論文 / 著書情報
Article / Book Information

題目(和文)	特許翻訳のためのバイリンガル知識の獲得に関する研究
Title(English)	
著者(和文)	田村晃裕
Author(English)	Akihiro Tamura
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9327号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:奥村　学,小林　隆夫,住田　一男,熊澤　逸夫,篠崎　隆宏,高村　大也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9327号, Conferred date:2013/9/25, 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

（博士課程）
Doctoral Program

論 文 要 旨
THESIS SUMMARY

専攻： Department of	物理情報システム　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学籍番号： Student ID Number		指導教員（主）： Academic Advisor(main)	奥村　学 教授
学生氏名： Student's Name	田村　晃裕	指導教員（副）： Academic Advisor(sub)	

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

近年，企業のグローバル化に伴い，日本以外の国において知的財産を守る必要性が高まっている．そのため，日本語以外の特許調査や特許出願を効率良く行う手助けとなる，特許の機械翻訳（特許翻訳）のニーズが高まっている．また，特許のデータは公開されていて利用しやすいこともあり，特許翻訳は，学術的な研究も盛んな分野である．そこで，本論文では，特許翻訳が抱える二つの問題を解決することで，特許翻訳の性能向上を試みる．

一つ目の問題は，特許では，一般的な用語ではない専門用語や造語が多く使われるため，特許翻訳では，翻訳モデルに含まれずに翻訳できない未知語が多いという問題である．この問題を解決するため，コンパラブルコーパスから翻訳対を獲得する手法を提案する．従来のコンパラブルコーパスからの翻訳対抽出手法は，シード翻訳対との同一文脈における共起関係（直接的関係）が類似する単語対を翻訳対として抽出する．そのため，シード翻訳対が小規模な場合，翻訳関係を判定する手がかりである，同一文脈で出現するシード翻訳対が少なくなり，性能が悪いという問題がある．そこで，シード翻訳対との直接的関係だけではなく間接的関係も利用して，翻訳関係を判定する手法を提案する．具体的には，まず，全ての単語間の直接的関係を辺で結んだグラフを生成する．本論文では，各辺が二単語間の同一文脈における共起関係を表す「共起グラフ」と，各辺が二単語の文脈間の類似関係を表す「類似グラフ」の二つを提案する．その後，グラフベースの手法であるラベル伝播を用いて，各単語に対し，全てのシード翻訳対との間接的な関連度を求める．この間接的な関連度は，グラフ上で距離が近いほど値が大きくなる．そして，全てのシード翻訳対との関係が似ている単語対を翻訳対として抽出する．日本語と英語の特許文書から作成したコンパラブルコーパスからの未知語に対する翻訳対抽出評価を通じて，提案手法は，従来手法よりも精度良く翻訳対を抽出できることを示す．抽出精度が高くなる分，多くの未知語に対する正しい翻訳対を抽出でき，未知語を削減できるため，特許翻訳の性能向上につながる．また，評価と考察を通じて，類似グラフは，単語間の偶然の共起関係がもたらす悪影響を緩和すると共に，同義語を巧みに同一視することで，共起グラフを用いた場合よりも更に高い抽出精度を達成できることを示す．

二つ目の問題は，複雑で長い文が多い特許翻訳に有効な統語情報に基づく機械翻訳では，統語情報として使われる品詞が翻訳に適しているとは限らないという問題である．この問題を解決するため，コーパスから翻訳に適した品詞を獲得する手法を提案する．従来のコーパスからの品詞導出手法は，品詞付与対象の言語（本論文では原言語）の状況にのみ基づいて品詞を導出し，翻訳相手の言語（本論文では目的言語）を考慮しないため，翻訳に有効な品詞を導出できないという問題がある．そこで，従来の無限ツリーモデルをベースに，翻訳相手の言語での振る舞いも考慮したバイリンガルな品詞を導出する手法を提案する．具体的には，原言語と目的言語の単語単位の対応関係に基づき，目的言語の情報を原言語の係り受け構造に埋め込むことで，シンボルをバイリンガル化する．そして，原言語と目的言語の両方の情報を持つバイリンガルなシンボルに基づいて品詞を導出する．本論文では，バイリンガルなシンボルの生成過程が異なる2種類のモデル

（「結合モデル」と「独立モデル」）を提案する．結合モデルでは，各隠れ状態は，原言語の単語とその単語に対応する目的言語の情報を結合させた結合文字列をシンボルとして出力する．そのため，目的言語での振る舞いが異なれば，異なるシンボルとなる．そして，品詞導出の際には，異なるシンボル出力確率に基づいて品詞を導出するため，目的言語における違いを反映した品詞を導出できる．独立モデルでは，各隠れ状態は，原言語の単語とその単語に対応する目的言語の情報を，別々，独立に出力する．そして，品詞導出の際には，目的言語のシンボル出力確率にも基づいて品詞を導出するため，目的言語における違いを反映した品詞を導出できる．日英特許翻訳による評価を通じて，統語情報に基づく機械翻訳であるForest-to-String 翻訳システムで，提案手法が導出した品詞を使うことにより，既存の品詞や従来手法が導出した品詞を使うよりも性能が良くなることを示す．また，独立モデルは，結合モデルで生じるシンボルのスパースネス問題を解決することで，結合モデルよりも翻訳により適した品詞を導出し，特許翻訳性能を更に改善できることを示す．

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	物理情報システム　専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of　（　工学　）
学籍番号： Student ID Number		指導教員（主）： Academic Advisor(main)	奥村　学 教授
学生氏名： Student's Name	田村　晃裕	指導教員（副）： Academic Advisor(sub)	

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

In recent years, there has been a growing need for patent machine translation due to the globalization of business. In this paper, we aim to improve the performance of patent machine translation by solving two problems.
The first problem is that there are many unknown words in patent machine translation, which cannot be translated. To solve this problem, this paper proposes a novel method of bilingual lexicon extraction from comparable corpora using graph-based label propagation. A previous study found that performance drastically decreases when the coverage of a seed lexicon is small. We address this problem by using indirect relations with bilingual seeds together with direct relations, in which each word is represented by a distribution of lexical seeds. The seed distributions are propagated over a graph that represents relations among words. Translation pairs are extracted by identifying word pairs with high similarities in the seed distributions. Evaluations on comparable corpora of English and Japanese patent documents show that our proposed graph propagation method outperforms conventional methods.
The second problem is that existing Part-of-Speech (POS) tagsets used in patent machine translation are not optimal for statistical machine translation (SMT). To solve this problem, this paper proposes a nonparametric Bayesian method for inducing POS tags for SMT using bilingual observations. In particular, we extend the monolingual infinite tree model (Finkel et al., 2007) to a bilingual scenario: each hidden state (POS tag) of a source-side dependency tree emits a source word together with its aligned target word, either jointly (joint model), or independently (independent model). Evaluations of Japanese-to-English translation on the NTCIR-9 data show that our induced Japanese POS tags for dependency trees improve the performance of a forest-to-string SMT system. Our independent model gains over 1 point in BLEU by resolving the sparseness problem introduced in the joint model.

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 2 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 2 copies of 800 Words (English).