# 日本語・英語ニュースを対象とした 文・単語の同時アライメント

加藤　直人
ＮＨＫ放送技術研究所
katou.n-ga@nhk.or.jp

## 1　はじめに

ＮＨＫの多言語ニュースコーパスは、２つの言語間で記事という単位ではアライメントされているものの、単語はもとより文の単位でもアライメントされていない。翻訳メモリや機械翻訳の研究をする上では、しかしながら、単語・文単位でもアライメントされているほうが望ましい。

本稿では、日本語と英語ニュースを対象として、自動的に文と単語のアライメントをする手法について述べる。提案手法は文アライメントの信頼度（文対応度）と単語アライメントの信頼度（単語対応度）に基づいており、さらにこの２つの信頼度が相互依存の関係にある。そして、Pantel らのEspresso[1]のように、それぞれの信頼度を繰り返し更新することによって、双方の精度が向上することを期待している。本稿では、また、提案手法の評価の第一歩として行った文アライメントについて、その実験結果を報告する。

## 2　日本語・英語対訳ニュース

ＮＨＫの日英対訳ニュースの例を図１に示す。ＮＨＫの日英対訳ニュースの詳細な分析に関しては文献[2]に譲り、ここでは簡単にその特徴について述べる。

日本語ニュースは記者が書いた原稿を元に、アナウンサーが読む原稿に書き換えられる。読んで伝えることを目的として書かれているため、新聞記事と比較すると冗長な表現も少なくない。

一方、英語ニュースは日本語ニュースを元に作成される。その際、１つの日本語ニュースを参考にすることが多いが、事態の進展とともに、その詳細が書かれた複数の日本語ニュースを参考にすることもある。英語ニュースの作成は、日本語ニュースを土台にニュースライターがその内容や背景などを理解した上で行われており、純粋な翻訳作業とは異なる。また、あまり重要でない情報は盛り込まず、初めてそのニュースを聞く外国人にもわかるように補足説明を入れることもある。したがって、日英対訳ニュースは記事全体としては同じ内容であるが、その表現方法は大きく異なる場合も多い。したがって、文や単語のアライメントが難しい場合もある。

図１の文アライメントを考えてみよう。次のような５つの文アライメントになると考えられる。

① J1 ⇔ E1
② J2 ⇔ φ
③ J3 ⇔ E2
④ J4 ⇔ E3,E4
⑤ J5 ⇔ φ

（φは対応する文がないことを表す）

この文アライメントからわかる特徴としては、日本語文に対応する英語文がないことがある（②、⑤）、日本語文１文に対して英語文が複数文対応することがある（④）、文アライメントは交差しない[*1]などがあげられる。

次に単語や節（句）という小さい単位で見てみよう。文アライメント①を見ると、「インドネシアのスマトラ島沖の巨大地震と津波を教訓に…被害を防ぐため」という長い日本語節が、英語では“in responding to earthquakes and tsunamis”と、かなり簡略化されている。具体的には、日本語にある「インドネシアのスマトラ島沖の巨大」や「教訓」などの情報が英語側では省略されている。逆に、英語側では“Japan's NHK”と、NHKに“Japan”という情報が補足されている。情報の補足に関しては他に、文アライメント③を見ると、「来月」と相対的な時制表現が、英語側では“Feburary”と絶対的な表現となっていることがわかる。このように、単語アライメントは文アライメントに比べてさらに複雑である。

*1 実際には（他の例では）、少数ではあるが、文アライメントが交差している場合もある。

T| アジアの放送局が津波情報の会議
D|2005/01/06 18:11

| | |
|---|---|
| J1 インドネシアのスマトラ島沖の巨大地震と津波を教訓にアジア各国の放送局が被害を防ぐためにどのような放送をすべきかを話し合う会議が来月、東京で開かれることになりました。<br>J2 今回の津波の被害を受けて国連などはインド洋沿岸地域を対象に津波警報システムを設ける方向で議論を進めています。<br>J3 これを受けて、NHKとABU（エイビーユー）＝アジア太平洋放送連合は津波の被害を防ぐためにどのような放送をすべきかを話し合う会議を来月二十八日から三日間、東京・渋谷のNHK放送センターで開くことになりました。<br>J4 会議にはインドネシアやタイ、スリランカなど今回の津波の被災国を含むアジアの十三の国の放送局や津波の専門家などが参加し、▼各国の放送局が今回の災害をどのように伝えたか検証したうえで、▼津波情報を迅速に伝えるにはどうすればいいのか議論することにしていて、NHKは会議を通じて災害時のニュース速報などのノウハウを提供することにしています。<br>J5 また、最終日には東海地震に備えて津波対策を進めている静岡県を視察することになっています。 | E1 Broadcasters from around Asia, including Japan's NHK, plan to discuss how their broadcasts can be more effective in responding to earthquakes and tsunamis.<br>E2 The Asia-Pacific Broadcasting Union, or ABU, will hold a three-day conference from February 28th at NHK's headquarters in Tokyo.<br>E3 Officials from broadcasters in 13 ABU member countries, including Indonesia, Thailand and Sri Lanka, will attend the conference, along with tsunamiexperts.<br>E4 They will examine the way in which reports about the recent quake and tsunamis were broadcast.<br>E5 They will also look at ways to ensure that tsunami information can be broadcast as quickly as |

図１ ＮＨＫの日英対訳ニュースの例

## ３ 文と単語の同時アライメント

### 3.1 Espresso

提案手法を説明する前に、まずEspressoについて簡単に説明する。

Espressoは意味関係抽出アルゴリズムである。“種”となる初期のインスタンスを人手で与え、コーパスからその意味関係をもつパターンを抽出するとともにインスタンスの拡張を行う。

例えば、初期のインスタンスとして

(I1) “Leonardo da Vinci” + “Mona Lisa”

を人手で与え、コーパスに

(S1) Leonardo da Vinci painted Mona Lisa
(S2) Leonardo da Vinci is the painter of Mona Lisa
(S3) Vincent van Gogh is the painter of Sunflower

という文が含まれるものとする。

インスタンス(I1)はコーパス中の(S1)と(S2)に共通するので、Espressoはパターンとして

(P1) “painted”と“is the painter of”

を抽出する。次に、得られたパターン(P1)から、今度は同じパターンを持つ(S3)により、新たなインスタンスとして

(I2) “Vincent van Gogh” + “Sunflower”

を抽出する。さらに、抽出されたインスタンスからまた新たなパターンを抽出する。このような処理を繰り返すことにより、Espressoは新たなパターンとインスタンスを抽出する。

パターンやインスタンスを抽出する際には、必ずしも正しいものばかりが得られるわけではない。そこで、Espressoでは信頼度を定義し、その値が大きいパターンとインスタンスを抽出している。信頼度は、パターンとインスタンスがそれぞれ相互依存にある関係として、次式のように定義されている。

$$r_{\pi}(p) = \frac{\sum_{i \in I} \left( \frac{pmi(i,p)}{\max_{pmi}} \times r_{\iota}(i) \right)}{|I|} \quad (1)$$

$$r_{\iota}(i) = \frac{\sum_{p \in P} \left( \frac{pmi(i,p)}{\max_{pmi}} \times r_{\pi}(p) \right)}{|P|} \quad (2)$$

式(1),(2)では、パターンの信頼度が向上すればインスタンスの信頼度が向上し、逆にインスタンスの信頼度が向上すればパターンの信頼度が向上するようになっている。Espressoでは、あらたなパターンやインスタンスが抽出されたとき、これらの信頼度を更新している。

## 3.2 提案手法

提案手法は基本的にはEspressoと同じであり、Espressoのパターンが文アライメントに、インスタンスが単語アライメントに対応している。

### 3.2.1 対訳辞書

Espressoでは“種”となるインスタンスを人手で与えているが、提案手法では“種”となる単語アライメントを対訳辞書で与えた。対訳辞書による単語アライメントは人手を介していないので、必ずしも正解ばかりではない。

対訳辞書にはＥＤＲの日英対訳辞書、英日機械翻訳で利用していた辞書 [4] を用いた。

さらに、対訳辞書処理の一環として、対訳辞書にあらかじめ登録しておくことが困難である固有名詞や数字の処理を行っている。固有名詞処理では、日本語形態素結果から得られる読みをローマ字に変換して英語文中の単語と照合し、類似度が高い場合は対訳であるとしている。数字処理では、ルールによって対訳を生成しており、例えば、「九千二百二十五」に対しては、英訳として“9,225”，“nine thousand two hundred twenty five”，“9.225-thousand”，“nine-thousand 225”等を自動生成している。

### 3.2.2 文アライメント

文アライメントはその信頼度（文対応度）に基づいて行う。文対応度 $r_s(s_i^J, s_i^E)$ は、順序不変度 $r_s^1(s_i^J, s_i^E)$ と単語一致度 $r_s^2(s_i^J, s_i^E)$ という２つの積で、式 (3) のように定義した。

$$(3)\quad r_s(s_i^J, s_i^E) = r_s^1(s_i^J, s_i^E) \times r_s^2(s_i^J, s_i^E)$$

ここで、順序不変度は、

$$(4)\quad r_s^1(s_i^J, s_i^E) = \exp(-\alpha \mid i - j \mid)$$

（$\alpha$は定数）

単語一致度は、

$$(5)\quad r_s^2\left(s_i^J, s_j^E\right) = \frac{\sum_k R_w\left(w_k^J, w_l^E\right)}{\left|s_j^E\right|}$$

$$R_w\left(w_k^J, w_l^E\right) = \frac{1}{1+\exp\left[-\beta\left(r_w\left(w_k^J, w_l^E\right) - r_0\right)\right]}$$

$$w_k^J \in s_i^J, w_l^E \in s_j^E$$

（$\beta$ , $r_0$ は定数）

順序一致度では、文アライメントは上から順にとれる場合が多いという特徴を表現している。一方、単語一致度とは、文中で対訳となる単語が含まれる割合である。対訳となるか否かは、対訳辞書とともに、次に説明する単語アライメントの結果も使う。また、文アライメントを行う際には、英語文１文に対して日本語文（文対応度が最大のもの）を１文選択するものとした。これは、ＮＨＫの日英対訳ニュースでは英語のほうが一文あたりの単語数が少ないため、英語文から日本語文を１つ決めることのほうが容易なことによる。さらに、文対応度が小さいときは、対応する日本語文がないとした。

### 3.2.3 単語アライメント

単語アライメントは、対訳辞書にない場合には単語アライメントの信頼度（単語対応度）に基づいて行う。単語対応度 $r_w(w_k^J, w_l^E)$ は、低頻度ペナルティ $r_w^1(w_k^J, w_l^E)$ と単語共起度 $r_w^2(w_k^J, w_l^E)$ という２つの積で、式 (6) のように定義した。

$$(6)\quad r_w(w_k^J, w_l^E) = r_w^1(w_k^J, w_l^E) \times r_w^2(w_k^J, w_l^E)$$

ここで、低頻度ペナルティは、式 (7) で定義した。

$$(7)\quad r_w^1(w_i^J, w_i^E) = 1\text{-}\exp[\,-\gamma\,(freq(w_k^J, w_l^E) - f_0)$$

（$\gamma$ , $f_0$ は定数）

低頻度ペナルティは、対訳共起 $freq(w_k^J, w_l^E)$ が低頻度のときには、次の単語共起度が大きくなり過ぎるという傾向を緩和するために導入した。単語共起度は、文アライメントを前提にしている、一般的な単語共起の指標 [5] を使用した。例えば、対数尤度比、カイ２乗値、相互情報量などがあるが、後述する実験では対数尤度比を使用した。ただし、提案手法では共起頻度の代わりに文対応度を用いている。すなわち、単語共起度は式 (8) のように定義した。

$$\begin{aligned}(8)\quad r_w^2\left(w_k^J, w_l^E\right) = {} & l(a) + l(b) + l(c) + l(d) \\ & - l(a+b) - l(a+c) - l(b+d) \\ & + l(a+b+c+d)\end{aligned}$$

ここで、

$$l(x) = x \log x$$

$$a = freq\left(w_k^J, w_l^E\right) = \sum_{i,j} r_s\left(s_i^J, s_j^E\right) \delta\left(w_k^J \in s_i^J, w_l^E \in s_j^E\right)$$

$$b = freq\left(w_k^J, \neg w_l^E\right) = \sum_{i,j} r_s\left(s_i^J, s_j^E\right) \delta\left(w_k^J \in s_i^J, w_l^E \notin s_j^E\right)$$

$$c = freq\left(\neg w_k^J, w_l^E\right) = \sum_{i,j} r_s\left(s_i^J, s_j^E\right)\delta\left(w_k^J \notin s_i^J, w_l^E \in s_j^E\right)$$

$$d = freq\left(\neg w_k^J, \neg w_l^E\right) = \sum_{i,j} r_s\left(s_i^J, s_j^E\right)\delta\left(w_k^J \notin s_i^J, w_l^E \notin s_j^E\right)$$

提案手法では、式 (5) と式 (8) で信頼度が依存関係にある。

## ４　実験

提案手法の評価の第１ステップとして、文アライメントの評価実験を行った。日英対訳ニュースには 1,098 記事を用い、評価データには人手で文アライメントをつけた 20 記事 *2 を用いた。

図２に文アライメントの評価結果を示す。ここで、繰り返し回数 0 回というのが対訳辞書のみによる結果である。図２を見ると、繰り返し処理により、文アライメントの精度が向上したのがわかる。しかし、繰り返しを続けると精度が上下しており、安定していない。

図３に、はじめに得られた（＝繰り返し回数１回目で使われた）単語アライメントの結果を示す。これらはもちろん対訳辞書に未登録な対訳である。単語アライメントは１対１の単語同士で行っているため、「自衛隊， Self-Defense」ように単語対応が部分的なものとなっているものもある。「きょう， on」の単語アライメントは、2 章で述べた時制表現の変換による。すなわち、ＮＨＫの英語ニュースでは「きょう」は、“on Monday”のように‘曜日’で翻訳されるためである。

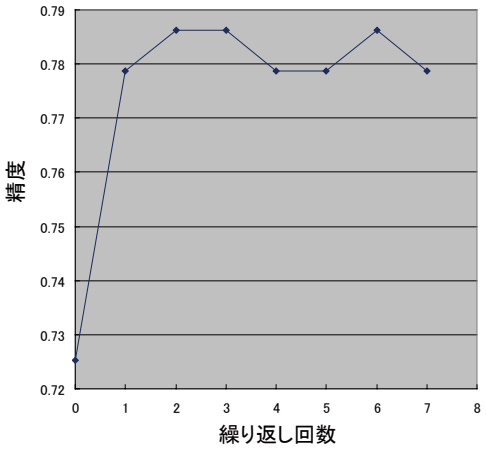


図２ 文アライメントによる評価結果

*2 今回の実験では、１つの英語文に対して２つ以上の日本語文が対応する文が 2 つあった。したがって、精度は 100%にはならない。

## ５　おわりに

ＮＨＫの日英対訳ニュースを対象にして、文と単語のアライメントを行う手法について述べた。また、文アライメントの評価実験を行い、その有効性を確認した。今後は、まず単語アライメントの評価する必要がある。また、提案手法では複数のテキストを用いて信頼度を計算しているが、単一のテキストを用いた手法 [6] との比較も行いたい。

## 参考文献

[1]Patrick Pantel et al. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. Procs. of Coling-ACL06, pp.113-120, 2006.

[2] 荒牧英治ほか. 用例ベース翻訳のための日英アライメント確信度と日本語類似度を用いた訳語選択. 自然言語処理, Vol.11, No.1, pp.107-124, 2004.

[3]NHK「ニュース 7」「ニュース 9」の二カ国語放送制作現場を拝見. 翻訳辞典 2000 年度版, アルク地球人ムック, pp.67-74, 2000.

[4] 畑田のぶ子ほか. 衛星放送・英日機械翻訳システムの辞書整備. 第 46 回情報処理学会全国大会, pp.157-158, 1993.

[5]Yuji Matsumoto et al. Lexical Knowledge Acquisition. Handbook of Natural Language Processing (Marcell Dekker, Inc.), 2000.

[6] 春野雅彦. 辞書と統計を用いた対訳アライメント. 情報処理学会論文誌, Vol.38, No.4, pp.719-726, 1997.

| 単語対応度 | 日本語，英語 |
|---|---|
| 0.459640 | 自衛隊，Self-Defense |
| 0.420947 | さん，Mr |
| 0.397275 | きょう，on |
| 0.383139 | さん，Ms |
| 0.371515 | 労働省，ministry |
| 0.354764 | ディスカバリー，Discovery |
| 0.346939 | 宿営，camp |
| 0.318483 | 行方，miss |
| 0.312190 | や，and |
| 0.302951 | 警視庁，police |
| 0.302057 | 交通省，ministry |
| 0.301101 | センチ，centimeter |
| 0.296409 | 羽田空港，Haneda |

図３ 自動単語アライメントで得られた例