

(19) 日本国特許庁(JP)

(12) 特許公報(B2)

(11) 特許番号

特許第3820452号  
(P3820452)

(45) 発行日 平成18年9月13日(2006.9.13)

(24) 登録日 平成18年6月30日(2006.6.30)

(51) Int. Cl.

G06F 17/28 (2006.01)

F I

G06F 17/28

U

請求項の数 3 (全 15 頁)

(21) 出願番号	特願2003-160464 (P2003-160464)	(73) 特許権者	301022471
(22) 出願日	平成15年6月5日(2003.6.5)		独立行政法人情報通信研究機構
(65) 公開番号	特開2004-362305 (P2004-362305A)		東京都小金井市貫井北町4-2-1
(43) 公開日	平成16年12月24日(2004.12.24)	(74) 代理人	100103827
審査請求日	平成15年6月5日(2003.6.5)		弁理士 平岡 憲一
特許法第30条第1項適用	2003年3月18日 言語処理学会発行の「言語処理学会第9回年次大会 発表論文集」に発表	(74) 代理人	100119161
前置審査			弁理士 重久 啓子
		(72) 発明者	馬 青
			東京都小金井市貫井北町4-2-1 独立行政法人通信総合研究所内
		(72) 発明者	張 玉潔
			東京都小金井市貫井北町4-2-1 独立行政法人通信総合研究所内

最終頁に続く

(54) 【発明の名称】 対応付け装置及びプログラム

(57) 【特許請求の範囲】

【請求項1】

一方の言語の一定量の文書データを格納するコーパスデータと、  
 他方の言語から一方の言語に翻訳する辞書を格納する翻訳辞書と、  
 入力された対訳文の単語のコーディングを行うデータコーディング手段と、  
 前記入力された対訳文の単語を自動でマップする自己組織化マップ手段とを備え、  
 前記データコーディング手段は、前記入力された対訳文の一方の言語の単語を前記コーパスデータ中の前記入力された対訳文の一方の言語の単語及びその前後の単語である共起語と共起頻度で定義すると共に、前記入力された対訳文の他方の言語の単語を前記翻訳辞書を用いて一方の言語の訳語候補を求め、該求めた訳語候補から前記コーパスデータを利用して共起語と共起頻度で定義し、

前記自己組織化マップ手段は、前記共起語と共起頻度で定義した入力された対訳文の単語から前記共起語と共起頻度により単語間の距離を求め、該求めたある単語との距離が小さい単語ほど、前記ある単語に対してこれら以外の単語より小さい距離に配置するようにして、前記入力された対訳文の単語の自動マップを行うことを特徴とした対応付け装置。

【請求項2】

前記データコーディング手段は、前記共起語として前記コーパスデータ中の前記入力された対訳文の一方の言語の単語及びその前後1つずつの単語とすることを特徴とした請求項1記載の対応付け装置。

【請求項3】

10

20

コーパスデータとして一方の言語の一定量の文書データを格納する手段と、

翻訳辞書として他方の言語から一方の言語に翻訳する辞書を格納する手段と、

前記入力された対訳文の一方の言語の単語を前記コーパスデータ中の前記入力された対訳文の一方の言語の単語及びその前後の単語である共起語と共起頻度で定義すると共に、前記入力された対訳文の他方の言語の単語を、前記翻訳辞書を用いて一方の言語の訳語候補を求め、該求めた訳語候補から前記コーパスデータを利用して共起語と共起頻度で定義するデータコーディング手段と、

前記共起語と共起頻度で定義した入力された対訳文の単語から前記共起語と共起頻度により単語間の距離を求め、該求めたある単語との距離が小さい単語ほど、前記ある単語に対してこれら以外の単語より小さい距離に配置するようにして、前記入力された対訳文の単語の自動マップを行う自己組織化マップ手段として、

コンピュータを機能させるためのプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、日中対訳文（日本語とその中国語の翻訳文）等の対訳文を入力し、意味に基づく二言語の単語のアライメント（対応付け）を自動で行う対応付け装置に関する。

【0002】

【従来の技術】

対訳コーパスから翻訳知識を抽出するためには、文レベルだけでなく単語レベルでのアライメントも必要である。対訳コーパスが単語レベルでアライメントされていれば、辞書に載っていない、ドメインや時期などに依存する訳語が得られたり、複数の訳語候補へのスコアリングができたり、更には単語の対訳関係をもとにして、句や節単位の対応関係といった翻訳パターンが自動獲得されることが期待できる（例えば、非特許文献1参照。）。

【0003】

このように、アライメントは自然言語処理の分野で非常に重要かつ基本的な研究課題である。関連する研究としては、Brownらが考案した一連の統計モデル（例えば、非特許文献2、3参照。）、それから、ダイナミックプログラミングを用いる手法（例えば、非特許文献4参照。）や、最近では文脈情報を導入した統計手法（例えば、非特許文献5参照。）、さらには構造化アライメント法（例えば、非特許文献6、7、8参照。）が挙げられる。

【0004】

【非特許文献1】

Brown, Ralf D.: Automated dictionary example-based translation, Proceedings of the Seventh International Conference on Theoretical and Methodological Issues in Machine Translation, pp. 111-118. 1997.

【非特許文献2】

Brown, PF., Cocke, J., Della Pietra, SA., Della Pietra, VJ., Jelinek, F., Mercer RL., Roossin, P.: A statistical approach to language translation, COLING'88, pp. 71-76, 1988.

Brown, PF., Della Pietra, SA., Della Pietra, VJ., Mercer RL.: The mathematics of statistical machine translation: parameter estimation, Computational Linguistics, Vol. 19, No. 2, pp.263-311, 1993.

【非特許文献4】

Dagan I, Church KW, Gale WA.: Robust bilingual word alignment for machine aided translation, Proceedings of the Workshop on Very Large Corpora, pp. 1-8, 1993.

【非特許文献5】

Varea, IG., Och, FJ, Casacuberta: Improving alignment quality in statistical machine translation using context-dependent maximum entropy models, COLING2002, pp. 1051-1057, 2002.

10

20

30

40

50

## 【非特許文献 6】

Kaji, H., Kida, Y., Morimoto Y.: Learning translation templates from bilingual text, COLING'92, pp. 672-678, 1992.

## 【非特許文献 7】

Matsumoto, Y., Ishimoto, H, Utsuro, T.: Structural matching of parallel texts, A CL'93, pp. 23-30, 1993.

## 【非特許文献 8】

Imamura, K.: Hierarchical phrase alignment harmonized with parsing, NLP RS2001, p. 377-384, 2001.

## 【0005】

## 【発明が解決しようとする課題】

上記従来のもは、いずれも、共起語などの統計情報や文法的構造に基づくアプローチであり、意味に基づくものではない。よい対訳とは直訳ではなく、意味に基づくものである。このため、これまで提案されてきた統計や文法的構造に頼るアライメントの手法の限界は明らかであり、よい対訳とはいえないものであった。

## 【0006】

本発明は、このような従来の問題点の解決を図り、意味に基づく単語アライメントを目指し、日中等の対訳文を入力とした二言語の意味マップの自動構築を行うことを目的とする。

## 【0007】

## 【課題を解決するための手段】

図1は本発明の原理説明図である。図1中、1aはデータコーディング手段、2は翻訳辞書、3はコーパスデータ、4aは自己組織化マップ手段である。

## 【0008】

本発明は、前記従来の課題を解決するため次のような手段を有する。

## 【0009】

(1): 一方の言語の一定量の文書データを格納するコーパスデータ3と、他方の言語から一方の言語に翻訳する辞書を格納する翻訳辞書2と、入力された対訳文の単語のコーディングを行うデータコーディング手段1aと、前記入力された対訳文の単語を自動でマップする自己組織化マップ手段4aとを備え、前記データコーディング手段1aは、前記入力された対訳文の一方の言語の単語を前記コーパスデータ3中の前記入力された対訳文の一方の言語の単語及びその前後の単語である共起語と共起頻度で定義すると共に、前記入力された対訳文の他方の言語の単語を前記翻訳辞書2を用いて一方の言語の訳語候補を求め、該求めた訳語候補から前記コーパスデータ3を利用して共起語と共起頻度で定義し、前記自己組織化マップ手段4aは、前記共起語と共起頻度で定義した入力された対訳文の単語から前記共起語と共起頻度により単語間の距離を求め、該求めたある単語との距離が小さい単語ほど、前記ある単語に対してこれら以外の単語より小さい距離に配置するようにして、前記入力された対訳文の単語の自動マップを行う。このため、二次元で可視化して、正確な対応付けが自動ででき、また2番目に近い単語をすぐ見つけることができる。

## 【0010】

(2): 前記(1)の対応付け装置において、前記データコーディング手段1aは、前記共起語として前記コーパスデータ3中の前記入力された対訳文の一方の言語の単語及びその前後1つずつの単語とする。このため、共起語の処理データ数を少なくすることができる。

## 【0011】

## 【発明の実施の形態】

## (1): 対応付け装置の説明

図2は対応付け装置の説明図である。図2において、対応付け装置には、データコーディング部1、翻訳辞書2、コーパスデータ3、SOM部(自己組織化マップ部)4が設けて

10

20

30

40

50

ある。データコーディング部 1 は、コーパスデータ 3 と翻訳辞書 2 を用いて個々の単語を多次元ベクトルにコーディングするものである。翻訳辞書 2 は、ある国語を他の国語に変換する辞書である。コーパスデータ 3 は、新聞等のある言語の一定量の文書データである。SOM部 4 は、データコーディング部 1 がコーディングしたデータより、単語（ノード）の自動配置（マップ）を行うものである。

【0012】

図 3 は対応付け処理フローチャートである。以下、図 3 の処理 S 1 ~ S 4 に従って日本語と中国語の対訳文の単語の対応付け処理を説明する。

【0013】

S 1 : データコーディング部 1 に、単語分割された対訳文が入力される（なお、単語分割されていない対訳文が入力された場合は、形態素解析器などであらかじめ単語分割する）。

10

【0014】

S 2 : データコーディング部 1 は、コーパスデータ 3（例えば、8 年分の毎日新聞）を利用して、日本語文の単語を共起語情報のセット（共起語と共起頻度）で定義する。ここで、共起語とは、コーパスデータ 3 中のその単語自身及びその周辺（前後）の単語である。

【0015】

S 3 : データコーディング部 1 は、中国語文の単語を、翻訳辞書 2 を用い日本語の訳文候補を求め、この訳文候補をコーパスデータ 3 を利用して共起語情報のセット（共起語と共起頻度）を求める。すなわち、中国語文の単語を日本語の共起語情報のセット（共起語と共起頻度）で定義する。

20

【0016】

S 4 : SOM部 4 は、前記処理 S 2 と処理 S 3 で定義された日本語文の単語の共起語情報のセットと中国語文の単語の共起語情報のセットを用い、二次元上に、各単語を自動でマップする。

【0017】

このように、中国語単語も日本語の共起語で定義されているので、中国語と日本語を区別する必要はなくマップを行うことができる。

【0018】

以下、日本語と中国語の具体的対訳文の例により対応付け装置が作成する意味マップを説明する。

30

【0019】

(2) : 対訳コーパスにおける単語アライメントの意味マップの説明

1) 目標

本発明者らはこれまで、日本語や中国語において、意味的に近い単語どうしは近いところに、意味的に遠い単語どうしは離れたところに配置されるような、単言語の意味マップの自動構築手法を提案してきた（例えば、馬青，神崎享子，村田真樹，内元清貴，井佐原均：日本語名詞の意味マップの自己組織化，情報処理学会論文誌，Vol. 42, No. 10, pp. 2379-2391, 2001. 及び Ma, Q., Zhang, M., Murata, M., Zhou, M., Isahara, H.: Self-Organizing Chinese and Japanese Semantic Maps, The 19th International Conference on Computational Linguistics (COLING'2002), Taiwan, pp. 605-611, August, 2002. 参照）。もし、対訳文を入力とした二言語（あるいは多言語）の意味マップが自動的に構築できれば、その意味マップから単語のアライメントが簡単に取れるであろう。そして、単言語の意味マップと同様、その結果は可視性や連続性を有するため、一対多や多対一のアライメントの取り扱いが容易になる。さらに、二言語の意味マップは例えば対訳コーパスを用いた外国語の学習支援や外国語の作文支援などにも応用できる。もっとも、よい対訳は直訳ではなく意識によるものが多いため、これまで提案されてきた統計や文法的構造に頼るアライメントの手法の限界は明らかであり、最終的には意味に基づく方法を模索する必要があろう。

40

【0020】

50

本発明では、意味に基づく単語アライメントを目指し、日中対訳文を入力とした日中二言語の意味マップの自動構築手法を提案する（なお、現在の意味マップは、基本的に共起情報に基づいて構築される。）。

【0021】

提案手法の有効性を確かめる実験には、京大コーパスVer3.0とその中国語訳の対訳コーパスを用いる。また、意味マップの自動構築に必要な学習データは1991年～1998年の8年分の毎日新聞から得られるものとした。

【0022】

2)自己組織化神経回路網モデルの説明

意味マップの自動構築マシンとしてはKohonenの自己組織化神経回路網モデルである自己組織化マップ部4（Self-organization Map, 略してSOM）（Kohonen, T.: Self-organizing maps, Springer, 2nd Edition, 1997.）を用いる。SOMは高次元入力を持つ2次元配列のノードで構成され、以下に述べる自己組織化によって、高次元データをその特徴を反映するように2次元空間にマッピングすることができる。

【0023】

【数1】

入力  $x = [\xi_1, \xi_2, \dots, \xi_N]^T \in \mathcal{R}^N$  ならば、個々のノード  $i$  はそれぞれ参照ベクトル  $m_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{iN}]^T \in \mathcal{R}^N$  を持つものとする。

【0024】

但し、参照ベクトルの要素  $\mu_{ij}$  はノード  $i$  と入力要素  $\xi_j$  の間の重みであり、自己組織過程において少しずつ修正される。入力ベクトル  $x$  が与えられたとき、まず、その入力をすべてのノードの参照ベクトルと比較し、ユークリッド距離の一番短いノードを活性化する。マッピング処理段階ではこのノードのみ活性化される。このノードを勝者ノードと呼ぶ。即ち、勝者ノード  $c$  は以下の式1のように選ばれる。

【0025】

【数2】

$$c = \operatorname{argmin}_i \{ \|x - m_i\| \} \quad \text{式 1}$$

【0026】

一方、自己組織化過程では、グローバルに自己組織化が行われるように、勝者ノードだけでなくその近傍のノードも活性化させ、リラックス処理を行う。即ち、活性化されたすべてのノードに対し、それらの参照ベクトルを入力ベクトルに近づくように修正を行う。

【0027】

【数3】

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \quad \text{式 2}$$

【0028】

ここで、 $t$  は学習回数で、 $h_{ci}(t)$  は、例えば以下の式3のように定義された近傍関数である。

【0029】

【数4】

10

20

30

40

$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right) \quad \text{式 3}$$

但し、 $r_c \in \mathcal{R}^2$  と  $r_i \in \mathcal{R}^2$  はそれぞれ勝者ノード  $c$  と近傍ノード  $i$  の位置ベクトルである。

【0030】

従って、項  $r_c - r_i$  は近傍ノード  $i$  が勝者ノード  $c$  から離れて行くにつれ、 $h_{ci}$  が小さくなり  $m_i(t)$  の修正量が小さくなることを意味する。また、 $\alpha(t)$  は学習率で、 $\sigma(t)$  は近傍の大きさ（半径）である。これらは時間と共に単調に減少していく関数であればよい。

【0031】

通常、学習過程は「整列」フェーズと「微調整」フェーズからなる。「整列」フェーズにおいては  $\alpha(t)$  と  $\sigma(t)$  の初期値を共に大きく取り、時間と共に減少して行く。ノードの配置の基本形はこのフェーズで形成される。一方、残りのフェーズでは、 $\alpha(t)$  と  $\sigma(t)$  は小さい値のまま長時間をかけて、初期フェーズで形成された基本形を微調整する。

【0032】

3) 単語アライメントの意味マップの自己組織化の説明 20  
(目的)

単語アライメントの意味マップの自己組織化とは、以下のような対訳文が与えられたとき、何らかの教師なし学習データを用いることによってそれらの文に出現するすべての単語が意味に応じて一枚のマップに自動配置されることである。

【0033】

(日) 経営 トップ が 低 成長 時代 定着 を 実感 して いる こと を うかがわ せた 。

【0034】

(中) 由此 可以 看出 ， 最高 経営者 深感 經濟 仍 停 留 在 低 速 增 長 時代 。

【0035】

(データの説明)

日中機械翻訳プロジェクトの一環として、京大コーパスVer3.0をベースとした日中の対訳コーパスを構築中である。対訳文はこの対訳コーパスから取り出したものである。京大コーパスはもともと形態素解析済のものなので、日本語文は形態素解析済のものをそのまま使うことにした。一方、中国語訳文については、北京大学の形態素解析ツール（周強，段慧明：現代漢語語料庫加工中の切詞与詞性標注処理，中国計算機学報，Vol.85，1994. 参照）を用いて単語分割及び品詞の付与を行った。

【0036】

異なる言語を同じ評価尺度で取り扱えるようにするために、中国語の訳文に現れる中国語の単語については、「漢日辞典」（吉林大学、吉林教育出版社）及び「中日大辞典」（愛知大学、大修館書店）（なお、「漢日辞典」にエントリがない場合のみ「中日大辞典」を利用した。）より人手で最大5個まで（この最大5個の訳語は以下の優先順序で選択した：(1) 日本語文にも現れるもの；(2) 元の中国語単語と品詞が一致するもの；(3) 辞書に載っている順番；(4) 京大コーパスに現れたもの。但し、形容動詞の訳語はその語幹のみを、形容詞の訳語をその中止形を、動詞の訳語をその原形を用いることにした。）の日本語訳語を付与し、それらの訳語を代わりに用いることにした。そうすると、上記中国語の訳文が以下ようになる。その結果、例えば上記中国語訳文のそれぞれの単語に以下のような日本語候補が付与された。

【0037】

(中) 由此：これによって  
 可以：ことができる / てよい  
 看出：見抜く / 看破  
 最高：最高 / 最も高い  
 経営者：経営者  
 深感：実感  
 経済：経済 / 生活 / 経済的  
 仍：依然として / いまなお  
 停留：滞在 / 止まる  
 在：で / に / している / しつつある  
 低速：低  
 増長：増長 / ふえる  
 時代：期 / 時代

10

。：。

## 【0038】

このような方法を用いることによって、日本語という単一言語で表される対訳文が得られる。但し、この例からも分かるように、「これによって」や「ことができる / てよい」など、ほとんどの日本語訳が日本語の原文に存在していない。従って、対訳文の言語が統一されたとしても、単純に単語間の表層表現でアライメントをとることは無理である。

## 【0039】

自己組織化に用いる実際の学習データは以下のようにして得た。日本語文に現れる日本語の単語については、1991年～1998年の8年分の毎日新聞から得られた共起語（その単語自身及び前後一つずつの単語）を用いて定義し、自己組織化の学習データとした。一方、中国語文に現れる中国語の単語は、それらに付与された日本語の訳語候補の共起語（それぞれの訳語候補及び前後一つずつの単語）を用いて定義し、自己組織化の学習データとした。次では学習データの具体的な構成及びSOMの入力ベクトルへのコーディングについて述べる。

20

## 【0040】

(データコーディングの説明)

日中对訳文が、次のように与えられたとする。

30

## 【0041】

【数5】

$$J_1, J_2, \dots, J_m$$

$$C_1 : J_{11} / \dots / J_{1,n_1}, \dots, C_n : J_{n1} / \dots / J_{n,n_n}$$

## 【0042】

但し、 $J_i$  ( $i = 1, \dots, m$ ) は日本語の文を構成する単語、 $C_i$  ( $i = 1, \dots, n$ ) はその訳文を構成する単語、 $J_{ij}$  ( $i = 1, \dots, n, j = 1, \dots, n_i$ ) は  $C_i$  の  $j$  番目の訳語候補、 $n_i$  ( $1 \leq n_i \leq t$ ) は  $C_i$  の訳語候補の数、 $t$  は最大候補数（この例においては  $t = 5$ ）である。日本語文の単語  $w_i$  ( $i = 1, \dots, m$ ) は、以下の式4のように共起語情報のセットで定義される。

40

## 【0043】

【数6】

$$w_i = J_i = \{a_1^{(i)}, f_1^{(i)}, \dots, a_{\alpha_i}^{(i)}, f_{\alpha_i}^{(i)}\} \quad \text{式 4}$$

但し、 $a_j^{(i)}$  と  $f_j^{(i)}$  は  $J_i$  の共起語と正規化（つまり、 $\sum_{j=1}^{\alpha_i} f_j^{(i)} = 1$ ）

された共起頻度で、 $\alpha_i$  は  $J_i$  と共起する単語の数である。

【0044】

一方、中国語訳文の単語  $w_j$  ( $= C_j$ ) は以下の式5のように共起語情報のセットで定義 10  
される。

【0045】

【数7】

$$w_j = C_j = \{J_{j1}, \dots, J_{j,n_j}\} = \{a_1^{(j)}, f_1^{(j)}, \dots, a_{\alpha_j}^{(j)}, f_{\alpha_j}^{(j)}\} \quad \text{式 5}$$

但し、 $a_i^{(j)}$  と  $f_i^{(j)}$  は  $J_{j1}, \dots, J_{j,n_j}$  のいずれか（あるいは複数個）の共起語と正規化された共起頻度（複数個と共起している場合はそれぞれの共起頻度の和）で、 $\alpha_j$  は 20  
 $C_j$  と共起する単語の数である。

【0046】

つまり、一つの訳語候補とでも共起していれば、元の中国語の共起語と見なされる。

【0047】

このように、中国語単語も日本語の共起語で定義されているので、中国語と日本語を区別 30  
する必要がなく、これまで提案してきた単言語の意味マップの構築に関するすべてのデータコーディング法を用いることが可能である。本発明では、対訳文に現れる任意の両単語  $w_i$  と  $w_j$  の意味的距離  $d_{ij}$  を以下の式6に示す頻度重み付け法で求める。

【0048】

【数8】

$$d_{ij} = \begin{cases} \frac{(F_i - F_{ij}) + (F_j - F_{ij})}{F_i + F_j - F_{ij}} & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad \text{式 6}$$

【0049】

但し、 $F_i$  と  $F_j$  はそれぞれ  $w_i$  と  $w_j$  が持つ共起語の数  $i$  と  $j$  の拡張で、 $F_{ij}$  は  $w_i$  と  $w_j$  の共通する共起語の数  $c_{ij}$  の拡張である。これらは以下の式7で求められる。 40

【0050】

【数9】



$$F_i = \sum_{x=1}^{x=\alpha_i} f_x^{(i)} \quad \text{and} \quad F_{ij} = \sum_{x=1}^{x=c_{ij}} f_x^{(ij)} \quad \text{式 7}$$

但し、 $f_x^{(i)}$  は単語  $w_i$  と共起語  $a_x^{(i)}$  との共起頻度、 $f_x^{(ij)}$  は単語  $w_i$  と  $w_j$  と共起語  $a_x^{(i)}$  との共起頻度である ( $x = 1, \dots, \alpha_i$ )。

10

【 0 0 5 1 】

このようにして、距離  $d_{ij}$  を要素とする相関行列が求められる。そして、個々の単語  $w_i$  を相関行列  $D$  の  $i$  行目の要素で構成される多次元ベクトルにコーディングする。

【 0 0 5 2 】

【 数 1 0 】

$$V(w_i) = [d_{i1}, d_{i2}, \dots, d_{iN}]^T \quad \text{式 8}$$

ここで、 $N$  は対訳文の単語の総数、すなわち  $N = m + n$  で、 $V(w_i) \in \mathbb{R}^N$  は SOM の入力である。

20

【 0 0 5 3 】

4) 具体的な実験結果の説明

データ：前記3)の(データの説明)に述べた対訳文(10ペア)を単語のアライメント実験の対象とした。学習データは、前記3)の(データの説明)に述べた方法で得た。前記3)の(データの説明)に挙げた対訳文を例としてみれば、単語の総数は  $N = m + n = 16 + 15 = 31$ 、共起語ののべ総数は62,627、異なり総数は22,077であった。このうち、日本語文の「。」と中国語訳文の「。」(実際、ピリオドのアライメントは必要ないが、ここでは機械的に処理するというので、省かないことにした。)の共通する共起語がもっとも多く(4,180個)、日本語文の「うかがわ」と中国語訳文の「，」の共通する共起語がもっとも少なかった(5個)。

30

【 0 0 5 4 】

SOM：実験には  $13 \times 13$  の2次元配列のSOMを用いた。入力の次元  $N$  は対象単語の数と同様、31であった。整列フェーズにおいては、学習総回数  $T$  を10,000に、学習率の初期値(0)を0.1に、そして、近傍の初期半径(0)を13に設定した。微調整フェーズにおいては、学習総回数  $T$  を100,000に、学習率の初期値(0)を0.01に、そして、近傍の初期半径(0)を7に設定した。

【 0 0 5 5 】

結果：図4は単語アライメントの意味マップの説明図である。図4において、前記3)の(目的)に挙げた対訳文への単語アライメントの意味マップを示している。但し、単語の前に「J」がついているのが日本語文の日本語であり、「C」がついているのがその訳文の中の中国語である。この意味マップから、日本語を中心にそれぞれの日本語と一番距離の近い中国語を取り出すことにより、以下の表1に示す単語間のアライメント結果が得られる。

40

【 0 0 5 6 】

表1：意味マップから得られるアライメントの結果

日本語	中国語	正解
J:経営	C:経営者	-
J:トップ	C:停留	C:最高
J:が	C:在	-
J:低	C:低速	C:低速
J:成長	C:停留	C:増長
J:時代	C:時代	C:時代
J:定着	C:増長	C:停留
J:を	C:。	-
J:実感	C:深感	C:深感
J:して	C:在	-
J:いる	C:可以	-
J:こと	C:看出	-
J:を	C:。	-
J:うかがわ	C:看出	C:看出
J:せた	C:可以	C:可以
J:。	C:。	C:。

10

20

30

## 【0057】

上記表1の結果は、図4の意味マップから一番近い距離にあるもののみを選び出している。もし、二番目近いもしくは三番目近い単語なども用いれば、アライメントの結果として複数候補が得られる。但し、分かりやすくするために右側に正解のアライメントも示している。この表からは（J：低、C：低速）、（J：時代、C：時代）、（J：実感、C：深感）、（J：うかがわ、C：看出）、（J：せた、C：可以）、（J：。、C：。）が正しくアライメントされているのが分かる。このうち、（J：うかがわ、C：看出）、（J：せた、C：可以）に関しては、日本語と中国語の日本語訳語候補との表層表現が違うものである。その他のアライメント結果は厳密に言えばすべて間違っているが、この中にも興味深いものが存在する。

40

## 【0058】

例えば、「J：成長」は「C：停留」とアライメントされているが、意味マップをみると、二番目に近いのが実は「C：増長」である。つまり、二番目の候補を含めると、正解になる。同様に、「J：定着」と「J：トップ」はそれらの二番目候補がそれぞれ「C：停留」と「C：最高」になっていて正解である。また、（J：こと、C：看出）と（J：を、C：。）の間違ひは、そもそもそれらの日本語に対応する中国語が（訳文に現れ）なかったためであり、単語分割の不一致により生じる（J：経営、C：経営者）のような間違ひも含め、アライメント技術だけでは対応しきれない問題である。

## 【0059】

50

(主成分分析による単語アライメントの意味マップの説明)

図5は主成分分析による単語アライメントの意味マップの説明図である。主成分分析結果である図5とSOMを用いる図4とを比較すれば、主成分分析の結果が劣っていることがわかる。例えば、表層表現の違う(「J:うかがわ、C:看出」)が得られていないし、「J:成長」に関しては、二番目の候補をいれても正しくアライメントできない。そして、単語が偏ったりして全体の配置のバランスが悪く、意味マップの特徴である可視性や連続性に問題がある。また、階層クラスタリングも行って見たが、その結果はかなり自己組織化された意味マップの結果に似てはいるが、「J:うかがわ、C:看出」が得られていないなど、やや劣っている。そして、意味マップと違って、グループの中の単語間の距離が分からないため、二番目の候補などを得るのが簡単ではない。

10

【0060】

(ベースライン手法との比較の説明)

ベースライン手法は、自己組織化マップ部4を用いなくて意味的距離 $d_{ij}$ の値が最も近いものに対応付ける手法である。この結果は、以下の表2のアライメント結果が得られる。

【0061】

表2: ベースライン手法のアライメントの結果

日本語	中国語	正解
J:経営	C:経営者	-
J:トップ	C:経営者	C:最高
J:が	C:在	-
J:低	C:低速	C:低速
J:成長	C:時代	C:増長
J:時代	C:時代	C:時代
J:定着	C:深感	C:停留
J:を	C:在	-
J:実感	C:深感	C:深感
J:して	C:在	-
J:いる	C:可以	-
J:こと	C:深感	-
J:を	C:在	-
J:うかがわ	C:深感	C:看出
J:せた	C:可以	C:可以
J:。	C:。	C:。

20

30

40

【0062】

前記表1の意味マップの手法は、「J:成長」と「J:停留」を誤り、表2のベースライン手法では、「J:成長」と「J:停留」の他に「J:うかがわ」の対応づけも誤っている。すなわち、ベースラインの手方の方が一個余分に誤っている。小規模な実験ではある

50

が、この実験ではSOMを用いる意味マップの手法の方がベースラインよりも精度が高いことがわかる。

【0063】

5)まとめ

本発明は、意味マップを用いることによって、意味に基づくアプローチを目指した新しい単語アライメント手法を提案している。提案手法の有効性は小規模な実験によって確かめられた。今後は、客観的な数値評価を導入し既存手法との大規模な比較実験を行うとともに、既存手法との融合も含め実用レベルのアライメント技術の開発を行っていく予定である。

【0064】

このように、本発明は、二次元に可視化されているので2番目に近い単語を直ぐ見つけることができ、対応付けもすぐできる(翻訳事例を多くたくわえることにより、辞書に載っていないドメインや時期などに依存する訳語を自動獲得することができる。)

【0065】

なお、前記実施の形態では、日本語と中国語の対訳文の単語の対応付けについて説明したが、他の言語の対訳文の単語の対応付けに適用することもできる。

【0066】

(3): プログラムインストールの説明

データコーディング部1、データコーディング手段1a、翻訳辞書2を格納する手段、コーパスデータ3を格納する手段、SOM部4、自己組織化マップ手段4a等は、プログラムで構成でき、主制御部(CPU)が実行するものであり、主記憶に格納されているものである。このプログラムは、一般的な、コンピュータで処理されるものである。このコンピュータは、主制御部、主記憶、ファイル装置、表示装置、キーボード等の入力手段である入力装置などのハードウェアで構成されている。このコンピュータに、本発明のプログラムをインストールする。このインストールは、フロッピィ、光磁気ディスク等の可搬型の記録(記憶)媒体に、これらのプログラムを記憶させておき、コンピュータが備えている記録媒体に対して、アクセスするためのドライブ装置を介して、或いは、LAN等のネットワークを介して、コンピュータに設けられたファイル装置にインストールされる。そして、このファイル装置から処理に必要なプログラムステップを主記憶に読み出し、主制御部が実行するものである。

【0067】

【発明の効果】

以上説明したように、本発明によれば、次のような効果がある。

【0068】

(1): データコーディング手段で、入力された対訳文の一方の言語の単語をコーパスデータ中の前記入力された対訳文の一方の言語の単語及びその前後の単語である共起語と共起頻度で定義すると共に、前記入力された対訳文の他方の言語の単語を翻訳辞書を用いて一方の言語の訳語候補を求め、該求めた訳語候補から前記コーパスデータを利用して共起語と共起頻度で定義し、自己組織化マップ手段で、前記共起語と共起頻度で定義した入力された対訳文の単語から前記共起語と共起頻度により単語間の距離を求め、該求めたある単語との距離が小さい単語ほど、前記ある単語に対してこれら以外の単語より小さい距離に配置するようにして、前記入力された対訳文の単語の自動マップを行うため、二次元で可視化して正確な対応付けが自動ででき、また2番目に近い単語をすぐ見つけることができる。

【0069】

(2): データコーディング手段で、共起語としてコーパスデータ中の入力された対訳文の一方の言語の単語及びその前後1つずつの単語とするため、共起語の処理データ数を少なくすることができる。

【0070】

(3): コーパスデータとして一方の言語の一定量の文書データを格納する手段と、翻

10

20

30

40

50

訳辞書として他方の言語から一方の言語に翻訳する辞書を格納する手段と、入力された対訳文の一方の言語の単語を前記コーパスデータ中の前記入力された対訳文の一方の言語の単語及びその前後の単語である共起語と共起頻度で定義すると共に、前記入力された対訳文の他方の言語の単語を、前記訳辞書を用いて一方の言語の訳語候補を求め、該求めた訳語候補から前記コーパスデータを利用して共起語と共起頻度で定義するデータコーディング手段と、前記共起語と共起頻度で定義した入力された対訳文の単語から前記共起語と共起頻度により単語間の距離を求め、該求めたある単語との距離が小さい単語ほど、前記ある単語に対してこれら以外の単語より小さい距離に配置するようにして、前記入力された対訳文の単語の自動マップを行う自己組織化マップ手段として、コンピュータを機能させるためのプログラム又はプログラム記録したコンピュータ読取可能な記録媒体とするため、このプログラムをコンピュータにインストールすることで正確な対応付けが自動でできる対応付け装置を容易に提供することができる。

10

【図面の簡単な説明】

【図1】本発明の原理説明図である。

【図2】実施の形態における対応付け装置の説明図である。

【図3】実施の形態における対応付け処理フローチャートである。

【図4】実施の形態における単語アライメントの意味マップの説明図である。

【図5】実施の形態における主成分分析による単語アライメントの意味マップの説明図である。

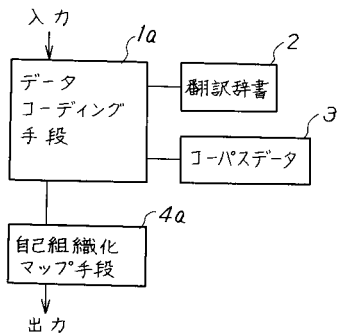
20

【符号の説明】

- 1 a データコーディング手段
- 2 訳辞書
- 3 コーパスデータ
- 4 a 自己組織化マップ手段

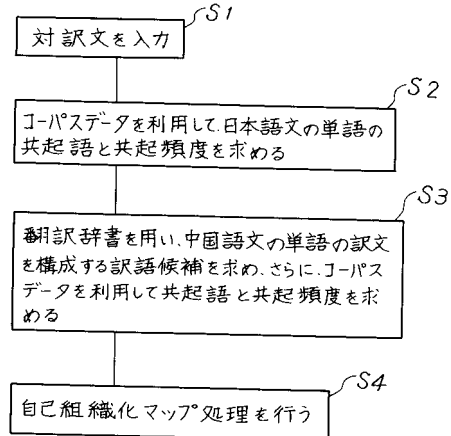
【図1】

本発明の原理説明図



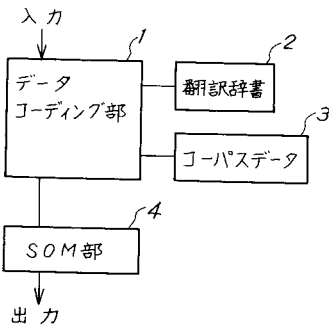
【図3】

対応付け処理フローチャート



【図2】

対応付け装置の説明図





---

フロントページの続き

- (72)発明者 村田 真樹  
東京都小金井市貫井北町4 - 2 - 1 独立行政法人通信総合研究所内
- (72)発明者 井佐原 均  
東京都小金井市貫井北町4 - 2 - 1 独立行政法人通信総合研究所内

審査官 和田 財太

- (56)参考文献 特開平10 - 063669 (JP, A)  
馬青・神崎享子・村田真樹・内元清貴・井佐原均, 日本語名詞の意味マップの自己組織化, 情報処理学会論文誌, 日本, 社団法人情報処理学会, 2001年10月15日, Vol. 42, No. 10, p. 2379 - p. 2391  
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, 日本, 社団法人情報処理学会, 1997年 4月15日, Vol. 38, No. 4, p. 719 - p. 726

- (58)調査した分野(Int.Cl., DB名)  
G06F 17/27-17/28