T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

題目(和文)	統計的機械翻訳における翻訳モデルの学習手法に関する研究（論文要旨）
Title(English)	
著者(和文)	上垣外英剛
Author(English)	Hidetaka Kamigaito
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10542号, 授与年月日:2017年3月26日, 学位の種別:課程博士, 審査員:高村 大也,新田 克己,中本 高道,奥村 学,長谷川 晶一
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10542号, Conferred date:2017/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

Powered by T2R2 (Tokyo Institute Research Repository)

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	知能システム科学　　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学生氏名： Student's Name	上垣外 英剛	指導教員（主）： Academic Advisor(main)	高村大也
		指導教員（副）： Academic Advisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

機械翻訳は、入力された文書を、目的の言語に自動で翻訳するという課題である。電子計算機が発明されて以降、機械翻訳に関して様々な研究が行われ、数多くの手法が開発されてきた。本研究ではその中でも統計的機械翻訳の枠組みに注目し、その構成要素の一つである、翻訳モデルに残された課題の解決を行う。本論文で扱う翻訳モデルに残された課題は次の二つである。
一つ目の課題は、既存の教師なし学習に基づく単語アライメントの学習手法では、明示的に機能語と内容語を区別していないために、言語学的に遠い言語対における機能語と内容語の対応が誤り易いという課題である。
二つ目の課題は、階層的フレーズベース翻訳において、翻訳モデルを格納するためのルールテーブルが、単語アライメントから網羅的に抽出することによって生成されるために、翻訳時に不要な対応を多く含むこととなり、ルールテーブルのサイズが増大してしまうという課題である。
本論文では上記の二つの課題に対して、人手によるアノテーションを必要としない、教師なし学習に基づく解決手法を提案する。
第1章の序論では、機械翻訳の歴史と、統計的機械翻訳における翻訳モデルの概要を説明し、上記の二つの課題への導入を行う。
第2章の関連研究では、提案手法の理解を助けるために、まず、統計的機械翻訳とその学習手法について、翻訳モデルを中心に説明する。次に、一つ目の提案手法と関連する、単語アライメントに関する研究の説明を行い、最後に、二つ目の提案手法と関連する、翻訳モデル削減手法についての説明を行う。
第3章の、頻度制約を考慮した制約付きEM に基づく教師なし単語アライメントでは、一つ目の課題を解決するために、事後確率正則化学習法の制約として、機能語と内容語を明示的に区別して扱うことが可能な制約を提案した。提案した制約は出現頻度に基づいて機能語と内容語を識別するため、対象とする言語に依存しない。評価の結果、提案した手法は、従来手法である対称化制約と比較し、日本語と英語の言語対において、単語アライメント精度と翻訳精度が向上する事を、それぞれ自動評価の尺度であるAERとBLEUを用いて確認した。
第4章の階層的バックオフ過程に基づくHiero文法の学習では、上記の二つ目の課題を解決するために、階層型Pitman-Yor過程に基づくバックオフを行うモデルを提案した。提案手法と従来のベイズ的な手法を比較するために、ドイツ語/フランス語/スペイン語/日本語から英語への翻訳評価を実施した結果、提案手法は同等のルールテーブルサイズの元で、BLEUにおいて同等かより高い翻訳精度を示した。また、提案手法と従来のヒューリスティックな手法との比較評価についてもドイツ語から英語への翻訳を行う実験を実施し、提案手法は10分の1以下のルールテーブルサイズで、BLEUにおいてより高い翻訳精度を示すことを確認した。さらに提案手法とヒューリスティックな手法に統計的な検定に基づくルールテーブル削減手法を適用したものとの比較評価についても、ドイツ語から英語への翻訳を対象に実施した。その結果、提案手法は、より少ないルールテーブルサイズで、同等の翻訳精度を示すことを確認した。
第5章の、結論と今後の課題では、第3章と第4章で提案した手法が、実際に上記の翻訳モデルに存在する二つの課題を解決できているかについて論じた。そして、提案手法に残された課題について、統計的機械翻訳とニューラルネットワークに基づく翻訳の両面からの関係性を論じることで、機械翻訳における提案手法の位置付けをより明確にした。
上記のように、本論文では、統計的機械翻訳における翻訳モデルに存在している二つの課題に対処するための、教師なし学習に基づく手法を提案し、それらの手法が実際に効果的である事を実験によって示し、またその結果について論じた。

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	Computational Intelligence and Systems Science	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineering)
学生氏名： Student's Name	Hidetaka Kamigaito	指導教員（主）： Academic Advisor(main)	Hiroya Takamura
		指導教員（副）： Academic Advisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

In this thesis, we propose methods to solve the problems of translation model in the statistical machine translation system. There are mainly two problems in learning the translation model.
The first problem is that in linguistically different language pairs, function words are difficult to align with each other with current unsupervised word alignment models. To solve the problem, this thesis proposes a constraint to discriminate function words and content words by using the frequency of each word in training corpus with the posterior regularization framework. Experimental results show that the proposed method achieved higher word alignment quality and translation quality on Japanese and English language pairs.
The second problem is that in hierarchical phrase-based machine translation, the size of the rule table tends to be excessively large because of its heuristic method for rule extraction. To resolve the problem, this thesis proposes a hierarchical back-off model for Hiero grammar, an instance of a synchronous context free grammar (SCFG), on the basis of the hierarchical Pitman-Yor process. The model can extract a compact rule and phrase table without resorting to any heuristics by hierarchically backing off to smaller phrases under SCFG. Experimental results show that the proposed model achieved higher or at least comparable translation quality with comparable rule-table sizes against a previous Bayesian model on various language pairs; German/French/Spanish/Japanese-to-English. When compared against heuristic models, our model achieved comparable translation quality on a full size of the German and English language pair with less than 10% of rule-table size. Comparison to the strong significance pruning method also shows that proposed method achieved comparative translation quality with a smaller rule-table size.
As described above, this thesis proposes methods based on unsupervised learning to deal with two problems existing in the translation model of statistical machine translation system, and the effectiveness of proposed methods is verified through experiments.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

注意：論文要旨は、東工大リサーチリポジトリ（T2R2）にてインターネット公表されますので、公表可能な範囲の内容で作成してください。
Attention: Thesis Summary will be published on Tokyo Tech Research Repository Website (T2R2).