関連語－訳語関連行列を用いた訳語選択と後処理による機械翻訳システムへの適用

綱川　隆司　　**梶　博行**
静岡大学情報学部情報科学科

はじめに

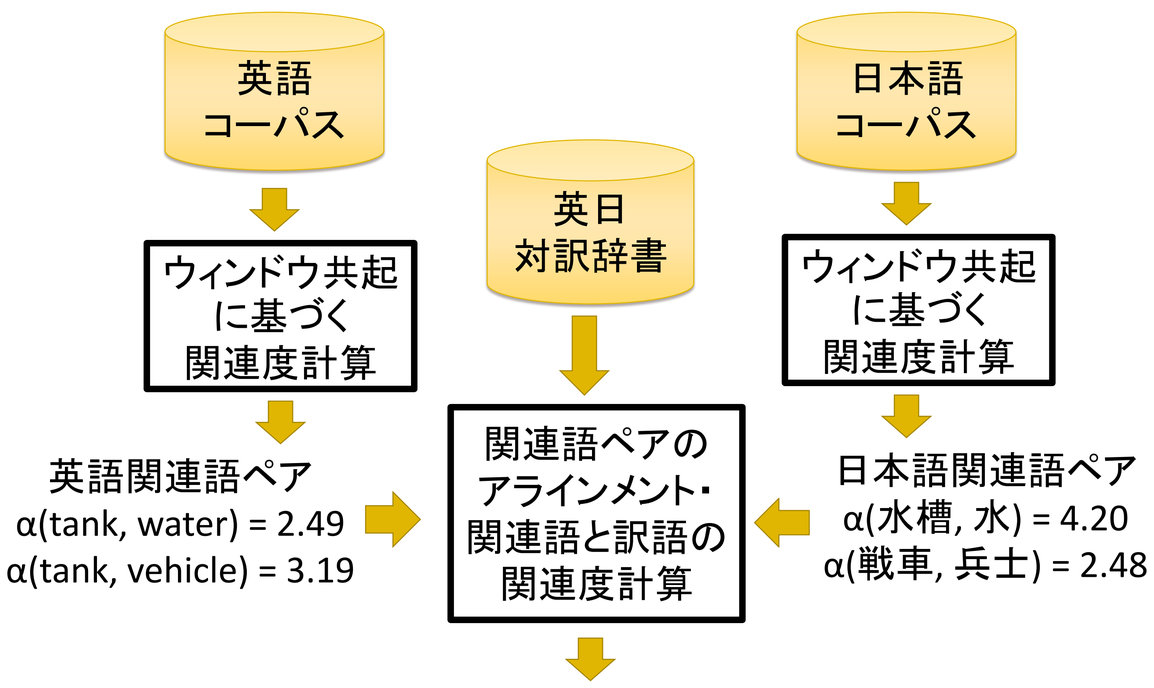
● 翻訳過程において、入力文中の語・句の訳の適切な選択（訳語選択）は最も重要な要素の一つ
　－ 多義語の翻訳の場合、訳語候補の中から入力文で意図した意味の訳語を選択しなければならない
　－ ある語に対して似た意味の訳語が複数ある場合、分野・文脈および使用する用語の統一を考えて出力する必要がある

● 既存の機械翻訳システムでは、訳語の選択について限定的な考慮しかしていない
　－ ルールベースシステムの場合、訳語ごとに訳語選択ルールを記述していく必要があり多くの労力を必要とし、かつ多数のルールの管理が困難になる
　－ 統計ベースシステム・用例ベースシステムの場合、頻度の高い訳やよく使われる単語列に含まれる語が選択されやすい

→ 各訳語がどのような語と関連性があるかをあらかじめコーパスから求めた“関連語－訳語関連行列”を用いて、入力文中の周りの語から適切な訳語を選択する

● 訳語選択の予備実験として、既存の機械翻訳システムの出力結果に対して後処理（訳語の置き換え）を試みる
　－ ルールベースと統計ベースの翻訳システムの結果を比較する

関連語－訳語関連行列の計算



関連語－訳語関連行列

<tank>	水槽	戦車	…
water	3.71	1.30	…
vehicle	1.15	4.83	…
…	…	…	

関連語・訳語間の関連度の反復計算

● 語 x の第 i 関連語 $x'(i)$ と第 j 訳語 $y(j)$ の関連度:「相互に関連のある関連語は同じ訳語を支持する」という仮説に基づき、関連のある他の関連語と訳語の関連度を用いて定義

$$C_n(x'(i),y(j)) = \alpha(x'(i),x)\cdot\frac{\sum_{x''\in A(x,x'(i))}\alpha(x'(i),x'')\cdot C_{n-1}(x'',y(j))}{\max_k \sum_{x''\in A(x,x'(i))}\alpha(x'(i),x'')\cdot C_{n-1}(x'',y(k))} \quad [1]$$

ただし、$\alpha(x'(i), x)$は$x'(i)$とxの関連度（相互情報量）, $A(x, x'(i))$は対象語xと関連語$x'(i)$に共通の関連語の集合

● 初期値C_0: 対訳辞書を介したアラインメント数に比例させる（曖昧性なくアラインメントがとれる関連語ペアが種となる）

$$C_0(x'(i),y(j)) = \begin{cases} \dfrac{a(x'(i),y(j))}{\sum_k a(x'(i),y(k))} & \cdots \sum_k a(x'(i),y(k)) \neq 0 \\ 0 & \cdots \sum_k a(x'(i),y(k)) = 0 \end{cases} \quad [2]$$

$$a(x'(i),y(j)) = \begin{cases} 1 & \cdots (x,x'(i))\text{とアラインメント可能な関連語ペア}\ (y(j),y')\ \text{が存在する} \\ 0 & \cdots (x,x'(i))\text{とアラインメント可能な関連語ペア}\ (y(j),y')\ \text{が存在しない} \end{cases}$$

予備実験－翻訳結果に対する後処理による訳語選択の適用

対象コーパス: 日英新聞記事対応付けデータ (JENAAD) (Utiyama and Isahara, 2003), 150000対訳
　－ ランダムに抽出した300対訳を英日翻訳
　－ 統計ベースシステムでは残りを訓練／開発データとして使用
　－ 関連行列の計算に別のコーパス・訓練コーパスをそれぞれ使用
　　・ 別コーパス: New York Times（2004年）および毎日新聞（2004年）

翻訳システム: Excite翻訳、NEC CROSSROAD、Moses (Koehn et al., 2007)
　－ Moses以外については、特にチューニングを行わない

訳語選択方法: 対象語 x に対して、第i訳語 t_i のスコアを以下の式で定義

$$Score(x,t_i) = \sum_{c:x\text{の周辺語}} \frac{1}{\sqrt{(x\text{と}c\text{の距離})}} C_x(c,t_i)$$

$C_x(c,t_i)$: 対象語xの関連行列中の関連語c・訳語t_iの値

入力文に x、出力文に t_i が含まれており、t_i のスコアが1位でない場合、1位の訳語に置き換える

自動評価結果 (BLEU/NIST)

	Excite	CROSSROAD	Moses
適用前	0.0430/2.643	0.0433/2.563	0.0854/3.799
後処理（別コーパス）	0.0411/2.599	0.0406/2.519	
後処理（訓練コーパス）	0.0429/2.462	0.0433/2.560	

翻訳例

入力文	it also advised amending clause 2 of article 9 in the future while maintaining a peaceful policy .
正解例	そのうえで 、**将来** 的 に は 世界 平和 に 貢献 する べく 、憲法 九 条 二 項 の **改正** が 望ましい と し て いる 。
Moses	また 八月 **改正** 九 条 二 項 で は 今後 、平和 を 維持 し つつ 、だ 。
Excite	また 、それ は 、**将来** 、平和な 政策 を 維持 して いる 間 、Article9 を 訂正 条項 2 に 知らせ ました 。
CROSS-ROAD	平和な 方針 を 維持 する 間 に 、それ は また **未来** に 第 9 条 を **改正** 条項 2 に 通知 し ました 。

後処理（別コーパス）: 将来 → 先物、未来 → 先物
後処理（訓練コーパス）: 改正 → 刑法

おわりに

● 現時点では訳語－関連語関連行列の結果がうまく求められていない
　－ 別コーパスから求めた場合、語彙のずれから逆効果となった
　－ 訓練コーパスから求めた場合、現時点ではほとんどの訳語対が関連行列において一対一になり、効果がない
→訳語－関連語関連行列の改善をまず行うべき
● 翻訳システム中で関連行列を動的に適用する手法を検討したい