辞書と統計を用いた対訳アライメント

春野 雅彦　　山崎 毅文
NTT コミュニケーション科学研究所

1 はじめに

テキストの電子化が進むと共に、対訳コーパスを利用した語彙知識獲得、機械翻訳などの研究が盛んになっている。こうした研究では大量の文対応付きコーパスを仮定しているため、対訳コーパスの文対応付けを高精度で自動的に行なうことが重要である。

これまでに行なわれた文対応付けの研究は大きく2つの流れに分類出来る。1つは文に含まれる単語数 [2] や文字数 [3] などの簡単な属性に基づく方法、他方は2言語の訳語対を利用する語彙に基づく方法である。前者は動的計画法を利用した効率的な実装が可能なこともあり広く利用されているが、その対象となるのは英語 - フランス語などの構造的に類似した言語間の硬い翻訳に限られる。日本語 - 英語などの構造が異なる言語では文字体系、文法体系の違いによってこれらの手法の精度は著しく落ちる。

一方、語彙に基づく方法は2言語間の訳語対を利用して文対応付けを行なうもので、[4] は統計的手法を用いた訳語対の獲得と文対応付けを徐々に行なう反復緩和法を提案している。この手法はヨーロッパ言語間では高い精度を達成するが、構造的に異なる言語ではうまく機能しない。これは主に文法の体系が異なるためである。ヨーロッパ言語間では冠詞、前置詞、否定辞などの機能語や副詞を文対応付けの手がかりとして利用できるが、構造の異なる日本語 - 英語間ではそうはいかない。つまり統計的に獲得された訳語対だけでは十分な文対応付け精度は達成出来ない。[5] は既存の対訳辞書と統計的手法を併用した文対応付け手法を提案している。まず対訳辞書のみによる動的計画法で文対応付けを行ない、統計的訳語対を獲得し、対訳辞書と統計的訳語対の両者を利用してもう一度動的計画法を繰り返す。この手法の問題点は動的計画法で利用する評価関数の意味が不明確であること、2パスであるため精度が対訳辞書の内容に大きく左右されることである。

本稿では上記の問題点を解決するため、統計的手法と既存の対訳辞書を同時に用いる対訳文対応付けシステムを提案する。統計に基づく手法は文脈に応じた訳語を獲得出来るため、ロバストである。一方、対訳辞書による手法は、コーパスに1度しか出現しない単語の情報も利用出来る利点がある。このことから両者が文対応付けにおいて、お互いに相補的な役割を果たすことが分かる。

本稿の構成は以下の通りである。まず次章で本システムの概要を述べ、3章で文対応付けアルゴリズムを説明する。4章では新聞の社説、科学論文などに対して行なった実験結果について報告する。5章ではユーザーが容易に文対応付け結果を確認し、修正できる統合的環境 BACCS について述べ、6章でまとめる。

2 システム構成

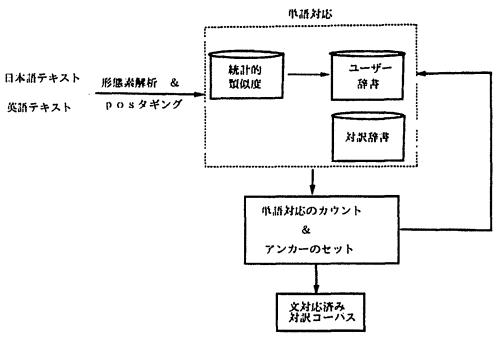

図 1: 文対応付けシステムの概要

図 1 に文対応付けシステムの概要を示す。お互いが対訳となっている日英のテキストがシステムの入力である。パターンマッチによって文の境界を確定した後、両言語の形態素解析を行う。[1] 英語の単語に関しては約100の規則を用いて原形を推定している。次にこうして得られた単語列から名詞、動詞、形容詞、副詞、未知語（日本語のみ）を抽出する。これは機能語を残しておくと文法構造の違いのために文対応精度が落ちるためである。

文対応付けアルゴリズムの初期状態は、既に対応している文のペア（アンカー）の集合である。これらのアンカーは記事や章の境界から決定され、最も一般的な場合にはテキストの最初と最後の文のみが初期アンカーを構成する（図 2 参照）。次に文対応可能なペアをこれらのアンカーから決定する。直観的に言えば、可能な文対応の範囲はアンカーの近くで小さく、アンカーから離れるほど大きくなる（図 2参照）。ここで重要なことは正しい文対応が可能な文対応候補の中に含まれていることである。

[1] 日本語、英語の形態素解析には各々 JUMAN システム [6]、Eric Brill の tagger[1] を使用した。これらのツールを提供して頂いた方々に感謝します。

続いて、文対応付けアルゴリズムは2種類の訳語対(統計的手法と対訳辞書から得られたもの)を用いて、文対応可能候補から新しいアンカーを見つける。決められた閾値以上の訳語対を含む文のペアを新たにアンカーとするのである。これら新しく見つかったアンカーによって、統計的訳語推定がより正確になる。この操作をパラメータの緩和を行ないながら繰り返すことで、信頼性の高いものから低いものへと順にアンカーが決定される。緩和法を用いることでアルゴリズムの終盤で起こる対応付け誤りが、全体の対応付け精度に及ぼす影響が少なくなる。

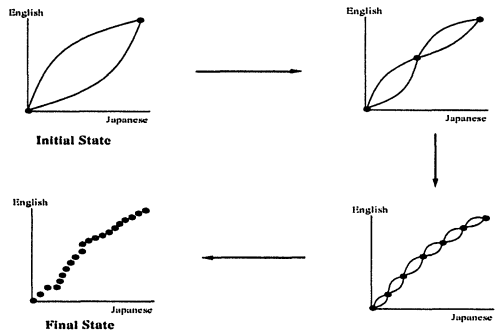

図 2: 文対応付けの過程

3 文対応付けアルゴリズム

3.1 統計的類似度

我々が用いる基本的なデータ構造は**文対応可能行列(ASM)**と**アンカー行列(AM)**の2つの行列である。ASMは対応可能な文のペアを表し、0と1から構成される。ASMの$i-j$要素が1であるのは日本文iと英文jが対応可能である場合であり、0であるのは対応不可能な場合である。これに対してAMのi-j要素は日本文iと英文jの文ペアが含む訳語対の'のべ総数'である。すなわち日本文iと英文jのペアがいくつの訳語対によって支持されたかを表す。

ここでp_iを日本文iとその対応可能な英文から成る集合であるとする。例えばp_2=$\{Jsentence_2$,$Esentence_2$, $Esentence_3\}$は、$Jsentence_2$が$Esentence_2$、$Esentence_3$と対応可能であることを意味する。

日本語単語w_{jpn}と英単語w_{eng}の出現に関する類似度を評価するために以下に示す分割表を考える。分割表は以下の4つの頻度分布から構成される: (a) w_{jpn}とw_{eng}の両方が現われたp_iの数、(b) w_{eng}のみが現われたp_iの数、(c) w_{jpn}のみが現われたp_iの数、(d) w_{jpn}とw_{eng}のどちらも現われないp_iの数。ただし1つのw_{eng}は(a)で高々1回しか数えないようにする。

	w_{jpn}	
w_{eng}	a	b
	c	d

分割表中におけるw_{jpn}と英単語w_{eng}の類似度を定量的に評価するために以下の(自己)相互情報量を導入する。

$$\log \frac{prob(w_{jpn}, w_{eng})}{prob(w_{jpn})prob(w_{eng})}$$

ここで各々の確率は以下のように与えられる。

$$prob(w_{jpn}) = \frac{a+c}{a+b+c+d} = \frac{a+c}{M}$$

$$prob(w_{eng}) = \frac{a+b}{a+b+c+d} = \frac{a+b}{M}$$

$$prob(w_{jpn}, w_{eng}) = \frac{a}{a+b+c+d} = \frac{a}{M}$$

良く知られているように相互情報量は出現頻度の低い事象に対して大きくなる性質がある。そこでその信頼性を評価するために以下のt-スコアを併用する。信頼性の低い相互情報量の値はt-スコアに閾値を設定することで排除可能となる。以上2つのパラメータを用いることで、他の類似度基準[4]に比べ、パラメータの緩和過程を細かく制御出来る。

$$t \approx \frac{prob(w_{jpn}, w_{eng}) - prob(w_{jpn})prob(w_{eng})}{\frac{1}{M}prob(w_{jpn}, w_{eng})}$$

3.2 対応付けアルゴリズム

以下のアルゴリズムによって2章で述べたアラインメント手法を実現している。

文対応可能行列(ASM)の初期化 この段階で初期的なASMを構成する。テキストがM文の日本文とN文の英文から成る場合、ASMは$M \times N$行列である。まず記事や章の境界を用いて初期的なアンカー集合を決定する。

次に、このアンカー集合を用いて可能な文対応の範囲を決定する。直観的には文対応はアンカーを結ぶ直線に近く分布するはずである。我々は2つのアンカーの中央辺りでは1文に対して$O(\sqrt[3]{L})$(Lは2つのアンカーの間に存在する文数)の文を対応付ける関数を用いる。これは統計学的に最大の分散が$O(\sqrt[3]{L})$でモデル化出来るからである[4]。

アンカー行列(AM)の構成 この段階では与えられた文対応行列(AM)と対訳辞書を用いてアンカー行列(AM)を更新する。ここでt_{high}、t_{low}、I_{high}、I_{low}をt-スコアと相互情報量に対する各々2つの閾値であるとする。また、ANCを文ペアがアンカーと認定されるために必要な最小の訳語対の数であるとする。

まず、ASMの対応可能文ペア中に含まれる全ての単語対に対してt-スコアと相互情報量を計算する。これ以降はt_{low}とI_{low}を越えるt-スコアと相互情報量を持つ単語対だけを考察の対象とする。ASM中の全ての対応可能ペア$Jsentence_i$と$Esentence_j$に対して以下の操作を行なう。

1. 次の3つの条件が成り立てばAMのi-j要素に3を加える。(1) $Jsentence_i$と$Esentence_j$がt_{high}、I_{high}を越える統計的訳語対(w_{jpn}とw_{eng})、あるいは対訳辞書の訳語対(w_{jpn}とw_{eng})を含んでいる。(2)w_{eng}が$Jsentence_i$と対応可能な他の英文に出現しない。
(3) $Jsentence_i$と$Esentence_j$が既に閾値ANCを越えている他のアンカーと交差しない。
2. 次の3つの条件が成り立てばAMのi-j要素に1を加える。(1) $Jsentence_i$と$Esentence_j$がt_{low}、I_{low}を越え、t_{high}、I_{high}以下の統計的訳語対(w_{jpn}とw_{eng})を含んでいる。
(2)w_{eng}が$Jsentence_i$と対応可能な他の英文に出現しない。(3) $Jsentence_i$と$Esentence_j$が既に閾値ANCを越えている他のアンカーと交差しない。

最初の手続きは対訳辞書の訳語対と信頼性の高い統計的訳語対を扱うものである。これらの訳語対に対してはAMに等しく3を付加している。2番目の訳語対は1番目の訳語対だけでは文対応付けに十分な数が得られないため導入される。これらの訳語対に対してはAMに1を付加する。この手続きで採用される訳語対は訳語としては誤っていることもあるが、文対応付けアルゴリズムの終盤で重要な働きをする。

ASMの更新 この段階で上記の手続きで構成したAMを用いてASMを更新する。AM中でANC個以上の訳語対を持つ文ペアはアンカーと見なされる。このアンカー集合から初期的なASMを構成したのと全く同様にして新しいASMが構成される。

我々のアルゴリズムはt_{low}、I_{low}、ANCを徐々に下げることで反復緩和法を実現している。

4 実験並びに評価

この章では様々な対訳コーパスに対して行なった文対応付け実験について、文対応の精度と統計的に得られた訳語対の観点から報告する。実験に用いたテキストの種類と文対応付け精度を表1にまとめた。最初2つのテキストは読売新聞の社説記事であり、WWWサーバーから得た[2]。3番目は経済評論、4番目はサイエンスの科学記事である。表中にはそれぞれのテキストの文数を示してある。

4.1 文対応付けの精度

統計、対訳辞書はそれぞれ統計的手法、対訳辞書によって得られた訳語対だけを用いて文対応付けを行なった結果である。ここでは対訳辞書として講談社和英辞典を利用した。表中でPRECISIONは自動的に付けられた文対応のうち正しいものの割合いを示し、RECALLは人手で付けた正解のうち対応付けプログラムで再現出来たものの割合いを示している。

テキスト1、テキスト2の短いテキストに対しては統計的手法は著しく成績が悪く、テキスト3、テキスト4に対しては統計的手法の方が対訳辞書のみの手法より成績が良い。これはテキストがある程度の長さを越えると統計によって文脈に依存した訳語対を抽出出来るためである。ここで注意すべきことは全てのテキストに対して提案手法が最も良い成績を収めていることである。

4.2 統計的に獲得された訳語対

この節では脳の非侵襲測定に関するテキスト4を例にして、どのような訳語対が統計的に得られるかを示す。表2にテキスト4から得られた訳語対のうち相互情報量の値が大きいものを示す。表中のNMR、MEG、PET、CT、MRI、Functional MRIはどれも脳を外部から測定するための装置であり、このテキストのキーワードである。ところがこれらの専門用語は我々が用いた対訳辞書には含まれていなかった。

またMEGの正しい日本語訳は脳磁図であるが、形態素解析システムがこれらを脳、磁、図に分解してしまっている。それにもかかわらず統計的手法によって磁や図とMEGの対応が捉えられている。このように統計を利用することで分野依存の訳語対を獲得し、文対応付けの精度を向上させることが可能となる。

5 統合的文対応付け環境BACCS

精度の高い対訳コーパスを大量に作るにはユーザが素早く対応付け結果を確認し、必要があれば修正を行なわなければならない。このためのツールとして我々は統合的文対応付け環境BACCS(Bilingual Aligned Corpus Construction System)を作成した。ユーザは次の操作手順で対訳コーパスを作成する。図3にBACCSの画面を示す。

1. 自動文対応付けプログラムを走らせて結果を画面に表示する。
2. プログラムによって出された信頼度(色で区別)を参考にして、結果を確認し、必要があれば修正する。

[2]データの使用を許可頂いた読売新聞社に感謝致します。

テキスト	日本文	英文	提案手法		統計		対訳辞書	
			PRECISION	RECALL	PRECISION	RECALL	PRECISION	RECALL
1 Root out guns at all costs	26	28	96.4%	96.3%	65.0%	48.5%	89.3%	88.9%
2 Economy facing last hurdle	35	42	95.3%	93.1%	61.3%	49.6%	87.2%	75.1%
3 The Post-Cold-War World	134	123	96.5%	97.1%	87.3%	85.1%	86.3%	88.2%
4 Visualizing the Mind	223	212	91.6%	93.8%	82.2%	79.3%	74.3%	63.8%

表 1: 文対応付けの結果

日本語	英語	相互情報量
記録	recording	3.58
リアルタイム	real	3.51
ニューロン	neuron	3.51
フィルム	film	3.51
グルコース	glucose	3.51
増加	increase	3.51
磁	MEG	3.51
解像度	resolution	3.43
電気	electrical	3.43
グループ	group	3.39
電気	recording	3.39
記録	electrical	3.39
言う	generate	3.33
提供	provide	3.33
図	MEG	3.33
NMR	NMR	3.17
ファンクショナル	functional	3.17
機器	equipment	3.17
臓器	organ	3.15
注射	compound	3.10
水	water	3.10
標識	radioactive	3.10
ＰＥＴ	PET	3.10
解像度	spatial	3.10
そのようだ	such	3.10
代謝	metabolism	3.06
言う	verb	3.04
科学者	scientist	2.95
地図	mapping	2.92
大学	university	2.92
思考	thought	2.90
化合物	compound	2.82
標識	label	2.82
オートラジオ	radioactivity	2.77
視覚	visual	2.77
信号	signal	2.77
リアルタイム	time	2.69
オートラジオ	autoradiography	2.67
能力	ability	2.63
ＣＴ	CT	2.63
聴覚	auditory	2.15
心	mental	2.05
ＭＲＩ	MRI	1.8

表 2: 統計的に得られた単語対応の一部

3. 統計的に得られた訳語対の中から適切なものをユーザー辞書に登録する。

6 まとめ

我々は日本語 - 英語のような構造の異なる言語間の対訳コーパスに対して高い精度で文対応付けを行なうための手法を提案した。本手法は統計的手法と対訳辞書から得られる訳語対を同時に利用することにより、分野や長さの異なる対訳コーパスに対して安定して高い精度の文対応付けを行なえた。

謝辞 本研究に関して辞書検索プログラムを提供、ならびに有益な議論をして頂いた宇津呂 武仁氏に感謝致します。

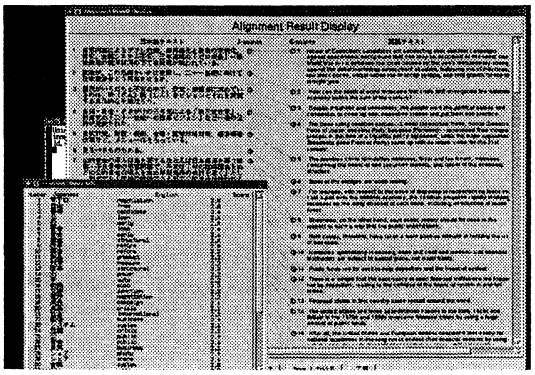

図 3: 統合的文対応付け環境 BACCS

参考文献

[1] Eric Brill. A simple rule-based part of speech tagger. In *Proc. Third Conference on Apolied Natural Language Processing*, pages 152–155, 1992.

[2] P F Brown et al. Aligning sentences in parallel corpora. In *the 29th Annual Meeting of ACL*, pages 169–176, 1991.

[3] W A Gale and K W Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102, March 1993.

[4] Martin Kay and Martin Roscheisen. Text-translation alignment. *Computational Linguistics*, 19(1):121–142, March 1993.

[5] Takehito Utsuro, Hiroshi Ikeda Masaya Yamane, Yuji Matsumoto, and Makoto Nagao. Bilingual text matching using bilingual dictionary and statistics. In *Proc. 15th COLING*, pages 1076–1082, 1994.

[6] 松本 裕治他. 日本語形態素解析システム *JUMAN* 使用説明書 *2.0*. TR94025, 奈良先端科学技術大学院大学, 1994.