対訳コーパスにおける単語アライメントの意味マップ

馬 青　張 玉潔　村田 真樹　井佐原 均
独立行政法人 通信総合研究所
{qma, yujie, murata, isahara}@crl.go.jp

1 はじめに

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 [1]．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 [2, 3]），それから，ダイナミックプログラミングを用いる手法 [5] や，最近では文脈情報を導入した統計手法 [4]，さらには構造化アライメント法（例えば，[6, 7, 8]）が挙げられる．いずれも，どちらかというと，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．

著者らはこれまで，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 [9, 10]．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かもしれない．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．

本稿では，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[1]．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1991-1998の8年分の毎日新聞から得られる．

[1] 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ので，提案手法も厳密に言えばまだ意味に基づく手法とは言えない．しかし，目指しているのは真の意味のマップであり，意味に基づくアライメント自体が重要なアイディアだと思うので，誤解を恐れず敢えて「意味に基づく」という言い方にした．

2 自己組織化神経回路網モデル

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（Self-organization Map，略して SOM [11]）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．

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．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 は以下のように選ばれる．

$$c = argmin_i\{||x - m_i||\} \tag{1}$$

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．

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

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

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

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．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．

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．

3 単語アライメントの意味マップの自己組織化

3.1 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．

(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 。
(中) 由此 可以 看出 ， 最高 经营者 深感 经济 仍 停留 在 低速 增长 时代 。

3.2 データ

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 [12] を用いて単語分割及び品詞の付与を行った．

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」[2]（愛知大学、大修館書店）より人手[3]で最大5個まで[4]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．

(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 经营者:経営者 深感:実感 经济: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 。:。

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．

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，1991-1998 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．次節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．

3.3 データコーディング

日中対訳文

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

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

が与えられたとする．但し，J_i $(i = 1, \cdots, m)$ は日本語の文を構成する単語，C_i $(i = 1, \cdots, n)$ はその訳文を構成する単語，J_{ij} $(i = 1, \cdots, n, j = 1, \cdots, n_i)$ 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（本報告においては $t = 5$）である．日本語文の単語 w_i $(= J_i)$ は以下のように共起語情報のセットで定義される．

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

但し，$a_j^{(i)}$ と $f_j^{(i)}$ は J_i の共起語と正規化[5]された共起頻度で，α_i は J_i と共起する単語の数である．一方，中国語訳文の単語 w_j $(= C_j)$ は以下のように共起語情報のセットで定義される．

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

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．

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

[2]「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．

[3]電子化された日中辞書が存在していないため，現状では人手に頼らざるを得なかった．

[4]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．

[5]つまり，$\sum_{j=1}^{\alpha_i} f_j^{(i)} = 1$.

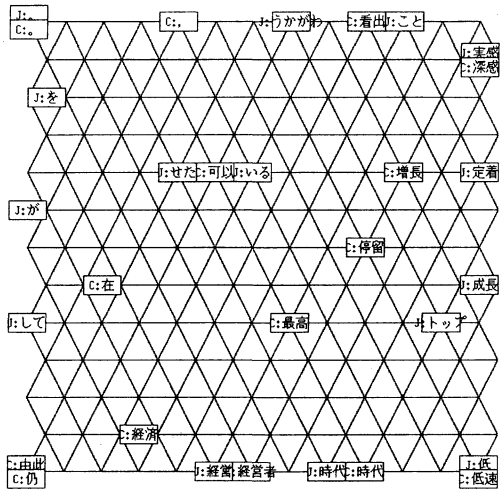

図 1: 単語アライメントの意味マップ

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．本稿では，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} を以下に示す頻度重み付け法で求める．

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

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．これらは以下の式で求められる．

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

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（$x = 1, \cdots, \alpha_i$）．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．

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

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．

4 実験結果

データ： 3.2 節に述べた対訳文（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 3.2 節に述べた方法で得た．3.1 節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「。」[6]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）．

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:经营者	-
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:增长
J:時代	C:时代	C:时代
J:定着	C:增长	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:。

SOM: 実験には 13×13 のの 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 $\alpha(0)$ を 0.1 に，そして，近傍の初期半径 $\sigma(0)$ 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 $\alpha(0)$ を 0.01 に，そして，近傍の初期半径 $\sigma(0)$ を 7 に設定した．

結果: 図 1は 3.1 節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す．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，表 1に示す単語間のアライメント結果が得られた[7]．但し，分かりやすくするため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：时代），（J：実感，C：深感），（J：

[6] 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．

[7] この結果は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．

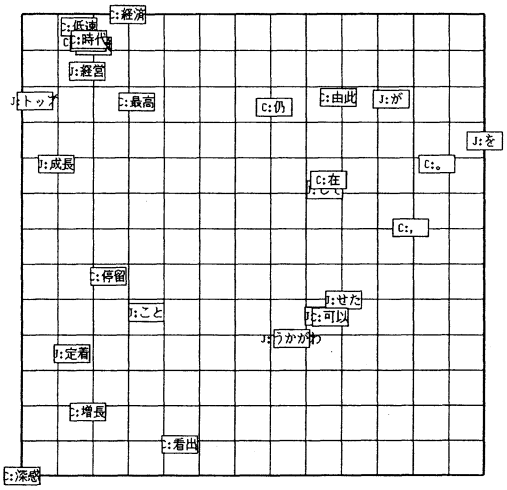

図 2: 主成分分析による単語アライメントの意味マップ

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．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．

図 2は主成分分析による単語アライメントの意味マップを示す．図 1と比較すれば，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．

5 結び

本稿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た．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．

参考文献

[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.

[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.

[3] 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.

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

[5] 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.

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

[7] Matsumoto, Y., Ishimoto, H, Utsuro, T.: Structural matching of parallel texts, *ACL'93*, pp. 23-30, 1993.

[8] Imamura, K.: Hierarchical phrase alignment harmonized with parsing, *NLPRS2001*, pp. 377-384, 2001.

[9] 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.

[10] 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.

[11] Kohonen, T.: Self-organizing maps, Springer, 2nd Edition, 1997.

[12] 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.