多言語を対象とした学習型機械翻訳システムの開発

越前谷 博［北海学園大学／助教］

背景・目的

近年、インターネットの急速な普及に伴い、母国語だけではなく世界中の様々な言語で表現された情報に迅速かつ容易に触れる機会が増大している。その際に、大きな問題となるのが言語の壁である。この問題を解決するために、機械翻訳システムに対するニーズが強まっている。日本語－英語間の機械翻訳システムにおいてはこれまでに最も多く研究されており、実用化が進んでいる。しかし、インターネット上の情報資源は日本語、英語だけにとどまらず、様々な言語を用いて配信されている。そのため、特定の言語間の機械翻訳システムだけではなく、多言語機械翻訳システムが大きな注目を集めている。

しかし、完全自動の多言語機械翻訳システムの開発は容易ではない。特に、翻訳のための言語資源である、辞書、文法規則、そして、対訳コーパスの乏しい言語においては、コストの面からも完全自動の多言語機械翻訳システムを実現することは容易ではない。そこで、本研究では、最初から完全自動の機械翻訳システムを目指すのではなく、不完全ではあるが徐々に学習により完成度の高い翻訳システムへと成長する、学習型機械翻訳システムの開発を目指す。学習が不十分な場合、ユーザは完全な訳文を得ることはできないが、システムは完全訳に最も近い訳文、即ち、類似文をユーザに提供することで、翻訳支援を行うことができる。また、本システムは対訳文のみから翻訳規則を効率よく学習することが可能であるため、対訳文を与えるだけで、完全自動の機械翻訳システムに近づくことが可能である。更に、対訳文として使用される言語を変更することで、様々な言語に対処することも可能となる。

内容・方法

本研究では、原言語文と目的言語文の組からなる対訳文より翻訳規則を自動獲得する、学習型機械翻訳システムを実現する。本システムは学習機能により翻訳規則を獲得できるため、あらかじめ対訳辞書や文法規則を人手で与える必要がない。本研究における翻訳規則の学習とは、対訳文のある部分を一般化することで対訳文中に内在する規則を獲得することである。このような学習型機械翻訳システムにおいて問題となるのは、対訳文の集合体からなる対訳コーパスより、いかに効率よく翻訳知識を獲得するかである。言語資源の乏しい言語においては大規模な対訳コーパスを収集することは容易ではないためこの問題の解決は非常に重要となる。

この問題に対して、本研究では、対訳文中の省略可能性に着目することで翻訳規則を効率よく獲得する。例えば、対訳文中に(a quiet room；静かな部屋)という句が存在し、また、他の対訳文中に(a room；部屋)という句が存在していた場合、英文においては“a”と“room”の間の部分を省略可能と考え、また、日本文においては、“部屋”の左側の部分を省略可能と考える。その結果、(a@room；@部屋)という一般化が可能となり、他の対訳文中にこのような句が存在する場合、“a”から“room”を一般化し、“部屋”の左側の部分を“部屋”も含めて一般化することで、より汎用的な翻訳規則を獲得できる。即ち、本研究では、対訳文中の句に相当する部分を一般化することにより、汎用的な翻訳規則を効率よく自動獲得することが可能になると考えられる。

結果・成果

英日の対訳文70文を学習データ、30文を評価データとして、翻訳評価を行ったところ、省略可能性に基づく翻訳規則の自動獲得は大きな翻訳精度の向上をもたらすには至らなかった[1]。そこで、自動獲得した(a @ room；@部屋)のような省略ルール間で更なる一般化を行うことで、より汎用的な抽出ルールを自動獲得する。例えば省略ルールとして(a @ hotel；@ホテル)が獲得されると、(a @ room；@部屋)との間の差異部分を更に一般化することで抽出ルールとして(a@；@)を自動獲得する。この抽出ルールは対訳文中の英文においては“a”から右側に存在する部分、日本文においては英文のその部分と対応関係にある部分をそれぞれ抽出してもよいとする、抽出のための範囲情報を有する対訳知識となる。この抽出ルールを第1抽出ルールとして対訳文に適用することで、より多くの対訳文の一般化を実現できる。更に、この第1抽出ルールが適用された対訳文を用いて、(in@；@に)といった第2抽出ルールの自動獲得も行う。この第2抽出ルールは、対訳文中の英文においては“in”の右側に存在する部分を、日本語文においては“に”の左側に存在する部分をそれぞれ抽出してもよいとする、範囲情報を有する。これらの第1抽出ルールと第2抽出ルールを自動獲得し、適用することで、より多くの翻訳規則の獲得が可能となった。

性能評価実験は、英日の対訳文948文を学習データ、762文を評価データとして用いた。更に、評価データを用いて得られた翻訳結果に対しては、人手による主観評価だけでなく、人手で作成した正解文である参照訳との類似度を計算することによりコンピュータが自動的に評価を行う、自動評価の両方で行った。通常、自動評価に使用する参照訳は1つのみでは不十分であるため、バイリンガルに参照訳の生成を委託し、1文に対し4つの参照訳を用いた。自動評価手法としては、BLEU と IMPACT を用いた。その結果、省略可能性を用いない機械翻訳システムと省略可能性を用いた機械翻訳システムとでは、省略可能性を用いた機械翻訳システムの方が、約4ポイントから6ポイント高い翻訳精度を示した。したがって、本研究で提案した、省略可能性に着目した翻

訳規則の自動獲得手法に基づく学習型機械翻訳システムの有効性が示された[2]。

[1] 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に基づく翻訳ルールの自動獲得手法、平成19年度電気･情報関係学会北海道支部連合大会講演論文集、pp.189–190, 2007年10月.
[2] 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、NL–183, pp.127–134, 2008年1月.

今後の展望

現在、対訳コーパスを用いた機械翻訳の研究が盛んに行われている。特に、統計ベースの機械翻訳手法はその代表的なものである。しかし、それらは大規模な対訳コーパスの使用が前提であるため、対訳コーパスの効率的な利用という点においては、疑問が残る、そこで、今後は、統計ベースの機械翻訳システムとの比較実験を行うことで、対訳コーパスの効率的な利用という観点での検証を行う。

また、今回は構文構造の大きく異なる英語－日本語間の対訳コーパスを対象に性能評価を行ったが、今後は他の様々な言語間の対訳コーパスを用いた性能評価実験を行うことにより、本研究の言語非依存性についても検証していく予定である。