統計的手法を用いた日英放送原稿の単語対応づけ

4L－1

熊野 正　田中 英輝　江原 暉将

NHK 放送技術研究所

1. はじめに

我々は現在，放送局の日英翻訳現場での利用を目指して，類似用例提示型翻訳支援システムの開発を進めている．これは日英の放送原稿の対訳コーパスを用い，ユーザの日本語もしくは英語表現によるリクエストに応じて適切な日英表現対を提示するシステムである．

本稿では，記事単位で対応がついた日英記事対を用いて日英単語の共起の度合を t-score を尺度として計算し，これを用いて日英記事対中の表現の対応関係を推定する手法について検討する．

2. 日英単語間の共起度の計算

我々が収集を行っている日英原稿データベースは，日英それぞれのニュース原稿からなっている．英語原稿は基本的には日本語原稿を人手で翻訳して作られているが，翻訳結果は自然な英語のニュース原稿になるように表現や内容が変更されており，元の日本語原稿の構造が保たれているとは限らない [2]．このため，日英の原稿は記事単位での対応関係は容易につけられるが，日英記事対中の各文の間に文対応関係をつけることは一般に困難で，対応の可能性の範囲をしぼっていくことで単語の対応関係を推定する手法（Kay の手法 [1] など）を採用することは難しい．

表 1:　日英記事対の統計的性質
（1 記事あたりの平均値）

	文数	単語数[1]（異なり単語数）
日本語	5.2	275　(130)
英語	7.5	153　(99)

そこで，記事単位で対応のついた日英記事対をそのまま用い，同じ記事対中にともに出てくる日英単語対を共起しているものと見なして共起関係の抽出を行い，有意な共起関係が抽出できるかどうか調査した．実験に用いた日英記事対は，1995 年 3 月から 11 月までの日英原稿の中で 1 対 1 対応のついた全記事対 4,731 対である．記事対の統計的な性質を表 1 に示す．

2.1. t-score の計算

日英単語の共起度を計算する尺度として，t-score を採用した．ここでは t-score を手法 1 のように計算した．この計算手法で，全記事対から抽出した日英単語（日本語: 1,301,130 語（うち異なり語: 23,835 語），英語: 725,605 語（うち異なり語: 25,571 語））間の t-score を計算した結果のうち，t-score 上位のものを表 2 に示す．

ある記事対（記事対番号 i）において，

$$\begin{cases} F_{\mathrm{j}}(i,w_{\mathrm{j}}) &= \text{記事対の日本語記事中に日本語単語 } w_{\mathrm{j}} \text{が出現した回数} \\ F_{\mathrm{e}}(i,w_{\mathrm{e}}) &= \text{記事対の英語記事中に英語単語 } w_{\mathrm{e}} \text{が出現した回数} \end{cases}$$

としたときに，この記事対における w_{j} と w_{e} の共起回数を以下のように定める．

$$F_{\mathrm{je}}(i,\{w_{\mathrm{j}},w_{\mathrm{e}}\}) = \frac{F_{\mathrm{j}}(i,w_{\mathrm{j}})F_{\mathrm{e}}(i,w_{\mathrm{e}})}{\sqrt{\sum_{w_{\mathrm{j}}}F_{\mathrm{j}}(i,w_{\mathrm{j}})\ \sum_{w_{\mathrm{e}}}F_{\mathrm{e}}(i,w_{\mathrm{e}})}}$$

従って，全ての記事対における $w_{\mathrm{j}},w_{\mathrm{e}}$ の出現確率は

$$P_{\mathrm{j}}(w_{\mathrm{j}}) = \frac{\sum_i F_{\mathrm{j}}(i,w_{\mathrm{j}})}{\sum_{w_{\mathrm{j}}}\sum_i F_{\mathrm{j}}(i,w_{\mathrm{j}})},\quad P_{\mathrm{e}}(w_{\mathrm{e}}) = \frac{\sum_i F_{\mathrm{e}}(i,w_{\mathrm{e}})}{\sum_{w_{\mathrm{e}}}\sum_i F_{\mathrm{e}}(i,w_{\mathrm{e}})}$$

となり，この単語対の t-score は以下のように計算される．

$$t(\{w_{\mathrm{j}},w_{\mathrm{e}}\}) = \frac{\sum_i F_{\mathrm{je}}(i,\{w_{\mathrm{j}},w_{\mathrm{e}}\}) - P_{\mathrm{j}}(w_{\mathrm{j}})P_{\mathrm{e}}(w_{\mathrm{e}})\sum_{\{w_{\mathrm{j}},w_{\mathrm{e}}\}}\sum_i F_{\mathrm{je}}(i,\{w_{\mathrm{j}},w_{\mathrm{e}}\})}{\sqrt{\sum_i F_{\mathrm{je}}(i,\{w_{\mathrm{j}},w_{\mathrm{e}}\})}}$$

手法 1:　t-score の計算手法

"Statistical Alignment between Japanese and English News Elements"
KUMANO Tadashi(kumano@strl.nhk.or.jp),
TANAKA Hideki, EHARA Terumasa
NHK Science and Technical Research Laboratories
1-10-11 Kinita, Setagaya-ku, Tokyo, JAPAN 157

[1] 単語数には日英とも記号などは含まれない．
日本語単語は品詞の異なる語は別単語として，英語単語は語形の異なる語は別単語として計算した．
また，我々の使用した日本語形態素解析プログラムの出力は自立語活用語尾を単独品詞として出力するので，ここでの日本語単語数の計算においてもそれに従った．

表 2: 日英単語対の t-score

アメリカ（カタカナ未定義語）	−	U-S	5.22
朝鮮（未定義語）	−	North	4.99
円（名詞）	−	yen	4.75
朝鮮（未定義語）	−	Korea	4.63
円（後置助数詞）	−	yen	4.41
北（名詞）	−	North	4.34
北（名詞）	−	Korea	4.00
核実験（名詞）	−	nuclear	3.98
日本（名詞）	−	Japanese	3.95
ドル（名詞）	−	yen	3.95
大臣（名詞）	−	Minister	3.90
⋮		⋮	⋮

t-score 上位の 100 単語対 (score = 5.22 − 2.40) について単語対応の正解率を調査してみたところ，35%（「朝鮮」− "North" のような複合語の部分にあたる語との対応を正解としても 62%）であった．

結果がよくない理由の 1 つは，日本語単語の方は形態素解析の結果から単語の品詞を特定したり語幹と語尾を区別しているが，英語単語の方は品詞の特定や語形の標準化を行っていないためであると考えられる．しかし英語側でこのような処理を行ったとしても，一方の言語での 1 語が相手言語では複数の語で表現されることもあり，またある語の最も適切な対訳は使用される記事に依存するため，単語単位の正確な対訳を決定することは難しい．

3. 日英表現間の対応関係の推定

我々が目指しているのは日英の文章間に適切な対応をつけることであり，文章中のある表現の相手言語での表現を特定するのに，単語間の正確な対応関係から出発する必要はない．

そこで，縦軸・横軸に日英単語を記事中での出現順序に並べ，各要素に該当単語対の t-score を置いたマトリクス（図 1 に一例を示す）を考えてみる．例えばこの図では，「河野外務大臣」と "Foreign Minister Yohei Kohno" の交わる部分の t-score がまとまって高く，この表現対が対応している可能性が高いことを表している．

このマトリクスを用いて，一方の言語でのある連続した表現が相手言語でどのように表現されているかを推測するには，マトリクス中でこの表現を含んだ「t-score の大きい要素の集まり」を見つけ，この集まりを相手言語の文章中に投影すればよい．例えば「核実験の停止を求める決議案」に対応する英語表現は，この表現に対応する t-score の大きい要素の集まりを考えることによって，"··· anti-nuclear testing resolution is put to the ···" の近辺ではないかと推測することができる．このような集合を認定するアルゴリズムについては今後の検討が必要である．

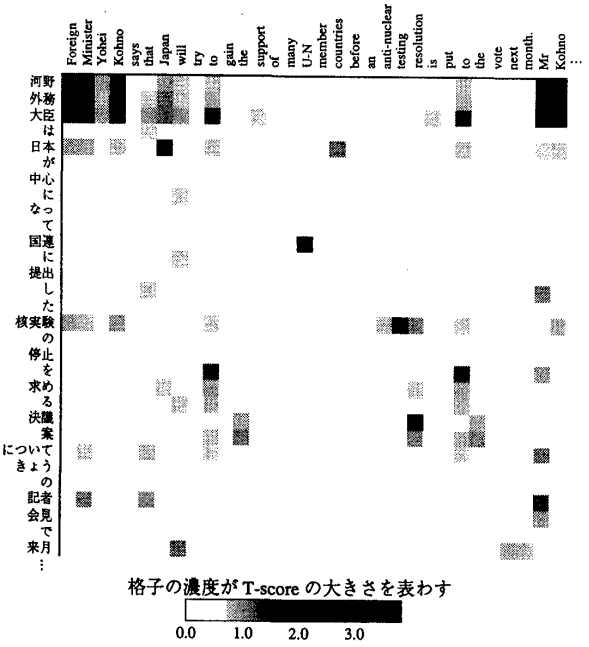

図 1: 記事中の出現単語間の t-score マトリクス

4. まとめ・今後の課題

本稿では，記事単位で対応がついた日英記事対から日英単語の共起の度合を，t-score を尺度として計算した．また，計算された t-score を用いて実際に記事対中の「複数単語からなる表現」の対応関係を推測する方法を検討した．

今後はこの表現対応関係の推測手法がどの程度精度よく実現可能かについて，大規模な実験を行いたい．また，既存の対訳辞書を利用する手法（宇津呂の手法 [3] など）との融合が可能かどうか，精度の向上をもたらすかどうかについても検討を行いたい．

参考文献

[1] Martin Kay and Martin Röscheisen. Text-translation alignment. *Computational Linguistics*, 19(1), pp. 121–142, 1993.

[2] 熊野正, 田中英輝, 金淵培, 浦谷則好. 日英ニュース原稿の対訳コーパス化に関する基礎調査. 言語処理学会年次大会, pp. 41–44, 1996.

[3] Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text, matching using bilingual dictionary and statistics. In *COLING 94*, pp. 1076–1082, 1994.