

Internal Use Only (非公開)

TR-SLT-0026

NHK ニュースのフレーズアライメント

Phrase Alignment of NHK News Corpus

荒牧英治
Eiji Aramaki

田中英輝
Hideki Tanaka

2002 年 9 月 30 日

概要

NHK ニュースコーパスの句間のアラメントを推定する研究を行った。NHK ニュースコーパスのフレーズ間のアラメントを推定する研究を行った。NHK ニュースコーパスは日英両言語で作成されているが、記事対応が付与されているだけで、文同士の対応は取られていない。そこで、まず文アライメントを行い、次に得られた文対応に対してフレーズアライメントを行った。その結果と考察を述べる。

(株) 国際電気通信基礎技術研究所
音声言語コミュニケーション研究所

〒619-0288 「けいはんな学研都市」 光台二丁目 2 番地 2 TEL : 0774-95-1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288, Japan
Telephone: +81-774-95-1301
Fax : +81-774-95-1308

©2002 (株) 国際電気通信基礎技術研究所

©2002 Advanced Telecommunication Research Institute International

目次

1	はじめに	1
2	文アライメント	2
2.1	文アライメントの先行研究	2
2.2	文アライメント手法	2
2.3	実験結果と考察	3
3	句アライメント	6
3.1	句アライメント手法	6
(3.1.1)	基本句を単位とした依存構造への変換	6
(3.1.2)	辞書の情報による対応関係の推定	7
(3.1.3)	未対応句の処理規則	8
3.2	実験結果と考察	8
4	まとめ	10
	参考文献	11

NATO・北大西洋条約機構は三日、日本時間のきょう夕方からベルリンで外相会議を開き、ヨーロッパでの平和維持活動などにあたる新しい部隊の創設をめざす議論に入りました。会議にはアメリカやフランス、ドイツなど加盟十六カ国の外相が出席しており、冒頭NATOのソラーナ事務総長は「新しい役割に向けNATOは今回の会議で重要な一歩を踏み出すことになる」と述べました。NATOは東西冷戦の終結を受けて、戦略転換をすすめており、今回の会議ではヨーロッパの民族紛争や地域紛争に機動的に対処する共同統合任務部隊の創設を決める見通しです。この部隊はアメリカの地上軍が参加しない場合でも、西ヨーロッパ各国で作るWEU・西欧同盟が作戦活動を行うことになっており今年の末、現在、ボスニアに展開している平和実施軍からアメリカ軍が撤退したあとその活動を引き継ぐことも予想されます。また、新たな共同統合任務部隊では一部とはいえヨーロッパ側が主導権をとることになるためこれまで、アメリカに反発してNATOの軍事機構から脱退していたフランスも、参加の意向を表明しています。NATOの外相会議は四日にはロシアや東ヨーロッパ諸国の外相を交え、NATOの拡大問題などについても討議することになっています。

Foreign ministers of 16 NATO nations are meeting in Berlin to discuss ways to make the Atlantic military alliance more adaptable to the post-Cold War era. Discussion will focus mainly on the creation of a small, solely European task force to deal with regional conflicts and peace-keeping activities in Europe. At the opening of the meeting on Monday, NATO Secretary General Javier Solana said an important step toward a new NATO will be taken at this meeting. Earlier on Monday, the U-S and European allies agreed to create a specifically European entity within the North Atlantic Treaty Organization. The new body would come under the control of the Western European Union or W-E-U, Europe's only defence set-up. The new task force would take over the mission of the U-S Forces set to withdraw later this year from the Implementation Force stationed in Bosnia. With the task force expected to increase European initiative in NATO, France has announced it will rejoin the organization it had left in protest of the strong U-S presence. On Tuesday, foreign ministers from Russia and eastern European nations will join NATO foreign ministers to discuss issues such as the expansion of NATO.

表 1: NHK ニュース記事の例

1 はじめに

NHK ニュースコーパスの句間のアラメントを推定する研究を行った。本稿ではその結果と考察を述べる。NHK ニュースコーパスは日英両言語で作成されており、記事対応が付与されている対訳コーパスである。表 1 に記事の例を挙げる。NHK ニュースコーパスは、このような 40083 記事ペアからなり、これらを翻訳知識として利用するために句アライメントを推定することが目標である。

本稿の構成は以下のとおりである。第 2 章で文アライメント処理について述べる。第 3 章で句アライメント手法について述べる。第 4 章にまとめを述べる。

手法	言語ペア	手法
Brown et al.[1]	英 - 仏	# of words
Gale and Church [4]	英 - 仏	# of characters
Wu(1994)	英 - 広東	# of characters
Key and Roscheisen [6]	英 - 仏 - 独	lexical
Chen(1993)	英 - 仏	lexical
Utsuro [9]	英 - 日	lexical
Haruno and Yamazaki [5]	英 - 日	lexical

表 2: 文アライメントの主な手法

2 文アライメント

2.1 文アライメントの先行研究

文対応推定には数多くの先行研究が存在する．まず，どのような文対応手法が NHK ニュースコーパスに適しているか調査した．

主要な先行研究とその手法を表 2 にまとめた．これまでに行われた文対応付けの研究は大きく 2 つに分類できる．1 つは文に含まれる単語数 [1] や文字数 [4] の情報に基づいた length-based と分類される手法である．この手法は，英語 - フランス語等の類似した言語間ではある程度の精度が報告されているが [4]，日本語 - 英語間のように，構造が大きくことなる言語間に用いられた例は少なく，NHK ニュースコーパスに対しては適切でない．

もう 1 つの手法は，lexical-method と分類される手法で，語彙の情報を用いるものである．語彙の情報としては，翻訳辞書の情報や，統計量による語の対応情報などが用いられる．本研究では，翻訳辞書を用いた文献 [9] とほぼ同じ手法を用いた．

2.2 文アライメント手法

文アライメント手法は，文献 [9] とほぼ同じ手法を採用しており，本稿では文献 [9] との違いについて述べる．

まず，本研究ではもとめる文間の対応の種類は，(日本語文数: 英語文数)=(1:1), (1:2), (1,3), (2,1), (2:2) の 5 種類を扱う．これは，人手で文対応を付与した 96 対訳記事を利用して調べてみた結果で，これら 5 種類の文対応で 80 % 以上を占めたからである (表 3)．

また，文対応スコアを両言語の文内に含まれる内容語が翻訳辞書で対応した割合として，次のように定める．

$$\text{文対応スコア} = \frac{W_d \times 100}{W_j + W_e}$$

日本語文数: 英語文数	文対応数
1 文:1 文	274 (39 %)
1 文:2 文	171 (24 %)
1 文:3 文	54 (8 %)
2 文:1 文	48 (7 %)
2 文:2 文	39 (6 %)
1 文:4 文	22 (3 %)
2 文:3 文	16 (2 %)
2 文:4 文	15 (2 %)
その他	66 (9 %)

表 3: 人手による文対応 (人手 A) の種類別の数

ただし, W_j は, 候補内の日本語の内容語の数, W_e は, 候補内の英語の内容語の数, W_d は候補内の辞書で対応がついた両言語の内容語の数である. また, この文対応スコアは, (2 文:2 文) の対応スコアは, これをばらして (1 文:1 文) + (1 文:1 文) の 2 つの文対応に分けた場合のスコアと等しくなる. よって, (2 文:2 文) の文対応スコアを 0.9 倍するペナルティを設ける. 利用した翻訳辞書は, EDR 日英対訳辞書, EDICT(一般的な日英対訳辞書), ENAMDICT(固有名詞の日英対訳辞書), アンカー日英対訳辞書, 英辞郎を用いた.

2.3 実験結果と考察

本手法で 96 記事の文対応付けを行った. この精度を調べるために, 文対応の正解データを 2 人によって作成した. これらを以降, 人手 A, 人手 B とよぶ. 評価は, 再現率, 適合率で行い, 人手 A と人手 B 同士も評価した (表 4). ただし, (1 文:n 文) の対応からは n 個の対応が得られるものとして計算した. 例えば, 文 J_1 と文 E_1, E_2, E_3 が対応している (1 文:3 文) の文対応では, 得られる対は $(J_1, E_1), (J_1, E_2), (J_1, E_3)$ の 3 個である.

$$\text{再現率} = \frac{\text{推定した文対応の中で正しい数}}{\text{正しい文対応の数}}$$

$$\text{適合率} = \frac{\text{推定した文対応の中で正しい数}}{\text{推定した文対応の数}}$$

結果は表 4 のようになった. 正解同士の適合率, 再現率は高く, 記事の文対応は可能であると考えられる. 一方, 本手法によって得られる文対応の適合率は 60 % 程度であり, 十分な精度とは言えない. 精度を下げる原因は次の 2 通りに分類される.

1. 実際の文対応が線状となっておらず, DP マッチングでは捕らえられない.

出力	正解	適合率 (%)	再現率 (%)
人手 A	人手 B	92.0	83.8
人手 B	人手 A	83.8	92.0
本手法	人手 A	60.7	65.8
本手法	人手 B	62.0	61.1

表 4: 文対応精度

2. 記事の後半部分など、対応する部分がない箇所を文対応として採用する。

これらのいずれの原因の場合でも、文対応スコアは低くなる。よって、文対応スコアを手がかりとして、文対応の高精度な部分とそうでない部分をフィルタリングすることが考えられる。フィルタリングに用いることができる指標として次の2つを調べた。

1. 文対応の種類

2. 文対応スコア

まず、(1)を調べた。実験では96記事から377文対応を抽出したが、このうち1文:1文の対応は145対応であり、半分弱を占める。この1文:1文の文対応にしばった場合の適合率は77.5%となり、文対応の全体の適合率よりも約15ポイント向上する。一方、他の文対応の種類でフィルタリングしても有意な精度の差異はない。よって、対応サイズを1:1に絞るフィルタリングは妥当である。

次に、(2)を調べた。文対応スコア別に文対応の精度と数を調べると表5のようになる。表中の括弧内は1:1の文対応だけに絞った場合である。文対応スコアが高くなるほど、文対応の精度は高くなっている。よって、文対応スコアによるフィルタリングは妥当であると考えた。

以上のような結果から、句アライメント処理では1:1対応に絞って、文対応スコア別に行った。

文対応スコア	対応の数	適合率 (%)
0 ~ 10	44(22)	19.0(22.7)
10 ~ 20	52(26)	45.9(61.5)
20 ~ 30	87(39)	72.8(91.0)
30 ~ 40	83(30)	78.7(93.3)
40 ~ 50	57(19)	83.3(100)
50 ~ 60	30(7)	81.2(100)
60 ~ 70	14(1)	75.0(100)
70 ~ 80	4(1)	87.5(100)
80 ~ 90	5(0)	54.8(-)
90 ~ 100	0(0)	- (-)

表 5: 文対応スコア別の対応の数と適合率 (括弧内は 1 文 : 1 文の文対応)

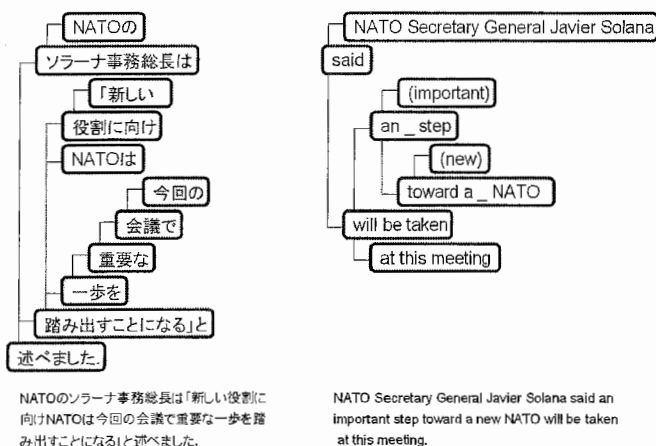


図 1: 基本句を単位とした依存構造

3 句アライメント

3.1 句アライメント手法

句アライメント手法は、文アライメントに比べ先行研究は少ない。本稿では、文献 [3] をもとにした手法を用いた。本手法の特徴は、基本句という日本語の文節に相当する単位を導入することと、基本句を単位とした依存構造を用いて構造的なアライメントを行う点である。

(3.1.1) 基本句を単位とした依存構造への変換

本手法では、対応を扱う単位として、基本句という単位を考える。ここで、基本句とは内容語を中心として機能語をまとめたもので、複合名詞や、名詞と前・後置詞、主動詞と助動詞などをひとまとまりとしたものである。この基本句を単にとした依存構造は次のようにして得る。

まず、対訳文中の両言語の文を統語解析し、その結果を基本句を単位とした依存構造に変換する。これは日英それぞれ次のように行なう。

日本語の文については、KNP[7]を用いて統語解析を行なう(図1)。その結果得られる「(接頭辞*)(内容語+)(機能語*)」という構造を持つ文節を基本句とする(*は0回以上の繰り返し、+は1回以上の繰り返し)。ただし、「(～に)ついて」や「(～に)おいて」など、機能的表現となっている内容語は、機能語とみなして直前の文節にまとめる。

英語の文については、Charniakの統語解析システム[2]を用いて解析を行なう。これは句構造を出力するので、各句構造規則に主辞を定義することにより依存構造に変換する。次に内容語を中心にして、その前後の機能語(前置詞・冠詞・不定詞など)をまとめる。これには次の規則を用いた。

1. 複合名詞をまとめる (例: the oil crises)

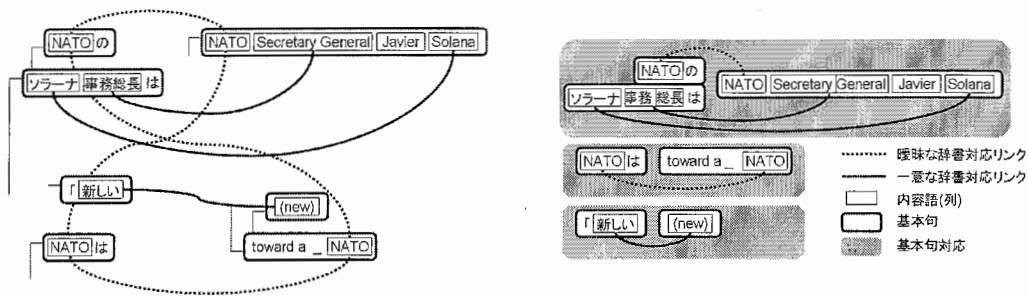


図 2: 辞書対応リンク

2. 機能語を内容語とまとめる (例: at this meeting)

3. 助動詞や助動詞的表現を主動詞とまとめる (例: had better study, will be taken)

日本語側で形容詞は1つの文節として扱われるため、英語側でも形容詞は1つの独立した基本句とする。例えば、“～ an important step ～”は、“an step”と“important”に分けて基本句とする。この規則によって基本句でもとの語順が保存されない場合には、“an _ step”, “(important)”のように表記する。

(3.1.2) 辞書の情報による対応関係の推定

翻訳辞書を用いて日英の内容語間に対応をつける。利用した翻訳辞書は文アライメントで用いたものと同じものである。対訳文中の両言語の語(列)が辞書に載っていれば、対訳文中の両言語の語(列)同士をリンク付ける。両言語とも内容語を含まない対応関係(例: 「で」⇔“at”など)や、片側の言語で内容語を含まない対応関係(例: 「進行中の」⇔“on”など)については、曖昧性が高いために辞書対応リンクを付与しない。

ある語(列)の訳語が相手側言語に複数存在すれば、語(列)は複数の辞書対応リンクを持つ。以降、このような曖昧な辞書対応リンクを曖昧な辞書対応リンクよぶ。一方、辞書対応リンクが1本だけ付与されている単語(列)において、その辞書対応リンクを一意的な辞書対応リンクよぶ。図2左は先の対訳文(図1)に張られる辞書対応リンクの例である。

次に辞書対応リンクを用いて、基本句の対応(基本句対応)を推定する。まず、一意的な辞書対応リンクで結ばれている基本句同士を採用する。次に、曖昧な辞書対応リンクで接続されている基本句は、辞書対応リンクの状態とすでに推定された基本句対応のうち周辺に存在するものを手がかりとして、もっともらしいものを採用する。先の例では、辞書対応リンクをもっとも多く持てる基本句対応の組み合わせから図2の右の3つの基本句対応が採用される。

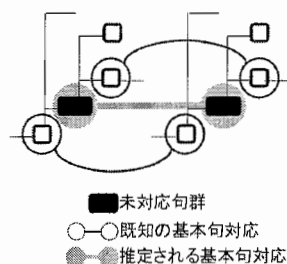


図 3: 未対応句の推定規則の例

(3.1.3) 未対応句の処理規則

先の処理で推定された基本句対応だけでは、対応付けられていない基本句 (以降、未対応句とよぶ) が残る場合がある。そこで、コーパスを調査して未対応句を処理する規則を作成した。処理は次の2つに分類され、規則に適合した場合に未対応句の対応付けを行う。

1. 未対応句同士を基本句対応とする (以降、新規対応とよぶ)
2. 未対応句をすでに推定された基本句対応に含めて新たな基本句対応とする (以降、拡張対応とよぶ)

図3は規則の一例である。この規則により、両言語の n 個以下の連続する未対応句 (以降、未対応句群とよぶ) の親同士が対応付けられていおり、子同士が対応付けられている場合に未対応句群同士を対応付ける。このような規則は6つあり、規則に適合する未対応句群が存在しなくなると、未対応句の処理は終了する。

3.2 実験結果と考察

文アライメントを行った96記事に含まれる1文:1文の対応の145対訳文ある。これらに対して文対応スコア別に基本句対応を推定した。未対応句処理の規則が未対応句を取り扱う個数 n は1とした¹。評価を行うために、145対訳文に対して、正しい対応関係を内容語単位で作成し (以降、この対応を内容語対応とよぶ)、これを用いて情報検索と同様に適合率と再現率の2つの尺度で行なった。ただし、出力は基本句対応であるのに対して正解は内容語対応なので、適合率は基本句対応の適合率、再現率は内容語対応の再現率とした。

$$\text{適合率} = \frac{(\text{内容語対応を完全に含んでいる基本句対応の数})}{(\text{システムが推定した基本句対応の数})}$$

$$\text{再現率} = \frac{(\text{基本句対応に完全に含まれている内容語対応の数})}{(\text{内容語対応の数})}$$

¹この値の精度がもっともよかった

文対応スコア	0 ～ 10	10 ～ 20	20 ～ 30	30 ～ 40	40 ～ 50	50 ～
適合率 (%)	22.2	35.5	54.5	63.3	69.6	67.7
再現率 (%)	39.2	49.2	68.0	66.1	78.7	75.0

表 6: 文対応スコアと句アライメント精度

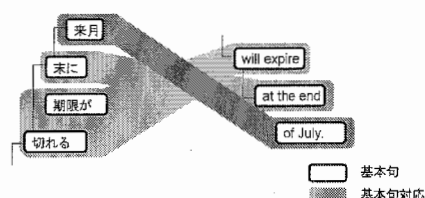


図 4: 未対応句を処理した対応の不正解

精度は、表 6 のようになった。文対応スコアが高くなるほど、適合率 - 再現率はおおむね高くなる傾向にあり、最高で適合率 69.6 %、再現率 78.7 % となる。

誤り原因を調べた結果、次の 2 つの場合に分類できる。

1. 対訳文の表現が異なる
2. 対訳文の構造が異なる

まず、(1) は、意味的に異なる表現が対応として推定される場合である。例えば、図 4 では、両言語で構造は一致しているため、未対応句同士の対応「来月」⇔ “of July” が規則により推定されている。このような問題をどう扱うかは今後の課題とする。

次に (2) 構造が異なるとは、両言語の依存構造が異なり、正しい対応が提案手法で推定されない場合である。この問題については、文献 [8] にて両言語の解析結果を照合し、適切な統語解析を選択するという手法が提案されており、本稿では取り扱わない。

4 まとめ

本稿は，記事同士の対応がとられている NHK ニュースコーパスの文アライメントと句アライメントを行った．句アライメント精度は最高でも適合率 69.6 % であり，十分な精度ではないが，より大規模な実験を行い，文対応スコアの高い部分を集めれば，句アライメント精度の高い文対応を抽出することが考えられる．今後の課題としては，句アライメント結果を翻訳知識としてどう利用するかを調べる予定である．

参考文献

- [1] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. Aligning sentences in parallel corpora. In *Proceedings of the ACL 1991*, pp. 169–176, 1991.
- [2] Eugene Charniak. A maximum-entropy-inspired parser. In *In Proceedings of NAACL 2000*, pp. 132–139, 2000.
- [3] Satoshi Sato Eiji Aramaki, Sadao Kurohashi and Hideo Watanabe. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of MT Summit VIII*, pp. 27–32, 2001.
- [4] William A. Gale and Kenneth W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, Vol. 19, No. 1, 1993.
- [5] Masahiko Haruno and Takefumi Yamazaki. High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the ACL 1996*, pp. 131–138, 1996.
- [6] Martin Kay and Martin Roscheisen. Text-translation alignment. *Computational Linguistics*, Vol. 19, No. 1, 1993.
- [7] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, 1994.
- [8] Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. Structural matching of parallel texts. In *Proceedings of the ACL 93*, pp. 23–30, 1993.
- [9] Takehito Utsuro, Hiroshi Ikeda, Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text matching using bilingual dictionary and statistics. In *Proceedings of the 15th COLING*, pp. 1076–1082, 1994.