動詞が頻繁に出現する。そのことが「前置詞の後ろには名詞(句)が現れる」という確率を下げてしまい、"to"以外の前置詞句の推定にも悪影響を与えてしまう。コーパスの品詞体系では不定詞を導く"to"には単独の品詞が与えられており、このような問題は生じない。そこで、翻訳エンジン側に近い論理的な品詞体系を新たに定義してその上で計算を行い、最後に出力を翻訳エンジンの体系に合わせるという手法をとった。

4. 確率的品詞推定の適用

翻訳エンジンの形態素解析の後工程として確率的品詞推定を適用してみた。すなわち、従来通りの形態素解析を行った後に品詞を推定してコストの調整や品詞の枝刈りを行うフェーズを設けた。それにより、既存の翻訳エンジンの枠組みを大きく変えることなしに以下のようなことが可能となった。

4.1. 単語コストの動的割り当て

「"time"の品詞は一般的に動詞より名詞で使われることが多い」といった漠然とした根拠で静的に与えられたコストを利用するのではなく、コーパスから得られた統計情報を根拠とし、「この文脈の"time"は名詞の確率が最も高い」といった情報を利用してコストを動的に割り振ってから構文解析へ渡すことができるようになった。

4.2. 長文の計算量の低減

短文ならばすべての可能性を調べて最適な解を選択すればよいが、長文の場合は翻訳時間が長くなりすぎて解析が終わる前にタイムアウトを起こしてしまう。したがって、なるべく早い段階であいまい性を排除することが重要であるが、品詞推定により形態素解析の段階で品詞の枝刈り(可能性の低い品詞候補の除去)を行うことができる。これによって、あり得ないと思われる可能性を排除し、計算量を減らすことができるようになった。

Applying statistical part-of-speech estimation to MT system.
Yoshiroh KAMIYAMA
Software Development Lab, IBM Japan, Ltd.
E-Mail: kami@jp.ibm.com

5. 実験と評価

試行錯誤の結果、過度の枝刈りを行うと、品詞推定を誤った場合に誤訳を招いてしまう上、仮に妥当な品詞列が選択されていたとしても翻訳精度が悪化する場合があることがわかった。これは、与えられた品詞列でその文を構文解析するための文法規則が足りないケースが出てくるためである。この場合、一つの構文木を構築することができず、誤った品詞を使ってでも構文解析に成功した場合よりもむしろ悪い結果となる。

そこで今回の実験では、過度の枝刈りによる解析失敗を避けるためにあいまい性を残した保守的な枝刈りを行った。cnn.com から長文 60 文を抜き出して翻訳したところ、品詞の枝刈りを行うと 15%程度翻訳速度が向上した。また、翻訳結果は改善が 13%、改悪が 7%、変化なしが 70%、変化ありだが訳質は同等が 10%であった。かなりあいまい性を残しているので、訳質は変化なしが多いが、翻訳速度が向上すればそれでもメリットは十分あると言える。

6. 確率的品詞推定のその他の応用

6.1. 対訳テキストの自動文対応付け

インターネットの普及に伴い、電子化された対訳テキストを利用できる機会が多くなっている。対訳テキストから抽出される知識は機械翻訳システムの精度向上に利用可能である。そのための第一歩として、対訳テキストを文に分割し、文レベルでの対応付けをすることが必要となる。そこで、確率的品詞推定を使って英日対訳テキストの自動文対応付けプログラムを実装した。以下に文対応付けのアルゴリズムを示す。

(1) 英語側、日本語側ともに文に分割する。
(2) 一方の言語の1文は他方の3文以内に対応するという制約の下、位置的に対応可能なペアを構成する。
(3) 対応可能なペアから訳語対を抽出する。ここで英語側を形態素解析して品詞推定を行い、推定された品詞の訳語と比較することによって訳語対を探す。
(4) 訳語対の数や内容からスコアを算出する。
(5) スコアが閾値以上のペアを確定し、閾値を下げて(2)へ戻る。

6.2. 文脈に応じた辞書引き

アプリケーションへの応用として、マウスポインタを単語の上に移動するだけで辞書引きができる機能への適用を行った。品詞推定を活用することにより、文脈に応じた品詞の訳語を表示することが可能となった。(図.1)

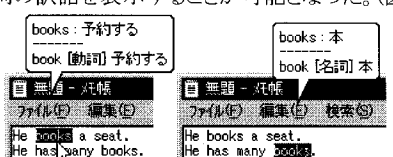

図.1 文脈に応じた辞書引きの例

6.3. 自動訳振り

英語がある程度は自力で読めるが、ときどき知らない単語に遭遇するといったレベルのユーザーを対象に、ブラウザ上で難しい単語だけに一括して訳を振る「自動訳振り」という機能を今回新たに考案した。ここでも品詞推定を活用することにより、文脈に応じた品詞の第一訳語を使って訳振りを行うことができる。(図.2)

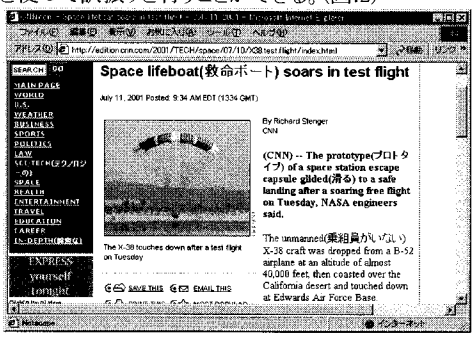

図.2 品詞推定を活用した訳振りの例

7. 今後の課題

応用範囲の拡大や適用方法の工夫の余地は多い。例えば品詞推定により次のような文法要素の認識ができれば訳質向上に有効である。(1) 文分割可能な場所(例えば副詞節の前など)。(2) 構文解析を乱す特定の副詞。(3) 挿入句の範囲。(4) 省略されている接続詞(that 等)や関係代名詞。

また、文の長さに応じた品詞の枝刈りによって速度と訳質の最良のバランスを探ることも必要であろう。

8. まとめ

既存の翻訳エンジンに対し、確率的なアプローチを導入する試みを行った。翻訳エンジンの品詞体系に適合するような確率的品詞推定のロジックの実装を行い、効果を検証した。また、訳質以外への応用を考案し、実装も行った。確率的品詞推定により得られる情報は、訳質やさまざまなアプリケーションにおいて有効であることが確認できたので、更に応用範囲を広げていきたい。

参考文献

[1] Takeda, K., Pattern-Based Context-Free Grammar for Machine Translation, Proc. of 34th ACL, pp. 144-151, 1996
[2] 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
[3] 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
[4] 宮平, パターンベース翻訳システム PalmTree の構文解析, 情報処理学会第 61 回全国大会, 2000
[5] D. Cutting, J. Kupiec, J.Pedersen, and P.Sibun: A Practical Part-of-Speech Tagger. In Proceedings of the Third Conference on Applied Natural Language Processing, Trento, Italy (April 1992)