博 士（情報科学） 彭 静

学 位 論 文 題 名

ACQUISITION OF TRANSLATION KNOWLEDGE FROM THE WEB

（ウェブからの翻訳知識の獲得）

学位論文内容の要旨

Although machine translation (MT) has been developed for many decades, none of the current MT systems seems to have improved the translation quality that satisfies human's requirements. The availability of large corpora and developed automatic corpus-based methods makes it possible to reduce one bottleneck of MT, that is, the lack of knowledge. The recent rapid development of the Web makes it an extremely large and valuable data source. In this thesis, we are mainly focusing on the two problems that are, whether the Web is a useful corpus and how to acquire translation knowledge from the Web.

Our work evolved from a pilot study of "Web as corpus." In the study, we investigated several questions. They are "Is the Web a corpus?", "How can we obtain N-gram counts from the Web?", "How is the Web N-gram coverage?", "How much is the indexed size of the Web?" and "Is the Web data perfectly normalized?" We knew some certain advantages of the Web. It is much larger and has better N-gram coverage than normal corpora. The content of the Web is constantly changing. Simple Web counts of N-gram have been proved to be useful for MT tasks. However, there are also several disadvantages. There is lots of noise data on the Web and the current search engines are reported to be not always stable and accurate. The Web data is also proved to be not perfectly normalized. Therefore, the Web estimates are obviously useful, but should be used with caution.

In this thesis, we combined the Web with several traditional approaches to resolve three MT routine tasks. The first is detection of the countability of English compound nouns. Knowledge of countability of nouns is very important in MT. Although, many approaches have been proposed for learning the countability of individual nouns, until recently, little attention has been given to the countability detection of compound nouns. The number of compound nouns is so large that it is impossible to collect all of them in one dictionary. Especially relatively new words have not yet reached any current dictionary. Thus using the Web-scale and constantly updating data is proved be a viable alternative to avoid the sparseness problem from normal corpora. We classified compound nouns into three classes: countable, uncountable, plural only. To detect which class a compound noun is, we proposed some simple, viable N-gram models whose parameters' values (Web counts) can be obtained with the help of Web search engine Google. Such Web-based models as filters were proved to be useful for improving the performance of the limited general rules of countability detection.

式の特性を改善する手法を提案している. これは ACK から得られる過去のチャネル推定値から, 実際の送信時刻におけるチャネルを予測することにより, 送信ベクトルの決定とリソース制御を行うものである. 本論文では, チャネル予測のために 3 通りの方式を提案している. 第 1 の方式は,ACK パケットを用いて推定されたチャネルを直線により外挿を行うものである. 第 2 の方式は,2 次関数を適用して予測するものである. 第 3 の方式は指数関数を用いて予測するものである. これら 3 通りの方式の特性を評価するため,HIPERLAN/2 を想定した通信システムについて計算機シミュレーションを行った. その結果, チャネルの予測を行わない方式の場合には, 最大ドップラー周波数が高くなるにつれて著しく特性が劣化するが, 提案したチャネル外挿方式を用いると誤り率特性が改善されることが確認された. 特に,2 次関数, あるいは, 指数関数を用いたときには, 高い最大ドップラー周波数においても劣化が少ないことが示された。

一方, 広帯域伝送系においては, 符号間干渉が発生することになる. この問題は OFDM(Orthogonal Frequency Division Multiplexing) 方式を用いることによって回避できる. 第 4 章では,OFDM 方式を用いた広帯域系への MIMO E-SDM 方式について論じている. 時変動フェージング環境に対応するため, 時間領域のチャネル (インパルス応答) を予測し, それらのフーリエ変換により各サブキャリアのチャネルを求める方式を提案している. このことを計算機シミュレーションを用いて評価を行い, 第 3 章と同じようにチャネル予測の有効性, 特に,2 次関数と指数関数の予測が優れていることを述べている. また, 通信系を簡易化するために全てのサブキャリアに共通の変調方式を用いても特性の劣化は少ないことが示されている.

第 5 章では, 提案方式の実験的検証のために行った伝搬実験の概要を述べている. 前章までの評価は Jakes モデルを仮定した計算機シミュレーションに基づいて行ったものであるが, 実際の伝搬路においては必ずしもこのモデルは成立しないため, 実環境での評価が必須となる. ここでは, 室内における MIMO チャネル測定法, および, 時変動伝搬環境におけるチャネルの自己相関とドップラースペクトルを明らかにしている.

上述の測定系により得られた時変動 MIMO チャネルデータを用いて, 第 6 章では狭帯域 MIMO E-SDM 方式におけるチャネル予測方式の評価を行った. その結果, 第 3 章で提案した方式が実伝搬路においても特性改善に有効であることが示された.

第 7 章では, 実測伝搬データを用いて MIMO-OFDM E-SDM 系におけるチャネル予測, および, 全てのサブキャリアに共通の変調方式を用いる簡易方式の評価を行った. 本論文で提案した方式は広帯域の実伝搬環境においても有効であることが明らかになった.

第 8 章は結論であり, 本論文の内容と得られた成果を要約している.

これを要するに, 著者は, 周波数利用効率改善が期待される MIMO E-SDM システムの時変動チャネル環境における特性評価とその改善法について重要な新知見を得たものであり, 無線通信工学に貢献するところ大なるものがある. よって著者は北海道大学博士 (工学) の学位を授与される資格あるものと認める.

Web counts have better N-gram coverage, but can be expected to contain noise introduced by a number of sources. On the other hand, corpus counts are much less noisy, but sparser than Web counts. Therefore, an interpolation scheme of combining the Web and a normal corpus estimate was proposed. This interpolation model was employed in the resolution of the second task addressed in our work that is correction of article errors in MT. The article selection is to decide when to use a (an), the, or zero article at the beginning of a noun phrase (NP). It is a complex problem in translation result generation in MT when source text is written in some languages, such as Chinese and Japanese, which do not have any articles or mark the countability. In our work, we considered articles and their headwords together and put them into 5 forms because determining an article is largely depended on the singular/plural form of its headword in this phrase. We assume that the article form with largest occurrence probability is most likely correct given a certain context. The occurrence probabilities can be obtained using the interpolation model combining the estimates of the Web and a corpus (BNC in our work). We evaluated the performance of using the pure Web estimates, pure corpus estimates and the interpolation model when given 4 different contexts. The interpolation model experimentally showed the best performance when appropriate interpolation parameters were chosen. We achieved a promising result on correction of article errors with much less parameters than those used in the previous research.

Beside the two approaches of combining the Web with rules and a corpus, we also investigated another possible scheme that incorporated the Web as a knowledge source into machine learning framework. The third task addressed in our work is the resolution of zero-anaphora (ZA) in Chinese text. In many natural languages, grammatical components that can be understood contextually by a reader are frequently unexpressed for discourse fluency. This phenomenon is ZA. ZA resolution is very important in MT. Since target languages such as English that cannot be adequately generated with omitted expressions, the antecedent of the ZA in the source language must be identified and made explicit. A learning classifier based on maximum entropy (ME) was proposed to determine whether a candidate is the correct antecedent or not. In our original ME-based classifier, we employed 13 regular features motivated by previous research. From the classification error analysis, our approach was found to suffer from semantic problems, such as semantic ambiguity and the lack of semantic knowledge, and these problems cannot be resolved using any current semantic dictionary. Two innovative features were constructed for extracting additional semantic information from the Web. The additional semantic information includes semantic consistency and the semantic relations of predicates. The values of the two features can be obtained by querying the Web using some patterns. We retrained and tested the advanced classifier with the regular features and the two additional features. The two Web-based features significantly improved the performance of classification. Our study showed the Web as a knowledge source could be incorporated effectively into learning framework and significantly improved the performance of the learning approach.

Although, the approaches proposed in this thesis are only crude, they are the first attempt of acquiring various kinds of translation knowledge from the Web. And combining the Web as a knowledge source with supervised methods is supposed as a promising direction, which we should currently purse.

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　長谷川　美　紀

学位論文題名

ACQUISITION OF TRANSLATION KNOWLEDGE FROM THE WEB

（ウェブからの翻訳知識の獲得）

本研究は，機械翻訳の質を向上させるための Web 上の知識を用いた翻訳タスクの解決手法について提案したものである。

近年，機械翻訳の精度は徐々に向上しているが，ユーザが満足するようなレベルまで達しているとはいえない。著者は，この問題を解決するためにいくつかの機械翻訳結果を用いてその原因を考察し，文法，語彙，意味，世界知識などの知識が不足していることが主な原因であることを明らかにした。さらに関連研究を比較検討し，この知識不足問題がまだ十分に解決されていないことを明らかにした。その問題の一つは，有効な知識を獲得するためにはコーパスを極めて大量に収集することが必要であるということである。Web 上のホームページは，画像のみのページを除いては自然言語で記述される。Web 全体を見れば一つの巨大なコーパスを構成していると考えることができる。それらの Web データの多くは頻繁に更新されており，新しいデータも次々と出現している。著者は，以上のことから，Web を翻訳知識を獲得するための理想的なコーパスと考え，その理想的なコーパスを利用して機械翻訳の性能を向上させることを本研究の目的としている。

Web をコーパスとして利用する ("Web as corpus") 際に，問題となるのが Web 上のデータの信頼性と有効性である。それらの問題に関するいくつかの予備実験と結論について第二章で述べられている。第二章で著者は，5 つの問題 (Web がコーパスとして利用可能かどうか，Web からの N-gram 出現頻度の抽出，Web 上での N-gram のカバー率の検討，検索エンジンにより検索できる Web サイズの推定，Web データの信頼性と妥当性の検討) の解決を通し，"Web as corpus"の性能の評価を行った。予備実験の結果によると，Web は普通のコーパスより規模が大きく，カバー率も高く，単純な Web 上の単語出現頻度に基づいて訳語候補を選択するができるという利点を持っているが，一方雑音データが多いという欠点がある。 さらに Web 上の検索エンジンの問題で Web(主に Web 上の単語出現頻度) から推測された N-gram の出現頻度と共起頻度などの信頼性と妥当性が普通のコーパスより劣っているという欠点も無視することができないということが明らかとなった。

そこで，本研究においては他研究と違い，直接 Web 上の単語出現頻度を利用するという方法を取らず，以下のような 3 つの手法の提案を行った。

1. Web 上の単語出現頻度をフィルタとして利用し，従来の rule-based アプローチと融合する。このようなフィルタにより，最適な候補の選択のためのルールを利用する前に可能性が低い N-gram 候補を取り除くことができる。このようなフィルタは規則が足りない場合に役立つと考えられる。

2. Web 上の単語出現頻度とコーパス上の単語出現頻度とは内挿モデルを用いて統合する。Web と通常のコーパスはそれぞれ利点と欠点を有している。適切なパラメータを有する内挿モデルにより両方の欠点をある程度解消することができると考えられる。
3. Web から抽出した知識は機械学習で利用される。通常の訓練コーパスから通常の特徴パラメータではなかなか得られない知識は Web から抽出し, Web 上の特徴パラメータに変更することで機械学習に導入することができると考えられる。

これらの手法を用いた翻訳知識の獲得は, 多様な方法で伝統的なアプローチと統合することができ, 従来の Web 上の単語出現頻度のみを用いた手法に対して, 新規性を打ち出すことができたと考えられる。次に著者は, 提案された手法を 3 つの翻訳タスク (可算性の検出, 冠詞誤りの校正, ゼロ代名詞の復元) の解決手法に応用し, それらの手法の有効性の確認を行った。この 3 つの翻訳タスクは機械翻訳の分野で最もよく研究されているタスクである。これらの翻訳タスクは意味的及び構文的知識, 分析型及び生成型問題, 英語及び中国語の処理に及んでいる。Web 上の知識を利用することにより rule-based と統計的なアプローチのみを用いてこれらの翻訳タスクを解決している既存の手法に対して, 新規性を有するものと考えられる。

第三章では, 英語複合名詞の可算性問題を扱っている。英語複合名詞に対して N-gram モデルと可算性規則を用いることにより, 可算, 不可算, 複数形名詞を検出する手法を提案した。また,N-gram モデルではパラメータとしてすべての単語の出現頻度と共起頻度を Web 上の単語出現頻度として簡単に推測することができるということが述べられている。第四章では, 英語の冠詞誤りを扱っている。Web 上の単語出現頻度とコーパス上の単語出現頻度を統合的に用いることにより内挿モデルにより冠詞の誤りを校正する手法を提案した。また, 提案された手法が単純 Web モデルとコーパス (BNC) モデルより有効であることを検討した。第五章では, 中国文に出現するゼロ代名詞 (zero anaphora) の復元問題を扱っている。Web から抽出された意味的な一致と述語間の意味関係などの知識を追加の特徴パラメータとして導入した ME-based (maximum entropy) 分類器を構築し, 中国文のゼロ代名詞を復元する手法の提案を行った。さらに, その手法の応用としてゼロ代名詞の復元による機械翻訳結果の自動校正手法の開発を行った。

本研究では主に以下に示す成果が得られている。
1. Web 上のデータによる推測の信頼性と妥当性を検討した。Web を用いた手法と伝統的なアプローチを結合した 3 つの手法の提案を行った。
2. 本研究により世界で初めて英語の単名詞ではなく複合名詞の可算性問題を扱った。最新の用語やその使用法が常に反映されている Web 上のデータは data sparseness 問題を避けることができ, 可算性検出ルールの性能を向上させることができることを示した。
3. 英語の冠詞誤りの自動校正では, 提案された Web 上の単語出現頻度とコーパス上の単語出現頻度を統合した内挿モデルにより Web を用いたものと通常のコーパスを用いたものの両方の欠点をある程度解消できることを示した。
4. 中国文のゼロ代名詞の復元問題の解決には意味上の一致と述語間の意味関係など通常の訓練コーパスからではなかなか得られない知識を Web 上から抽出することにより, より高い性能を持つ ME-based 分類器を生成できることを示した。この手法を用いて機械翻訳結果の自動校正を行った結果, 質の高い機械翻訳システムを構築することができることを示した。

以上を要約すると, 著者は Web 上のテキストデータを有効に活用することにより機械翻訳の質を画期的に向上させる手法を提案し, その性能評価実験により提案手法の有効性の確認を行った。本研究を通じて, 国際社会における異言語コミュニケーション技術の確立に貢献するところ大なるものがある。よって, 著者は北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。