(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2004-362305
(P2004-362305A)
(43)公開日　平成16年12月24日(2004.12.24)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/28	G06F 17/28	U	5B091

審査請求 有　請求項の数 3　ＯＬ　（全 16 頁）

(21)出願番号　特願2003-160464(P2003-160464)
(22)出願日　平成15年6月5日(2003.6.5)

特許法第３０条第１項適用申請有り　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

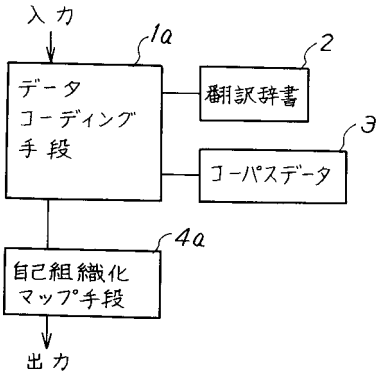
(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100103827
弁理士　平岡　憲一
(74)代理人　100097836
弁理士　福井　圀敞
(72)発明者　馬　青
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　張　玉潔
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

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)【要約】
【課題】意味に基づく単語の対応付けを自動で行うこと。
【解決手段】コーパスデータ３と、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は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周辺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は前記翻訳辞書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入力された対訳文の単語の自動マップを行う。
【選択図】　図１

本発明の原理説明図



【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は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周辺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は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】
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は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周辺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は、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　ｅｘａ

mple-based translation , Proceedings of the Seventh International Conference on Theor etical 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., J elinek, 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. 263-3 11, 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 max imum entropy models, COLING2002 , pp.1051- 1057, 2002.
【非特許文献6】
Kaji, H., Kida, Y., Morimoto Y.: Learning translation template s 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, ACL'93, pp. 23-30, 1993.
【非特許文献8】
Imamura, K.: Hierarchical phrase alignment harmonized with par sing, NLPRS2001, pp. 377-384, 2001.

【0005】

【発明が解決しようとする課題】

上記従来のものは、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は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周辺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は前記翻訳辞書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Ｓ１～Ｓ４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】
Ｓ１：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）。
【００１４】
Ｓ２：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】
Ｓ３：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。

【００１６】
Ｓ４：ＳＯＭ部４は、前記処理Ｓ２と処理Ｓ３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。
【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訳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。
【００２０】
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 $x = [\xi_1, \xi_2, \cdots, \xi_N]^T \in \Re^N$ ならば、個々のノード i はそれぞれ参照ベクトル $m_i = [\mu_{i1}, \mu_{i2}, \cdots, \mu_{iN}]^T \in \Re^N$ を持つものとする。

【００２４】
但し、参照ベクトルの要素μ$_{ij}$はノードｉと入力要素ξ$_j$　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】

$$c = argmin_i\{||x - m_i||\} \qquad \text{式 1}$$

【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】

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

【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】

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

但し、$r_c \in \Re^2$ と $r_i \in \Re^2$ 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、α（ｔ）　は学習率で、σ（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α（ｔ）　とσ（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、α（ｔ

）　とσ（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

（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　経済　仍　停留　在　低速　増長　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的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」にエントリーがない場合のみ「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のようになる。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
。：。

【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。

【００４０】

（データコーディングの説明）

日中対訳文が、次のように与えられたとする。

【００４１】

【数５】

$$J_1, J_2, \cdots, J_m$$

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

【００４２】

但し、J_i（ｉ＝１，…，ｍ）は日本語の文を構成する単語、C_i（ｉ＝１，…，ｎ）はその訳文を構成する単語、J_{ij}（ｉ＝１，…，ｎ，ｊ＝１，…，n_i）はC_iのｊ番目の訳語候補、n_i（$1 \leqq n_i \leqq t$）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（＝J_i）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。

【００４３】

【数６】

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

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される。

【００４５】

【数７】

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

但し、$a_i^{(j)}$ と $f_i^{(j)}$ は $J_{j1}, \cdots, J_{j,n_i}$ 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、α_j は 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 α_i と α_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。
【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（$x = 1, \cdots, \alpha_i$）。

【００５１】
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。
【００５２】
【数１０】

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

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

【００５３】
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ことで、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）。
【００５４】
ＳＯＭ：実験には１３×１３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α（０）　を０．１　に、そして、近傍の初期半径σ（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α（０）　を０．０１に、そして、近傍の初期半径σ（０）　を７　に設定した。
【００５５】
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。
【００５６】
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:。

【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。

【００５８】
例えば、「Ｊ：成長」は「Ｃ：停留」とアライメントされているが、意味マップをみてみると、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：。）の間違いは、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）のような間違いも含め、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】
（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みたが、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。

【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を用いないで意味的距離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:。

【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は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周辺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は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る、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は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周辺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は、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。

【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。

【符号の説明】

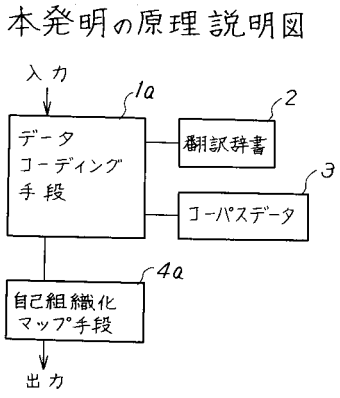
１ａ　データコーディング手段

２　翻訳辞書

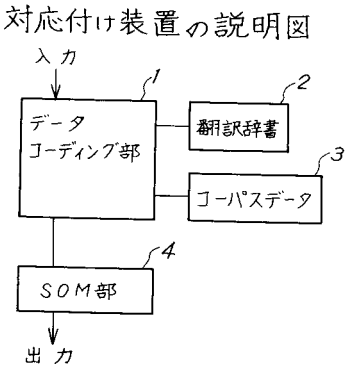
３　コーパスデータ

４ａ　自己組織化マップ手段

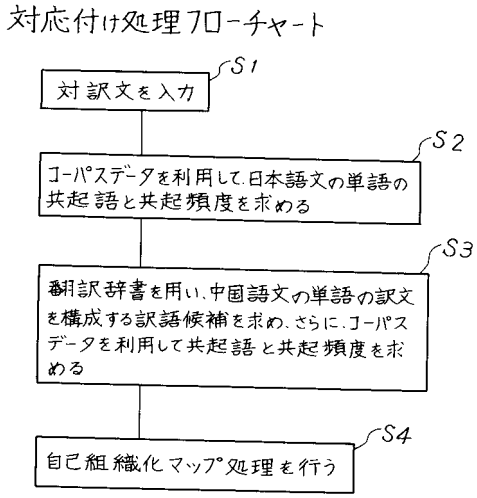
【図１】



【図２】

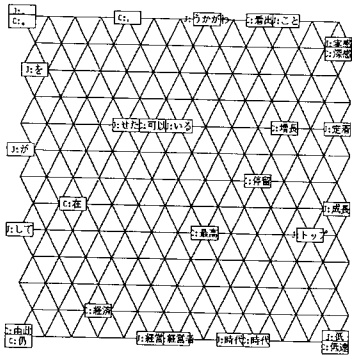


【図３】



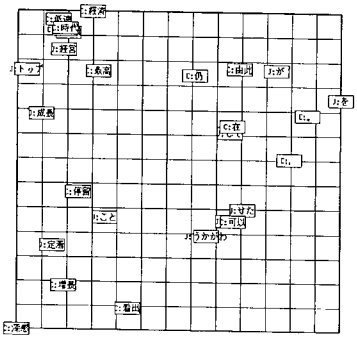
【図４】

単語アライメントの意味マップの説明図



【図５】

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



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

Ｆターム(参考) 5B091 AA07　BA03　BA14　CA05　CA12　CA24　CC05　CC16