単語の統語的関係を対訳辞書構築のためにいかに有効に利用できるか？

Daniel Andrade、博士課程２年生
東京大学大学院、情報理工学系研究科、コンピュータ科学専攻、辻井研究室

対訳辞書構築の動機 - 造語

同じ文のbag-of-wordsを持っている単語の区別

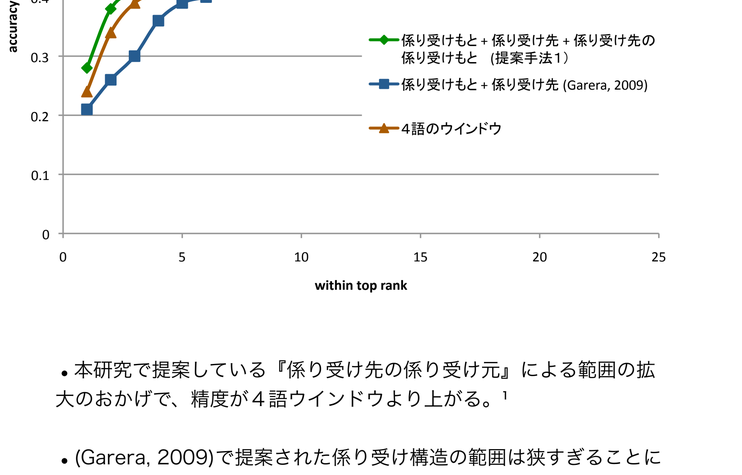
アディア：
同じトピックで共起する単語がbag-of-wordsで区別できないけれど、係り受け構造の周辺が異なる。同じトピックで共起する単語周辺のbag-of-wordsは同じだが、係り受け構造が異なる。

前の例：
…、カードキーでドアが開けなくなった。
…., the **side** door does not open with key.

sideと『キー』のbag-of-wordsの重複する単語は『ドア／door, 開く／open, 冬／winter』
keyと『キー』のbag-of-wordsの重複する単語も『ドア／door, 開く／open, 冬／winter』

しかし、係り受け構造で
sideの係り受け先は『ドア／door』
keyの係り受け先は『開く／open』で『キー』の係り受け先と同じ

日英の係り受け構造の変更



未知の単語の翻訳

日本語でも英語でも新しい単語が作られ、既存の単語の意味が広がる。

様々な単語の翻訳に役立つ手掛かりが考えられている：
翻字 :キー = "key"
複合: 食物 "dietary" + 繊維 "fiber" = "dietary fiber"

ただし、そういう手掛かりが言葉の翻訳に十分なわけではない：
例： 迷惑 "annoying" + メール "mail" != "annoying mail", 正しい翻訳： "spam"

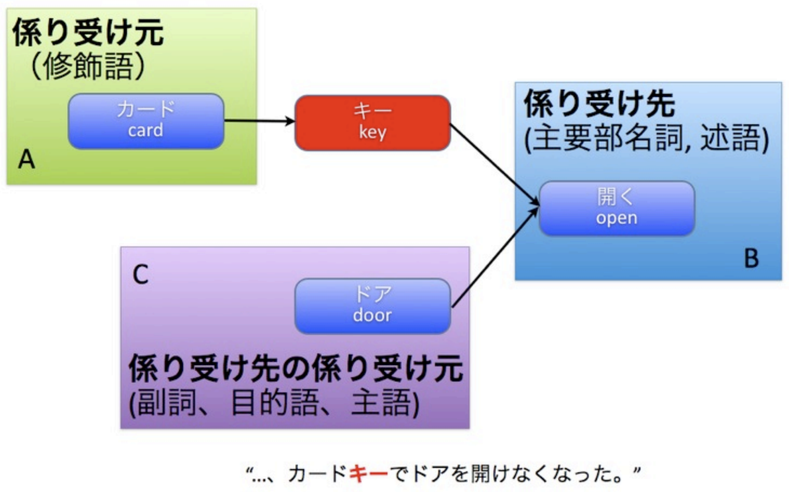
普遍的に文脈の情報が利用できる

基盤の仮説:
単語の意味が同じならば単語の文脈も同じである

例：キー／"key" の日本語と英語のコーパスから抜き出した文脈:

“冬で鍵穴が凍り付き、カードキーでドアが開けなくなった。”
“In winter when it is cold, the side door does not open with key.”

係り受け構造を注目語から見て三つのグループに分ける



提案手法

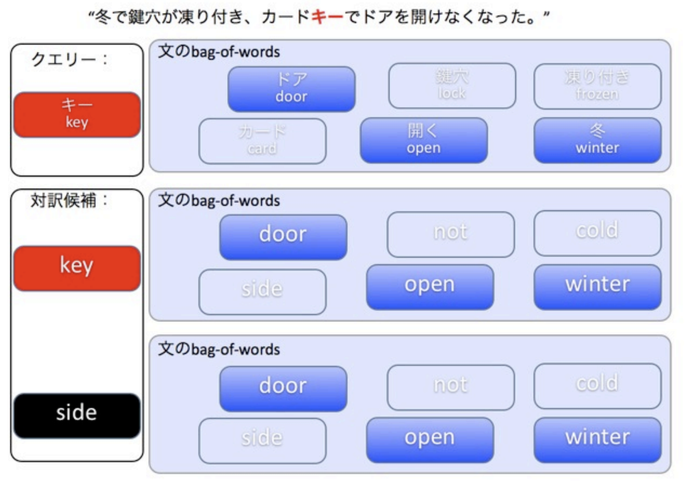
手掛かり１：原言語の文中で共起する単語は、目標言語においても共起する傾向がある。（従来研究）
手掛かり２：係り受け関係を持つ単語は意味的な関係を持つ為、言語横断の比較が可能 (本研究の着眼点)

手掛かり１の信頼性は低いが、文の共起頻度を統計的にうまく分析することにより、有意なビボット語の抽出ができる。
手掛かり２の信頼性は高いが、手掛かり１より局所的であるため、係り受け関係にない単語を抽出できない。

提案手法１：
基づいた手法は(Andrade, 2010)で発表されている手法である。
ただし、共起の範囲はA・B・Cで出現する単語である。（手掛かり２）

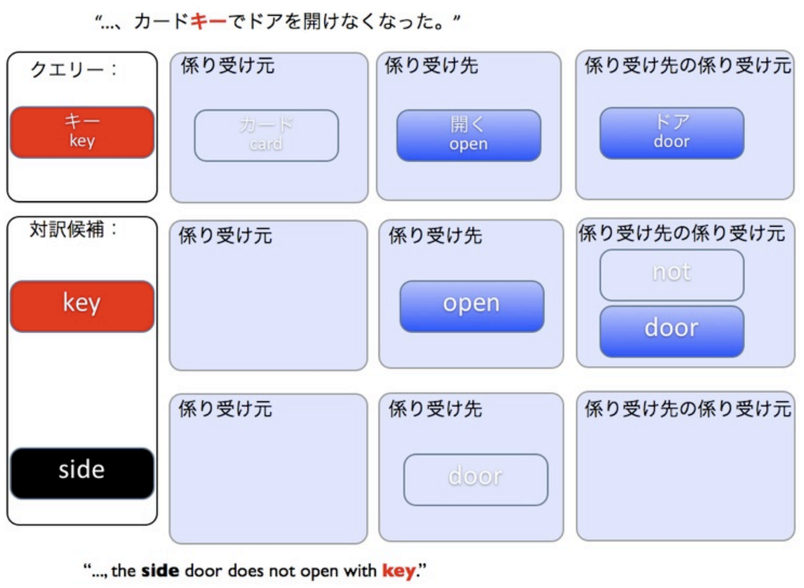
提案手法２：
提案手法１と同様の係り受け構造の情報を用い、文中の共起頻度の情報を組み合わせることである。
（手掛かり１ + 手掛かり２）

必要なリソース:
- Comparable corpora (違う言語で書いてトピックが似ているコーパスのペア。対訳コーパスが必要ではない。)
- 一般的な対訳辞書
(必要とする単語の周りの文脈を翻訳のために利用する。)



keyもsideも名詞のため、対訳候補と見なす。
問題：sideの文中の共起単語とkeyの文中の共起単語が同じで、区別できない。

日英の係り受け構造で区別されている対訳候補



実験のリソース

コーパスペア：
国道交通省とUSA National Highway Traffic Safety Administrationの車クレームの集合。MSTとKNPで係り受け解析

対訳辞書：
日英辞書JMDic

正解セット：
１００語の日本語の名詞と英語の対訳語

予備実験結果

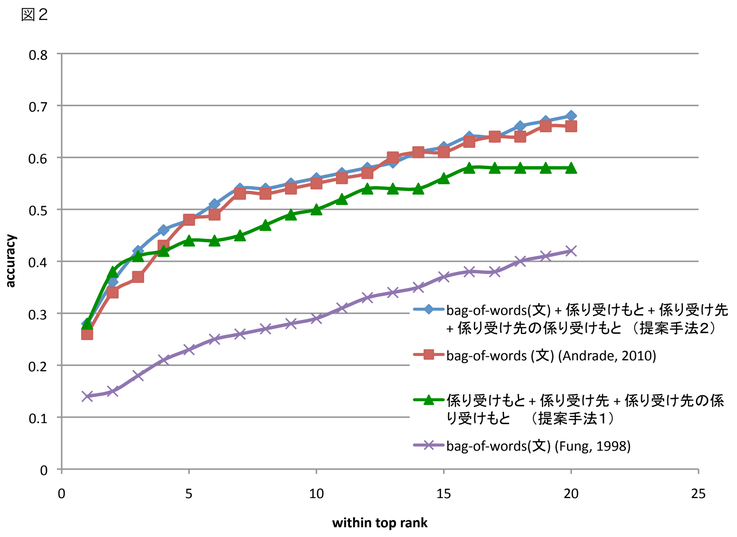
図１

- 本研究で提案している『係り受け先の係り受け元』による範囲の拡大のおかげで、精度が４語ウインドウより上がる。[1]

- (Garera, 2009)で提案された係り受け構造の範囲は狭すぎることによって、４語ウインドウの精度に及ばない。[2]

[1] 単語ウインドーサイズは４語、必要とする単語から２語左と２語右、（Haghighi、２００８; Garera, 2009)と同様。
[2] (Garera, 2009) で係り受け元の係り受け元と係り受け先の係り受け先の情報も使われており、本実験で使わない。

図２



- 提案手法（１）は 上位３位まで　bag-of-words（文）より精度が高い、ただし、上位４位から、bag-of-words(文)に及ばない。
- 提案手法（２）は提案手法（１）とbag-of-words(文）の情報を組み合わせることにより、上位４位からもbag-of-words(文）に負けない。

上位 n 出力までの精度（パーセント）

	Top1	Top2	Top3	Top10	Top20
bag-of-words(文)	26	34	37	55	66
提案手法１	**28**	**38**	41	50	58
提案手法２	**28**	36	**42**	**56**	**68**

実験の解釈

AとBの係り受けグループしか使わないとカバーされてない重要な単語が多く、精度は４語のウインドウを使ったモデルより低い。
さらに Cグループを使うと、４語ウインドウの使用より多くの重要な単語を抽出ができ、精度が上がる。(提案手法１、図１)

単純に文中の単語の共起頻度を使うモデルと比べると、提案手法１は上位３出力までは優位なだが、それ以降は精度が下がる。
共起の範囲は係り受け関係にある単語に制限されると、文の範囲より狭く、重要なピボット語を逃す場合があることが考えられる。
一方、単純に文中の単語の共起頻度をうまく統計的に分析すると有意な単語関係を見つけることができる。

その結果、文中の単語の共起頻度とそれぞれの係り受け構造中（提案手法１）にある単語の共起頻度の情報を組み合わせる手法も提案する。(提案手法２)
その手法は文中の単語の共起頻度を使う手法より一貫して精度が高い。(図２)

二つのクエリーの上位 5 位までの出力

手法	クエリー	Top1	Top2	Top3	Top4	Top5
bag-of-words(文)	樹脂	radiator	plastic	head	metal	tube
提案手法２		plastic	radiator	head	metal	tube
bag-of-words(文)	リコール	number	part	dealership	dealer	recall
提案手法２		number	part	recall	dealership	dealer

- 提案手法（２）の出力とbag-of-words(文）の出力において正しい対訳の位をペアワイズに比較すると、 提案手法（２）出力での正しい対訳語の位 は bag-of-words(文）出力の位より高い場合が多い。(結果は統計上有意、p-value 0.01)

まとめ

ある未知語と共起する既知単語(ビボット語）を翻訳し、目標言語における候補の文脈を比べ、類似性によって、対訳候補をランキングする手法が考えられている。(Fung, 1998)
ただし、未知語と共起するビボット語は全部ではなく、統計上有意なビボット語しか抽出しないことが有利である。(Andrade, 2010)
従来の研究は文脈として、文中の単語または単語ウインドウを用いている。

本研究の目的は文中の単語の共起の代わりに、係り受け関係にある単語の共起を用いることである。
対象語から見た係り受け構造に基づいて、ピボット語の出現位置を三つのグループに分ける：
係り受け元(A)、係り受け先(B)、係り受け先の係り受け元(C)
それによって対訳精度の向上が可能だと考えられる。

参考文献

このポスターで紹介している研究はまだ雑誌・国際学会で発表していないので、現在、雑誌の記事を書いている最中です。つきまして、この手法の詳細はそのうち公開します。
興味がある方は是非連絡してください： Daniel.Andrade.Silva@gmail.com (ダニエル)

Andrade, Daniel and Nasukawa, Tetsuya and Tsujii, Junichi. 2010. Robust Measurement and Comparison of Context Similarity for Finding Translation Pairs. Proceedings of the International Conference on Computational Linguistics

Fung, P. 1998.A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. Lecture Notes in Computer Science

Garera, N., C. Callison-Burch, and D. Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. Proceedings of the Conference on Computational Natural Language Learning,

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. Proceedings of the Annual Meeting of the Association for Computational Linguistics.