実例に基づいた翻訳

長岡技術科学大学　自然言語処理研究室　高橋寛治

はじめに

文献紹介

- 実例に基づいた翻訳
  - 佐藤 理史、長尾 真（京都大学工学部）
  - 情報処理学会第38回全国大会、333-334、1989-03-15

はじめに

- アナロジーによる翻訳
  - 機械翻訳システムの問題点を克服するため示された

- 具体的には？

類似の翻訳例を模倣することにより翻訳

⇩ 実例に基づいて推論

Memory-based Translation

Memory-based Translation(MBT)

基本的な考え方

（1）翻訳例のデータベースを用意

（2）2つの翻訳例間に距離を定義

（3）未知の翻訳の適切さの推論

データベースの形式

- 翻訳フレームと単語対は、辞書的規則の役割
  - キー
    - 翻訳フレーム
      - ソースとスロット数
    - 辞書対
      - ソース側

翻訳例の形式

	ソース	ターゲット
ヘッド	eat	たべる
スロット1	he	彼
スロット2	potato	じゃがいも

翻訳フレーム

翻訳フレーム	(eat たべる 2)
単語対1	(he 彼)
単語対2	(potato じゃがいも)

翻訳例間の距離

- あるD(翻訳例の集合全体)において、2つの翻訳例e_i,e_j間の距離$\Delta(D,e_i,e_j)$を以下のように定義

$$\Delta(D, e_i, e_j) = \begin{cases} \sum_{k=1}^{n} \delta(D, e_i.f, e_i.s_k, e_j.s_k) & \text{if } e_i.f = e_j.f \\ \infty & \text{otherwise} \end{cases}$$

- $d(D,p_i,p_j)$は、2つの単語対p_i,p_jの距離を表す
- $\omega_k(D,f_i)$は、翻訳フレームf_iのスロットkの重み

$$\delta(D, e_i.f, e_i.s_k, e_j.s_k) = d(D, e_i.s_k, e_j.s_k)\, w_k(D, e_i.f)$$

単語対間の距離

- スロット ー 単語対表
  - 2つの単語対の距離を定義

		単語対			
		1	2	3	...
ス	1:1	2	0	1	...
ロ	1:2	0	1	0	...
ッ	2:1	1	0	1	...
ト	.	.	.	.	...

- 左の表の列ベクトルをRiと表すとき、2つの単語対間の距離を定義

$$d(D, p_i, p_j) = \frac{1}{similarity(D, p_i, p_j)} - 1$$

$$similarity(D, p_i, p_j) = \frac{{}^t R_i R_j}{\|R_i\| \, \|R_j\|}$$

翻訳フレームのスロットの重み

- スロット数および競合する翻訳フレームの有無によって定義が異なる
- 競合するフレーム
  - (eat たべる 2)　(eat 侵す 2)
  - ソース側のヘッドが同じ、スロット数が同じ

翻訳プロセス

- 1. ソースの値とスロットの数から翻訳フレームを取り出す
- 2. ソースの値から単語対を取り出す。これを各スロットに対して行う
- 3. 1と2を組み合わせて、値がすべて埋まった翻訳フレームをつくる

- これにより、翻訳の候補が複数求まる
- 候補から、最も近い翻訳例をデータベースから探す

	ソース	ターゲット
ヘッド	eat	?
スロット1	he	?
スロット2	vegetable	?

実行例

- 翻訳例1179、翻訳フレーム数276、単語対数395

Source = (PLAY JAPANESE CARD)　*Weight-List* = (.207 .793)

Rank	*Target*	*Distance*	*Most Similar Translation*
1	(する 日本人 トランプ)	1.11(3.74 .429)	(PLAY TARO TENNIS) -> (する 太郎 テニス)
2	(ひく 日本人 トランプ)	7.41(25.3 2.74)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
3	(ひく 日本人 カード)	8.08(25.3 3.58)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
4	(する 日本語 トランプ)	207.(999. 0.0)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本語 カード)	208.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
5	(する 日本人 カード)	208.(999. 1.45)	(PLAY THEY CARD) -> (する 彼ら トランプ)
7	(ひく 日本語 トランプ)	209.(999. 2.74)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
8	(ひく 日本語 カード)	210.(999. 3.58)	(PLAY I VIOLIN) -> (ひく 私 バイオリン)
9	(演じる 日本人 トランプ)	797.(21.7 999.)	(PLAY SHE JULIET) -> (演じる 彼女 ジュリエット)
9	(演じる 日本人 カード)	797.(21.7 999.)	(PLAY SHE JULIET) -> (演じる 彼女 ジュリエット)
11	(演じる 日本語 トランプ)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)
11	(演じる 日本語 カード)	999.(999. 999.)	(PLAY HE ROMEO) -> (演じる 彼 ロメオ)

おわりに

- Memory-based Translationの基本的枠組み
- 翻訳というタスクを考えると
  - 個々の対応関係に基づいた置換プロセス
  - 基本原理や規則性よりも個別性が有意

- MBTのような翻訳例に基づいて翻訳を行うという考え方は非常に有望