ブートストラップ法に基づく日英対訳コーパスからの対訳用語自動抽出

金 仁哲　　小川 泰弘　　外山 勝彦
名古屋大学大学院情報科学研究科
{aronkim, yasuhiro, toyama}@kl.i.is.nagoya-u.ac.jp

1　はじめに

対訳辞書の構築には膨大なコストがかかるため，計算機による構築手法が盛んに研究されている．文対応付き対訳コーパスから対訳辞書を構築するためには，辞書見出し語の決定と，その訳語の抽出という二つの課題がある．前者には，単言語コーパスから出現頻度と連接頻度に基づいて専門用語を抽出する手法があり，後者には，ワードアライメントが主に用いられる．

対訳表現の抽出手法として，北村ら [1] は，まず，対訳コーパスを形態素解析した後，一定の頻度を超える自立語を辞書見出し語として抽出した．その後，拡張した *Dice* 係数を用いて日英の自立語間の類似度を計算し，類似度の高い対訳語ペアを抽出した．このように，辞書見出し語の決定と訳語抽出を逐次的に行い，高い精度での対訳表現自動抽出に成功している．

それに対して我々は，それぞれを逐次的に行うのではなく，辞書見出し語の決定と訳語抽出を同時に行う手法を提案する．これにより，二つの手法を統一的なモデルで扱うことが可能となる．また，辞書見出し語の決定の際に，その訳語に関する情報も利用することが可能になるという利点もある．

なお，日本語などの分かち書きされない言語においては，北村ら [1] のように辞書見出し語の決定の際に形態素解析を利用すると，形態素解析辞書に登録されていない語は抽出できないという問題がある．そこで我々は，形態素解析を用いないで非分かち書き文から見出し語を抽出するブートストラップ法 Monaka [2] に着目し，これを拡張することにより対訳語ペアを抽出する．

以下，2 章で Monaka の概略を述べ，それを拡張した提案手法を 3 章で示す．4 章では提案手法を用いた実験について述べ，5 章は本稿のまとめである．

2　Monaka アルゴリズム

Monaka は，ブートストラップ法に基づいて，非分かち書き文からシードインスタンスと同じ意味カテゴリに属する表現を抽出する手法である (図 1)．

まず，入力として与えられたシードインスタンスからパターンを抽出する．Monaka では，インスタンス

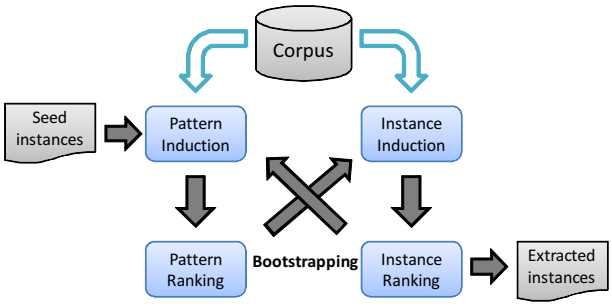

図 1: Monaka アルゴリズムの概略

に隣接する文字 n グラムをパターンとする．例えば，

S1： 委員及び臨時委員は学識経験のある者のうちから、内閣総理大臣が任命する。

という文からインスタンス “内閣総理大臣” に対応するパターンを抽出すると，“、#”，“ら、#”，“から、#” などの左側パターン及び “#が”，“#が任”，“#が任命” などの右側パターンが得られる．ここで，#はインスタンスのスロットを表す．

次に，パターンランキングのため，パターン p の信頼度 $r_\pi(p)$ を以下のように求める．

$$r_\pi(p) = \frac{1}{|I|} \sum_{i \in I} \frac{pmi(i,p)}{max_{pmi}} r_l(i) \ . \qquad (1)$$

ここで，I はインスタンスの集合である．$pmi(i,p)$ はインスタンス i とパターン p との自己相互情報量で，

$$pmi(i,p) = \log \frac{|i,p|}{|i,*||*,p|} \qquad (2)$$

によって計算する．ただし，$|i,p|$ は i と p の共起頻度を表し，$*$ はワイルドカードを表す．また，max_{pmi} は $pmi(*,p)$ の最大値である．

その後，インスタンスを抽出する．インスタンスは信頼度の高い上位 n 個のパターンを用いて抽出する．具体的には，パターンのスロットの位置に存在する文字 n グラムをコーパスから抽出する．例えば，文 (S1) にパターン “#が任命” を適用すると，“臣”，“大

臣”，...，“内閣総理大臣”，“、内閣総理大臣” などのインスタンスが抽出される．

最後に，インスタンスランキングを行う．インスタンス i の信頼度 $r_\iota(i)$ は，式 (1) と同様に，

$$r_\iota(i) = \frac{1}{|P|} \sum_{p \in P} \frac{pmi(i,p)}{max_{pmi}} r_\pi(p) \qquad (3)$$

により求める．ここで，P はパターンの集合である．

このようにパターンを定義することにより，分かち書きに頼らない抽出が可能となる．しかし，上述のように，正しく分かち書きされていないインスタンスが大量に抽出されてしまう．そのため，Monaka では，「信頼度の高いインスタンスは，信頼度の高い右側文脈と左側文脈に挟まれなければならない」という**両側隣接制約**を導入した．具体的には，各インスタンスに対して左側信頼度 r_l と右側信頼度 r_r を求める．左側信頼度 r_l は，式 (3) の P として左側パターンのみを用いて計算する信頼度である．右側信頼度 r_r も同様である．インスタンスの信頼度 r_ι は r_l と r_r の一般化平均を用いて，

$$r_\iota(i) = \sqrt[m]{\frac{1}{2}(r_l(i)^m + r_r(i)^m)} \qquad (4)$$

として求める．一般化平均は，算術平均や幾何平均など各種平均の一般化であり，m によって両側隣接制約の強さを調節する．例えば，m を 0 に近い値にすることにより，r_l と r_r の両方の信頼度が高いときにのみ r_ι が高くなるという制約を表現することができる．これにより，Monaka では分かち書きの正しいインスタンスを高い精度で抽出することに成功している．

Monaka では，抽出されたインスタンスをシードとして追加して上述のステップを繰り返すというブートストラッピングにより，インスタンスを増やしていく．

3 提案手法

本節では，前節において紹介した Monaka を拡張して，日英対訳コーパスから対訳用語を自動抽出する b-Monaka を提案する．b-Monaka は対象言語が分かち書き言語であるかどうかに関係なく，形態素解析を用いずに対訳表現を抽出することを可能にする．

なお，北村ら [1] の手法では，頻度の高い表現を抽出しているが，b-Monaka では，与えられたシードと同じ意味カテゴリに属する対訳表現を抽出する．

3.1 パターンとインスタンスの拡張

Monaka では単言語コーパスを対象とするが，b-Monaka では文対応付き対訳コーパスから対訳表現を抽出する．そのため，b-Monaka ではパターン p 及びインスタンス i を以下のように拡張してペアとする．

$$p = [p_j, p_e] \ , \ i = [i_j, i_e] \ .$$

ここで，p_j と p_e はそれぞれ日本語パターンと英語パターンであり，i_j と i_e はそれぞれ日本語インスタンスと英語インスタンスである．

3.2 パターン抽出の拡張

b-Monaka では対訳コーパス用に，Monaka のパターン抽出を以下のように拡張する．例として，

S2: ... 学識経験のある者のうちから、内閣総理大臣が任命する。/ ... members shall be appointed by the Prime Minister from among persons ...

という対訳文からのパターン抽出を考える．

まず，日本語文から日本語インスタンスに隣接する文字 n グラムを日本語パターンとして抽出し，それらの集合を P_j とする．同様に，英語文から英語インスタンスに隣接する単語 n グラムを英語パターンとして抽出し，それらの集合を P_e とする．次に，直積集合 $P = P_j \quad P_e$ を求める．P の要素が b-Monaka におけるパターンである．

例えば，文 (S2) からインスタンス [“内閣総理大臣”,“Prime Minister”] に対応するパターンとして [“ちから、#”,“# from among”] や [“#が任命する”,“appointed by the #”] などが抽出される．

3.3 インスタンス抽出の拡張

3.2 節と同様に，b-Monaka におけるインスタンス抽出を以下のように拡張する．

まず，日本語文から日本語パターンに基づいて日本語インスタンスを抽出し，その集合を I_j とする．次に，英語文から英語パターンに基づいて英語インスタンスを抽出し，その集合を I_e とする．次に，直積集合 $I = I_j \quad I_e$ を求める．I の要素が b-Monaka におけるインスタンスである．

例えば，文 (S2) にパターン [“#が任命する”,“appointed by the #”] を適用すると [“大臣”,“Prime Minister”]，[“総理大臣”,“Prime Minister from”]，[“内閣総理大臣”,“Prime Minister”] などのインスタンスが抽出される．

3.4 両側隣接制約の拡張

Monaka では，非分かち書きの文から形態素解析などを用いずに分かち書きの正しいインスタンスを抽出するため，両側隣接制約を導入している．

本手法では，両側隣接制約をそのまま適用できないため，両側隣接制約を拡張した．まず，それぞれの言語において両側隣接制約を掛ける．その後,「信頼度の高いインスタンスは，信頼度が共に高い日本語インスタンスと英語インスタンスから構成されなければならない」という**両言語制約**を与える．すなわち，日本語インスタンス i_j および英語インスタンス i_e に対して式 (4) を用いて求めた信頼度をそれぞれ $r(i_j)$，$r(i_e)$ とし，インスタンス全体の信頼度 r_ι は，一般化平均を用いて，

$$r_\iota(i) = \sqrt[m]{\frac{1}{2}(r(i_j)^m + r(i_e)^m)} \qquad (5)$$

と定義する．$r(i_j)$ と $r(i_e)$ を両側隣接制約式 (4) で置き換えて展開すると

$$r_\iota(i) = \sqrt[m]{\frac{1}{4}(r_l(i_j)^m + r_r(i_j)^m + r_l(i_e)^m + r_r(i_e)^m)} \tag{6}$$

となる．

Monaka では，インスタンスの信頼度はこのように一般化平均を用いて定式化しているが，実際にはパラメータ m が 0 に近いときにそのが性能が高い [2]．ゆえに，ここでは式を単純化し，m が 0 となる場合に相当する相乗平均を用いることにする．よって，インスタンスの信頼度は以下のようになる．

$$r_\iota(i) = \sqrt[4]{r_l(i_j)\ \ r_r(i_j)\ \ r_l(i_e)\ \ r_r(i_e)}\ . \tag{7}$$

4 実験

本節では，日英法令対訳コーパスに提案手法である b-Monaka を適用し，その性能について検討する．

4.1 実験条件

コーパス： 文対応付き日英法令コーパス 45,376 文 (法令 99 本) を用いる．

シードインスタンス： 職名を指す以下の五つのインスタンスをシードとして，最初に与える．

[“農林水産大臣”,“Minister of Agriculture, Forestry and Fisheries”], [“防衛大臣”,“Minister of Defense”], [“法務大臣”,“Minister of Justice”], [“厚生労働大臣”,“Minister of Health, Labour and Welfare”], [“内閣総理大臣”,“Prime Minister”].

各種パラメータ： インスタンスは毎回の繰り返しで 5 個ずつ抽出する．パターンは最初の繰り返しでは 100 個を抽出し，それ以降毎回の繰り返しで 50 個ずつ増やす．パターン抽出では，日本語文における文字 n グラム範囲と英語文における単語 n グラム範囲を 2 n 6 とする．インスタンス抽出では，日本語文における文字 n グラム範囲を 2 n 10 とし，英語文における単語 n グラム範囲を 1 n 10 とする．また，繰り返しは 10 回行う．

4.2 評価

実験により抽出された全 50 個のインスタンスを図 2 に示す．繰り返し回数ごとに {} で囲み，{} 内の 5 個のインスタンスは信頼度の高い順に並べている．

インスタンスは以下の三つの観点から評価した．その結果を表 1 に示す．

1. 抽出されたインスタンスの分かち書きは正しいか．
2. 日英インスタンスのペアは対訳として適切か．
3. シードと同じ意味カテゴリに属しているか．

表 1: 抽出インスタンスに対する評価

繰り返し回数	分かち書き		対訳	意味カテゴリ
	日本語	英語		
1	5	5	5 (0)	5 (0)
2	5	5	3 (2)	2 (2)
3	5	5	3 (2)	2 (2)
4	4	5	3 (1)	2 (1)
5	4	5	4 (0)	4 (0)
6	5	5	2 (3)	2 (3)
7	5	5	5 (0)	5 (0)
8	4	5	3 (0)	3 (0)
9	5	5	4 (1)	2 (1)
10	5	5	2 (2)	1 (2)
合計	47	50	34(11)	28(11)
精度	94.0%	100.0%	68.0% (90.0%)	56.0% (78.0%)

ここで，1. の分かち書きの正しさは主に日本語インスタンスに対する評価となる．ただし，英語インスタンスにおいても，“Minister of Health, Labour and Welfare,” のように最後に余分なコンマが付いているものが 3 個抽出された．これは英語の単語 n グラムの作成法に起因するものであり，b-Monaka のアルゴリズムに起因するものではないため，今回は正解とした．図 2 において下線が引いてあるインスタンスが，分かち書きの正しくないインスタンスである．

2. の対訳としての適切さは北村ら [1] と同様に，以下の基準で判定した．

正解： 対訳表現として適切なもの．例えば， [“内閣”,“Cabinet”] などは正解である．

半正解： 正解以外で，日英の各表現中に対応関係にある単語の組が少なくとも一つ存在するもの．例えば，[“法務大臣”,“Minister”] は半正解である．

不正解： 日英の各表現中に，対応関係にある単語の組が存在しないもの．図 2 では太文字で示した．

表 1 における対訳欄の数値は分かち書きが正しく，かつ対訳としても正解となるインスタンスの個数である．また括弧内は半正解のインスタンスの個数である．

3. の意味カテゴリに関しては，シードと近い意味をもつ職名や機関名を正解とした．シードと遠い意味と判定したインスタンスには図 2 において * を付与した．表 1 における意味カテゴリ欄の数値は，分かち書きが正しく，対訳としても正解であり，なおかつシードと近い意味をもつインスタンスの個数である．また，括弧内は対訳に関して半正解とされたインスタンスのうち，シードと近い意味をもつものの数である．今回では，半正解のインスタンスはすべてシードと近い意味をもっていた．

4.3 考察

本手法は形態素解析を用いなくても，94.0%の精度で正しく分かち書きされた日本語の見出し語を獲得で

{[“経済産業大臣”,“Minister of Economy, Trade and Industry”] [“国土交通大臣”,“Minister of Land, Infrastructure, Transport and Tourism”] [“経済産業局長”,“Director of Regional Bureau of Economy, Trade and Industry”] [“主務大臣”,“competent minister”] [“委員会”,“Commission”]} {[“厚生労働大臣”,“Minister”] [“公正取引委員会”,“Fair Trade Commission”] [“都道府県知事”,“prefectural governor”] [“公正取引委員会”,“Commission”] [“委託者保護基金”,“Consignor Protection Fund”]*} {[“経済産業大臣”,“Minister”] [“環境大臣”,“Minister of the Environment”] [“厚生労働大臣”,“Health, Labour and Welfare”] [“機構”,“Organization”]* [“厚生労働大臣”,“Minister of Health, Labor and Welfare”]} {[“、農林水産大臣”,“Minister of Agriculture, Forestry and Fisheries”] [“裁判所”,“court”] [“使用者”,“employer”]* [“経済産業大臣”,“Minister of METI”] [“行政官庁”,“relevant government agency”]} {[“、法務大臣”,“Minister of Justice”] [“検疫所長”,“quarantine station chief”] [“総務大臣”,“Minister of Internal A airs and Communications”] [“都道府県”,“prefectural government”] [“保健所長”,“director of a health center”]} {[“法務大臣”,“Minister”] [“環境大臣”,“Environment”] [“情報管理センター”,“Information Management Entity”] [“文部科学大臣”,“Minister”] [“最高裁判所”,“Supreme Court”]} {[“行政庁”,“administrative agency”] [“児童相談所長”,“child guidance center’s director”] [“商品取引員”,“Futures Commission Merchant”] [“協会”,“Institute”] [“商品取引所”,“Commodity Exchange”]} {[“、都道府県知事”,“prefectural governor”] [“内閣”,“Cabinet”] **[“経済産業大臣”,“Institute”]** [“家庭裁判所”,“family court”] [“入国警備官”,“immigration control o cer”]} {[“厚生労働大臣”,“Minister of Health, Labour and Welfare,”] [“申請書を厚生労働大臣”,“Minister of Health, Labour and Welfare,”] [“会員等”,“Member, etc.”] [“前条”,“preceding Article”]* [“資金管理業務”,“Deposit Management Business”]*} **{[“調査”,“Minister”]** [“農林水産大臣”,“Minister of Agriculture, Forestry and Fisheries,”] [“次に掲げる事項”,“matters”]* [“主務省令”,“competent minister”] [“自動車製造業者等”,“Vehicle Manufacturers, etc.”]}

図 2: b-Monaka により抽出されたインスタンス全 50 個

きた．対訳表現に関しては，半正解を含めれば 90.0%の精度で抽出できている．また，抽出されたインスタンスが与えられたシードに近い意味をもつかという点まで含めても，56.0%(半正解を含めれば 78.0%) の精度を達成した．

また，図 2 において “経済産業大臣” の対訳語として，“Minister of Economy, Trade and Industry” だけでなく，繰り返しの 4 回目において “Minister of METI” も抽出されており，複数の対訳語をもつ見出し語にも対応できることが分かる．

しかし，これには問題点もある．図 2 では繰り返しの 8 回目において [“経済産業大臣”, “Institute”] という誤った対訳も抽出された．これはコーパス中において “経済産業大臣” は “Institute” とも一定の頻度で共起し，さらに “Institute” が “Minister of Economy, Trade and Industry” と類似の文脈で出現するからである．このため，この対訳の信頼度も高くなり抽出された．この問題を解決するためには，片方のインスタンスが既にシードに含まれる場合，その信頼度を適当に減少させるなどの必要がある．

日本語の分かち書きに失敗した場合について検討する．Monaka においては，“、法務大臣” のようなインスタンスは左側信頼度が低いため，両側隣接制約により除外される．しかし，本稿で拡張した式 (7) では四つの信頼度の相乗平均を計算する．そのため，[“、法務大臣”, “Minister of Justice”] のようなインスタンスにおいては，$r_l(i_j)$ は低い値となるが，シード [“法務大臣”, “Minister of Justice”] の存在により，残りの三つの信頼度が高くなったため，結果的に r_ι が他のインスタンスよりも高くなってしまった．これを解決するためには，両側隣接制約を再検討する必要がある．

5 おわりに

本稿では，シードと同じ意味カテゴリに属する対訳表現の抽出手法である b-Monaka を提案した．本手法は，従来と異なり，辞書見出し語の決定と訳語の抽出を同時に行う新しい手法である．これは，対訳表現抽出技術の発想転換に貢献した．

b-Monaka の実装においては，計算量が膨大になるためパターンやインスタンスを足切りするなどの近似を行っているが，それによる悪影響も確認されている．よって，実装方法を改善し，b-Monaka の計算をより効率的に実現させる予定である．

それに加え，本手法を中英及び韓英など他の対訳コーパスに適用し，その有効性を検証する．

また，Monaka の改良版でより精度の高い g-Monaka [3] を利用することによる性能向上も目指す．

参考文献

[1] 北村美穂子，松本裕治: 対訳コーパスを利用した対訳表現の自動抽出，情報処理学会論文誌，Vol. 38, No. 4, pp. 727-736 (1997).

[2] Masato Hagiwara, Yasuhiro Ogawa, Katsuhiko Toyama: Bootstrapping-based Extraction of Dictionary Terms from Unsegmented Legal Text. *Proc. of the Second International Workshop on Juris-Informatics*, pp. 63-72 (2008).

[3] 萩原正人，小川泰弘，外山勝彦: グラフカーネルに基づく非分かち書き文からの意味的語彙カテゴリの抽出，言語処理学会第 15 回年次大会発表論文集，pp. 697-700 (2009).