SHARP

情報通信研究機構　委託研究

機械翻訳に用いる翻訳知識の自動獲得 概要紹介

2004年10月29日

シャープ株式会社
情報通信事業本部

紹介の概要

- はじめに
  - 研究の背景
  - 研究領域
- 2003年度研究の紹介
  - 基本方針
  - 手法の説明
  - 実験の説明
  - 実験結果
  - 抽出できた対訳の例
- まとめ
- 2004年度の目標

はじめに

- 本研究の目的
  - 機械翻訳開発者を支援するシステムを開発することを通じて、
    - ・機械翻訳の訳質向上
    - ・機械翻訳開発の効率向上…開発のスピードアップ

    → 多言語コミュニケーションをより豊かにする
  - ユーザを直接支援するシステムも視野に入れる。
    (ユーザによるカスタマイズ可能な翻訳システム、読解支援システムとの融合、等)
- この目的のために:
  - 対訳コーパスからの翻訳情報の高精度な抽出
    (文対応の精度が不十分でも使えるもの)

研究の背景

機械翻訳システムの辞書開発

…大量、迅速、低コストに

→　(半) 自動的な対訳情報獲得

統計的手法

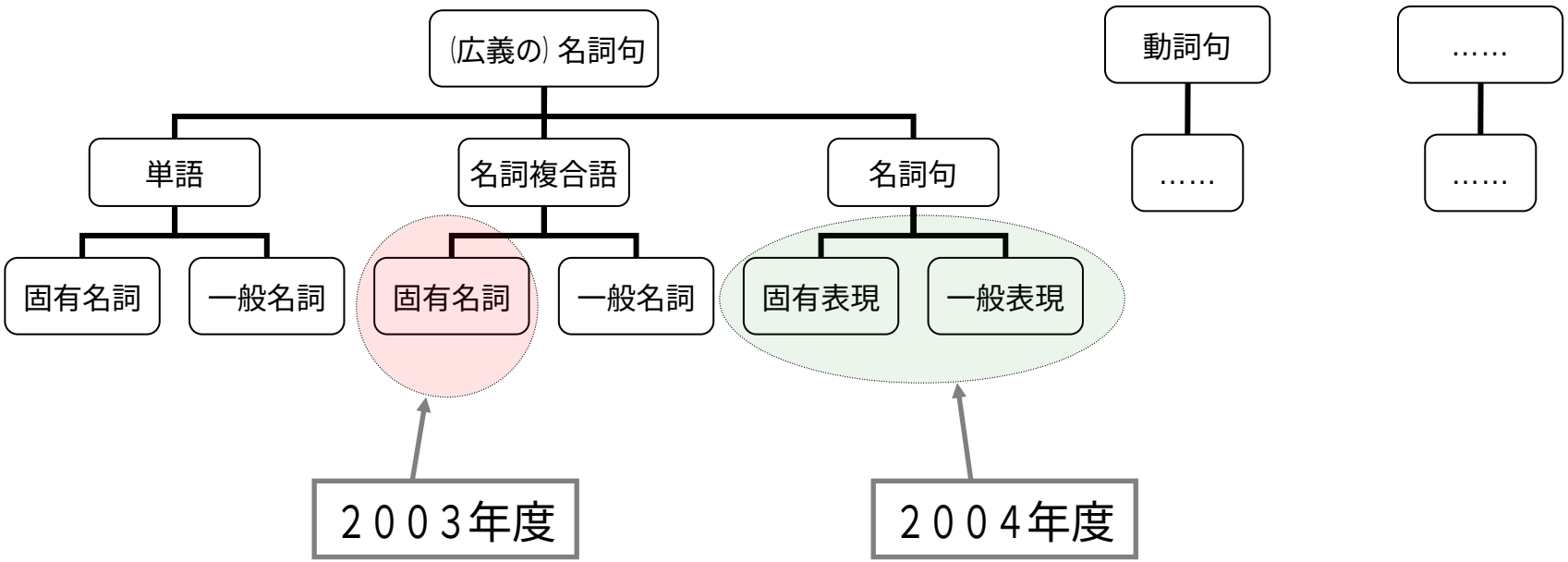
… 得られるのは高頻度表現

→ 現行の機械翻訳 (辞書・パターン) に存在する可能性が高い

現状の機械翻訳で訳せない

低頻度の対訳表現を獲得したい

対訳獲得の研究領域



※既存研究の状況(研究の新規性)や、翻訳応用への有用性に応じて、
対象領域を柔軟に変更していく必要性あり。

2003年度研究 —— 基本方針

- 現状の機械翻訳（英→日）で訳せない表現（訳文中に英語のまま残る）を、
  対応する日本語ネイティブ文と対比し、
  訳出されていない箇所（＝辞書未登録語）に対応する表現（訳語）を推定。
- 対象を「大文字始まりの連語」に限定。（固有名詞候補）
- 新聞記事（読売新聞・The Daily Yomiuri）を元にした日英対訳コーパスを利用して、「未登録固有名詞」の訳語を推定。

2003年度研究――手法の概要

新聞記事（読売新聞・The Daily Yomiuri）を元にした
日英対訳コーパス（機械処理による文対応付け済み）

その英語文を機械翻訳で英→日翻訳

→ 英語文(e1)・日本語文(h1)・機械翻訳文(m1)の三つ組を得る

(e1) The government says it is trying to provide more information and strengthen public relations by establishing an "International Cooperation Plaza" to collect and dispatch information, by holding symposiums and by publishing annual reports on how the government is carrying out ODA projects.

訳語

(h1) 政府は、総合的な情報収集・発信のための総合施設「国際協力プラザ」の開設、シンポジウムの開催、援助実施状況に関する年次報告書の発行など、情報公開や広報に努めているというが、事業そのものが不信をもたれては何もならない。

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

→ 日本語文（h1）・機械翻訳文（m1）を使用

2003年度研究――推定手法

- 品詞による絞り込み（名詞のみを対象）
- コーパス全体中の出現頻度による絞り込み
  対訳候補のうち、以下がすべて「1」のものを抽出
  - 機械翻訳文コーパス（日本語）における「未登録語」の出現頻度
  - 対訳コーパスの日本語文における「訳語候補」の出現頻度
  - 上記和文対における「未登録語」と「訳語候補」の同時出現頻度
- 訳語候補のスコア付け

1. 「未登録語の構成単語を機械翻訳した訳語を単純に合成したもの」と「訳語候補」との類似性（S_1）
2. 「未登録語のローマ字読み」と「訳語候補の読み」との類似性（S_2）
3. 「未登録語の近傍に現れる名詞の集合」と「訳語候補の近傍に現れる名詞の集合」との類似性（S_3）
4. 「訳語候補」と「同一の語」が機械翻訳文にも存在するか否か（S_4）
   → 当該「訳語候補」は、求めたい「未登録語」の訳語でない可能性大

候補のスコア付け――例1

1. 未登録語の構成単語の単純合成訳と訳語候補の類似性による優先順位付け

訳語候補①

(h1) 政府は、総合的な情報収集・発信のための総合施設「国際協力プラザ」の開設、シンポジウムの開催、援助実施状況に関する年次報告書の発行など、情報公開や広報に努めているというが、事業そのものが不信をもたれては何もならない。

訳語候補②

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

「未登録語」構成単語の訳語集合

- インターナショナル，協力，プラザ

「訳語候補①」の形態素集合

- 国際，協力，プラザ

「未登録語」と「訳語候補①」の形態素の和集合：国際，インターナショナル，協力，プラザ　(4個)

積集合：協力，プラザ　(2個)

ジャッカード係数 ＝ ½

「訳語候補②」の形態素集合

- 事業，そのもの

「未登録語」と「訳語候補②」の形態素の和集合：インターナショナル，協力，プラザ，事業，そのもの　(5個)

積集合：(0個)

ジャッカード係数 ＝ 0

候補のスコア付け――例1

2. 未登録語のローマ字読みと訳語候補の読みとの類似性による優先順位付け

訳語候補①

(h1) 政府は、総合的な情報収集・発信のための総合施設「国際協力プラザ」の開設、シンポジウムの開催、援助実施状況に関する年次報告書の発行など、情報公開や広報に努めているというが、事業そのものが不信をもたれては何もならない。

訳語候補②

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

「未登録語」のローマ字読み集合（日本語ローマ字読み不可の形態素はそのまま）

- international, cooperation, plaza　日本語ローマ字読み不可

「訳語候補①」の読み集合
- コクサイ, キョウリョク, プラザ

「未登録語」と「訳語候補①」の読みの和集合：international, cooperation, plaza, コクサイ, キョウリョク, プラザ（6個）
積集合：（0個）
ジャッカード係数 ＝ 0

「訳語候補②」の読み集合
- ジギョウ, ソノモノ

「未登録語」と「訳語候補②」の読みの和集合：international, cooperation, plaza, ジギョウ, ソノモノ（5個）
積集合：（0個）
ジャッカード係数 ＝ 0

候補のスコア付け――例1

3. 未登録語の近傍名詞集合と訳語候補の近傍名詞集合との類似性による優先順位付け

訳語候補①

(h1) 政府は、総合的な情報収集・発信のための総合施設「国際協力プラザ」の開設、シンポジウムの開催、援助実施状況に関する年次報告書の発行など、情報公開や広報に努めているというが、事業そのものが不信をもたれては何もならない。

訳語候補②

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

「未登録語」の近傍名詞集合
・出版/[業]/年報，情報，広報

最大前後各2単語の名詞　※複合語レベルでカウント
(その文の名詞の個数に応じて、採用する最大単語数を決定)

「訳語候補①」の近傍名詞集合　前後各4単語
・総合/施設，発信，情報/収集，総合/[的]，
開設，シンポジウム，開催，援助/実施/状況

「候補①」の近傍名詞集合の形態素の
和集合：総合，施設，情報…（14個）
積集合：情報（1個）
ジャッカード係数 ＝ 1/14

「訳語候補②」の近傍名詞集合
・広報，情報/公開，発行，
年次/報告/[書]，不信

「候補②」の近傍名詞集合の形態素の
和集合：広報，情報，公開…（9個）
積集合：広報，情報（2個）
ジャッカード係数 ＝ 2/9

候補のスコア付け――例1

4. 訳語候補と同一語の存在/非存在による優先順位付け

訳語候補①

(h1) 政府は、総合的な情報収集・発信のための総合施設「国際協力プラザ」の開設、シンポジウムの開催、援助実施状況に関する年次報告書の発行など、情報公開や広報に努めているというが、事業そのものが不信をもたれては何もならない。

訳語候補②

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

「訳語候補①」を構成する単語「国際」「協力」「プラザ」は、(m1) に存在しない。

「訳語候補②」を構成する単語「事業」「そのもの」は、(m1) に存在しない。

∴ この属性に関し、① と ② は同等に確からしい。
(①:0.5点、②:0.5点)

総合的評価

- 総合評価式

$$S = C + \sum_{i=1}^{4} (W_i \times S_i)$$

- 重みの値

	C	W_1 単語訳の合成	W_2 ローマ字	W_3 近傍名詞	W_4 同一語の存在
経験則	0	2	3	1	2
回帰分析	-4.58	20.75	15.04	3.58	2.81

総合評価――例1

(h1) 政府は、総合的な情報収集・発信のための①総合施設「②国際協力プラザ」の③開設、シンポジウムの開催、④援助実施状況に関する⑤年次報告書の発行など、情報公開や広報に努めているというが、⑥事業そのものが⑦不信をもたれては何もならない。

(m1) 政府によれば、それは、更に多くの情報を提供しようとしており、そして、どのように政府が ODA プロジェクトを実行していても、シンポジウムを開催することによる、そして、出版業年報による情報を集めて、ディスパッチするための「International Cooperation Plaza」を確立することによって広報を強化する。

未登録語

- 未登録語の構成単語の単純合成訳
- 未登録語のローマ字読み
- 近傍名詞集合
- 同一語の存在

訳語候補	正否	単語訳の合成		ローマ字		近傍名詞		同一語の存在		重み和	総合点
		素点	重み点	素点	重み点	素点	重み点	素点	重み点		
①総合/施設	×	0	0	0	0	0.071	0.256	0.5	1.40	1.66	-2.92
②国際/協力/プラザ	○	0.5	10.38	0	0	0.071	0.256	0.5	1.40	12.04	7.46
③開設	×	0	0	0	0	0.056	0.199	0.5	1.40	1.60	-2.98
④援助/実施/状況	×	0	0	0	0	0.143	0.511	0.5	1.40	1.91	-2.67
⑤年次/報告/書	×	0	0	0	0	0.143	0.511	0.5	1.40	1.91	-2.67
⑥事業/そのもの	×	0	0	0	0	0.222	0.796	0.5	1.40	2.20	-2.38
⑦不信	×	0	0	0	0	0.286	1.023	0.5	1.40	2.42	-2.16

候補のスコア付け――例2

2. 未登録語のローマ字読みと訳語候補の読みとの類似性による優先順位付け

訳語候補②

(h2) 新たに目撃されたのは倉敷市六口島の南東約三百メートルの備讃瀬戸で、

訳語候補① 坂出海上保安署の巡視艇「ことかぜ」乗組員が、サメ二匹を目撃。

(m2) Kotokaze の乗組員、Sakaide Maritime Safety Station に付けられた巡視船は、

Kurashiki の海岸から 300 メートルについて 2 匹のサメを目撃した。

未登録語

「未登録語」のローマ字読み集合（日本語ローマ字読み不可の形態素はそのまま）

- サカイデ，マリチメ，safety，station

「訳語候補①」の読み集合

•サカイデ，カイジョウ，ホアン，ショ

「未登録語」と「訳語候補①」の読みの和集合：サカイデ，マリチメ，safety，station，カイジョウ，ホアン，ショ（7個）

積集合：サカイデ（1個）

ジャッカード係数 ＝ 1/7

「訳語候補②」の読み集合

•コト，カゼ

「未登録語」と「訳語候補②」の読みの和集合：サカイデ，マリチメ，safety，station，コト，カゼ（6個）

積集合：（0個）

ジャッカード係数 ＝ 0

候補のスコア付け――例2

4. 訳語候補と同一語の存在/非存在による優先順位付け

訳語候補② 訳語候補③

(h2) 新たに目撃されたのは倉敷市六口島の南東約三百メートルの備讃瀬戸で、
訳語候補① 坂出海上保安署の巡視艇「ことかぜ」乗組員が、サメ二匹を目撃。
(m2) Kotokaze の乗組員、Sakaide Maritime Safety Station に付けられた巡視船は、
Kurashiki の海岸から 300 メートルについて 2 匹のサメを目撃した。

未登録語

「訳語候補①」を構成する単語「坂出」「海上」「保安」「署」は、
(m2) に存在しない。

「訳語候補②」を構成する単語「巡視」「艇」のうち、
「巡視」が (m2) に存在する。

「訳語候補③」を構成する単語「こと」「かぜ」は、
(m2) に存在しない。

∴ この属性に関し、① と ③ は、② よりも確からしい。
(①:0.5点、②:0点、③:0.5点)

総合評価——例2

(h2) 新たに目撃されたのは倉敷市六口島①の南東約三百メートルの備讃瀬戸②で、
坂出海上保安署③の巡視艇④「ことかぜ」⑤乗組員が、サメ⑥二匹を目撃⑦。
(m2) Kotokaze の乗組員、Sakaide Maritime Safety Station に付けられた巡視船は、
Kurashiki の海岸から 300 メートルについて 2 匹のサメを目撃した。

未登録語

- 未登録語の構成単語の単純合成訳
- 未登録語のローマ字読み
- 近傍名詞集合
- 同一語の存在

訳語候補	正否	単語訳の合成		ローマ字		近傍名詞		同一語の存在		重み和	総合点
		素点	重み点	素点	重み点	素点	重み点	素点	重み点		
①倉敷/市/六口島	×	0	0	0	0	0	0	0.5	1.40	1.40	-3.17
②讃/瀬戸	×	0	0	0	0	0	0	0.5	1.40	1.40	-3.17
③坂出/海上/保安/署	○	0	0	0.143	2.15	0.400	1.432	0.5	1.40	4.98	0.41
④巡視/艇	×	0	0	0	0	0	0	0	0	0.00	-4.58
⑤こと/かぜ	×	0	0	0	0	0.400	1.432	0.5	1.40	2.83	-1.74
⑥サメ	×	0	0	0	0	0.500	1.790	0	0	1.79	-2.79
⑦目撃	×	0	0	0	0	0.250	0.895	0.5	1.40	2.30	-2.28

2003年度研究――訳語対抽出の正解率

	第1位 (同点1位を含む)	上位2位
経験則	83.7% (221/264)	92.8% (245/264)
回帰分析	86.4% (228/264)	95.1% (251/264)

2003年度研究――獲得できた対訳の例

Growth Initiative	成長イニシアチブ
APEC Ministerial Meeting	ＡＰＥＣ閣僚会議
IAEA Safeguards Agreement	ＩＡＥＡ保障措置協定
De Gaulle	ドゴール
World Economic Summit	世界経済サミット
G-7 Action Plan	Ｇ７行動計画
Energy Sector	エネルギー部門
APEC Working Group	ＡＰＥＣ作業部会
JR Wakayama Station	ＪＲ和歌山駅
Postal Savings Law	郵便貯金法
Foreign Trade Control Law	外国貿易管理法
DPRK Government	北朝鮮政府
Basaman Palace	バスマン宮殿
Premier Li	李首相
Oji Sports Garden	王子公園競技場
Central Labor Relations Committee	労働委員会
Population Registration Act	人口登録法

2003年度の研究成果

- 商用機械翻訳システムの辞書に登録されていない複合語の訳語を、対訳コーパスを用いて抽出する方法を示した。
- 出現頻度の低い複合語を対象とし、訳語を獲得することができた。
- 訳語対抽出の正解率
  - スコア第１位が正解　…　８６．４％
  - スコア上位２位以内が正解　…　９５．１％
    (重み付けを回帰分析で決めた場合)

2004年度の目標

フレーズレベルの名詞句の対訳獲得

	2003年度研究	2004年度研究
英語の条件	機械翻訳で未訳出の固有名詞連語	前置詞や接続詞、形容詞などを含んだ名詞句
日本語の条件	名詞連続	連体の助詞や形容詞などを含んだ名詞句
頻度条件	低頻度	頻度に関わらず、未登録語
獲得する対訳	複合語レベル	フレーズレベル