1T−06　　ノイズを含む対訳コーパスからの対訳表現抽出法

加藤　直人　　江原　暉将
ＮＨＫ放送技術研究所
{katonao, eharate}@strl.nhk.or.jp

1　はじめに

機械翻訳システムを実用するためには，翻訳対象に固有の対訳表現を利用することが必要である．そこで，対訳コーパスから対訳表現を自動的に抽出する研究が盛んに行われている[Fung 97][熊野 97][北村 97][Haruno 98][Melamed 00][宇津呂 95]．これらは文単位，章単位などの範囲における共起情報に基づいている．共起情報は出現頻度が高い単語の抽出には有効であるが，低頻度の単語を抽出することは難しい．我々の対訳コーパス（ニュースの日英記事）でも，日本語を逐語的に翻訳している箇所は少ない，１つの記事はそれほど長くない等により，高頻度の単語は少ない．構文情報を使う方法[Kaji 92]が考えられが，ニュースの構文解析は非常に困難である．

本稿では，簡単な構文情報として単語出現位置を使って対訳表現を抽出する手法について述べる．単語出現位置は単語の並びを表すための値である．本手法は低頻度の単語にも対応することができるという特徴がある．

2　ニュースの対訳原稿

我々の対訳コーパス（ニュースの日英記事）の一例を図１に示す．日本語原稿はニュース用に記者が作成したものであり，英語原稿は翻訳者が日本語原稿を元にして作成したものである．英語原稿を作成する際には，逐語的に翻訳こともあるが，英語側で不要な表現は削除したり，他の日本語原稿を参照して新たに書き加えたりすることもある．したがって，日本語原稿と英語原稿では一致しない部分も多く，単語対応という観点からみるとノイズが多い対訳コーパスといえる．

3　単語出現位置を制約とした対訳表現抽出

本手法では英語原稿における各単語の日本語訳語を，日本語原稿と照合することにより対訳表現を抽出する．照合する際に構文情報として日本語原稿から得られる単語出現位置を使う．しかしながら，Fungらの手法と異なり，単語出現位置を，テキスト全体の特徴を表す共起情報として使うのではなく，翻訳のまとまりをとらえるための構文情報として使っている．

本手法は次の３つのステップからなる．

（１）英語原稿の訳語候補作成．
（２）日本語原稿における最適単語照合．
（３）対訳表現抽出．

以下ではそれぞれのステップについて説明する．

3．1　英語原稿の訳語候補作成

始めに英語原稿を形態素解析し，得られた各品詞に対する日本語訳語を求める．この際には，既存の我々の英日機械翻訳のモジュールである，英語形態素解析と英日対訳辞書を使っている．通常，英単語は複数の訳語候補がある．

Title¦政府・トルコ地震で専門家派遣へ
Date¦1999年09月03日12時14分
J1¦政府は、トルコ西部で起きた大地震の２次災害を防ぐため、被災した建物が余震で崩れ落ちる恐れがないかなどを診断する専門家のチームを、あさって（５日）から現地に派遣し、トルコ政府の関係者などに技術指導を行なうことになりました。
J2¦今回の派遣は、日本政府がトルコに対する緊急支援の一環として行なうもので、建設省の研究所や大阪府と兵庫県の職員、それに大学の研究者の、いずれも建物の耐震性についての専門家７人が、あさって（５日）からおよそ、２週間の予定でトルコに派遣されます。
J3¦一行は、現地で、トルコ政府や地元自治体などの職員に対して、耐震性の判断基準など、余震で被災した建物が倒壊する恐れがないかどうかを診断する技術を指導することにしています。
J4¦これによって、余震による２次災害を防ぐとともに、建物の危険度を個別に区分することで、被災者のための仮設住宅の数を見積もる作業も進めたいとしています。

日本語原稿

Title¦Japan to send quake-proof construction experts to Turkey.
Date¦1999/09/03 13:30
E1¦The Japanese government will send specialists to quake-stricken Turkey to help judge the resistance of damaged buildings to aftershocks.
E2¦The team of 7 (seven) specialists in quake-resistant building techniques will leave for Turkey on Sunday for a two-week stay.
E3¦The seven members are from Osaka and Hyogo prefectures, the construction ministry and universities.
E4¦They will give technical instructions to officials of the Turkish government and local offices to decide which buildings are susceptible to aftershocks.
E5¦The team will also estimate the number of temporary houses needed for victims.

英語原稿

図１　対訳コーパス（日英ニュース記事）

An automatic method for extracting translation patterns from noisy parallel corpora.
Naoto Katoh and Terumasa Ehara
NHK Science and Technical Research Laboratories

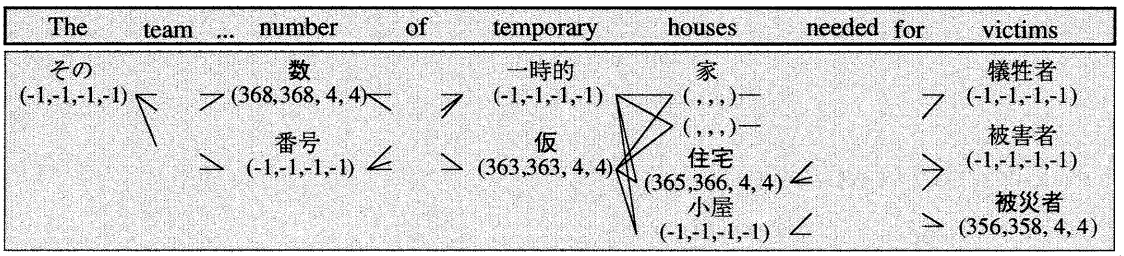

図２　英文原稿の訳語と日本語原稿との対応

３．２　日本語原稿における最適単語照合

次に日本語訳語候補を日本語原稿と照合する．この際に，先に述べたように訳語は複数の候補がある他に，日本語原稿と照合する際に複数箇所で一致する場合がある．したがって，最適な照合を求める必要がある．本手法では単語出現位置を使って次のようにして求める．まず，i番目の英単語の訳語（w_i）と$i+1$番目の英単語の訳語（w_{i+1}）との間の距離を４つの値（文字出現開始位置 ch^s，文字出現終了位置 ch^e，節番号 cl，文番号 st）に基づいて，文出現距離，文出現距離の和として次式のように定義した．

【訳語間距離】

$$distPosit(w_i, w_{i+1}) = \lambda_1 distCharPosit(w_i, w_{i+1}) + \lambda_2 distSentPosit(w_i, w_{i+1}) \quad (1)$$

（λ_1，λ_2（≥ 0）は定数，$\lambda_1 + \lambda_2 = 1$）

【文字出現距離】

$$distCharPosit(w_i, w_{i+1}) = \begin{cases} log(ch^s_{i+1} - ch^e_i) & if\ ch^s_{i+1} > ch^e_i \\ log|ch^s_{i+1} - ch^e_i| \times penalty & otherwise \end{cases} \quad (2)$$

【文出現距離】

$$distSentPosit(w_i, w_{i+1}) = \begin{cases} SentPosit(w_{i+1}) - SentPosit(w_i) & if\ SentPosit(w_{i+1}) \geq SentPosit(w_i) \\ (SentPosit(w_i) - SentPosit(w_{i+1})) \times penalty & otherwise \end{cases} \quad (3)$$

ただし，

$$SentPosit(w) = st + \frac{cl - 1}{clmax(w)} \quad (4)$$

$clmax(w)$：訳語 w が出現した文における節番号の最大値

（$penalty$（≥ 1）は定数）

ここで，文字出現開始（終了）位置は，日本語原稿の先頭の文字から順に番号をつけ，訳語が日本語原稿中で照合できた際，その文字の最初（最後）の番号である．節（文）番号は，その訳語が含まれる節（文）の番号である．訳語間距離では正順に並ぶ訳語や，同じ文や節にある訳語を優先するように定義している．すると，例えば，日本語で名詞連続となる表現は英語側でも名詞連続となることが多いが，このような構文情報を単語出現位置が連続するという特徴で表すことができる．

最適な単語対応は英文原稿における最初の訳語から最後の訳語までの距離の和が最小となるパスを動的計画法で求めればよい．

３．３　対訳表現抽出

対訳表現は最適な訳語対応からある評価関数に基づいて求める．例えば，評価関数として，訳語と日本語原稿との間で両側で文字が一致し，その間に挟まれるの１文字のみは不一致でもよいと定義する．すると，図２において，“temporary”の訳語「仮」と“houses”の訳語「住宅」は日本語原稿で一致し，その間には１文字しか不一致がないので，“temporary houses”の訳語として「仮設住宅」が得られる．

４　おわりに

単語出現位置を利用することにより，ノイズを含む日英対訳コーパスから，対訳表現を抽出する手法について述べた．今後は対訳表現抽出の評価関数をさまざま変え，大量の日英ニュース記事に対して対訳表現抽出の評価実験を行う予定である．また，本手法を使って文アライメントも行いたい．

参考文献

[Fung 97]Pascale Fung and Kathleen McKeown,“A Technical Word- and Term-Translation Aid Using Noisy Parallel corpora across Language Groups,” Machine Translation, Vol.12, No.1, pp.53-87, 1997.

[Kaji 92]Hiroyuki Kaji et al,“Learning Translation Templates from Bilingual Text,” Proc. COLING92, pp.672-678, 1992.

[熊野 97]熊野　明，平川秀樹，“対訳文書からの機械翻訳専門用語辞書作成，” 情報処理学会論文誌，Vol.35, No.11, pp.2283-2290, 1994.

[北村 97]北村美穂子，松本裕治，“対訳コーパスを利用した対訳表現の自動抽出，” 情報処理学会論文誌，Vol.38, No.4, pp.727-736, 1997.

[Haruno 98]Masahiko Haruno and Satoru Ikehara,“Two step Extraction of Billingual Collocations by Using Word-Level Sorting,” IEICE Trans Inf&Syst, Vol.E81-D, No.10, pp.1103-1110, 1998.

[Melamed 00]I.D. Melamed“Models of Translational Equivalence among Words,” Computational Linguistics, Vol.26, No.2, pp.221-249, 2000.

[宇津呂 95]宇津呂武仁，松本裕治，“対訳辞書および統計情報を用いた二言語対訳テキスト照合，” コンピュータソフトウェア，Vol.12, No.5, pp.12-21, 1995.