単語対応の導入

内山将夫@NICT
mutiyama@nict.go.jp

パラレルコーパスからの情報抽出

- 単語対応 (Word alignment) の抽出 (これ)
- 句対応 (Phrase alignment) の抽出
- 翻訳規則の抽出

単語対応 (Word Alignment) とはなにか

たくさんの対訳文対が，パラレルコーパスとして与えられたとき，各対訳文対において，日本語のどの単語が，英語のどの単語に対応しているかを同定する．
単語対応は，そもそも良く定義できない問題である．つまり，対訳文が与えられたとき，どの単語とどの単語が対応するかは，人間がみても良くわからないのが普通である．
しかし，今のところ，単語対応を求めることは，コーパスベースの翻訳にとって，本質的な問題である．

単語対応の例

- 「今日」「は」「良い」「天気」「だ 」
- 「It」「is」「fine」「today」

において

- 「今日」と「today」は対応する.
- 「良い」「天気」と「fine」は対応する.

その他の「は」「だ」や「It」「is」は，どう判断したら良いだろうか？

課題

適当な対訳文10文について，単語対応を求めること.

単語対応の確率モデル (Brown et al. 1993)

$P(\mathbf{f}|\mathbf{e})$ = フランス語の文$\mathbf{f}$が英語の文$\mathbf{e}$から生成される確率

この $P(\mathbf{f}|\mathbf{e})$ を，英単語と仏単語の対応確率から求めたい．まず，$\mathbf{e}$ と $\mathbf{f}$ は，それぞれ l,m 個の単語からなる．

$$\mathbf{e} = e_1 e_2 \ldots e_i \ldots e_l = e_1^l \tag{1}$$

$$\mathbf{f} = f_1 f_2 \ldots f_j \ldots f_m = f_1^m \tag{2}$$

次に，各仏単語 f_j が，ただ一つの英単語に対応するとして，その英単語の位置を a_j とする．

$$\mathbf{a} = a_1 a_2 \ldots a_j \ldots a_m = a_1^m \tag{3}$$

このとき，f_j に対応する英単語は，e_{a_j} である．もし，f_j に対応する英単語がない場合には，$a_j = 0$ とする．この f_j は，NULL単語から生成されたと考える．したがって，$a_j \in \{0, 1, \ldots, l\}$ である．ある対応,

	f_1	f_2	f_3	f_4	f_5
NULL					x
e_1		x			
e_2	x		x		
e_3				x	

において，$f_1 \to e_2$, $f_2 \to e_1$, $f_3 \to e_2$, $f_4 \to e_3$, $f_5 \to \mathrm{NULL}(e_0)$ より $a_1 = 2$, $a_2 = 1$, $a_3 = 2$, $a_4 = 3$, $a_5 = 0$ である．

ある生成モデル (IBM-Model 1,2 の原型)

$$
\begin{aligned}
P(\mathbf{f}|\mathbf{e}) &= \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}|\mathbf{e}) \\
P(\mathbf{f},\mathbf{a}|\mathbf{e}) &= \text{仏文}\mathbf{f}\text{と対応}\mathbf{a}\text{が英文}\mathbf{e}\text{から生成される確率} \\
&= P(m|\mathbf{e}) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \\
&\quad P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e})
\end{aligned}
$$

$$P(m|\mathbf{e}) = \text{英文}\mathbf{e}\text{が}m\text{単語の仏文になる確率}$$

$P(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) =$ 仏文が f_1^{j-1} まで生成され，それぞれが，a_1^{j-1} の位置の英単語につながっているとき，j 番目の仏語が a_j 番目の英単語につながる確率

$P(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}) =$ 上述の条件に加えて，j 番目の仏語が a_j 番目の単語につながるときに，j 番目の仏単語が f_j である確率

これは，確率の式を厳密に展開したものであることに注意する．上述の式を簡単化したものがIBM Model 1,2 である．別の形の式展開をするとIBM Model 3,4,5となる．