焼きなまし法を用いた対訳単語対抽出

3N－7

菊地弘晶 佐藤健吾 斎藤博昭 中西正和*

1 はじめに

近年，電子化された大規模なコーパス (言語現象の調査，統計処理などを行う為に収集された言語データ) が出現している．コーパスを統計的な手法を用いて表層的に解析し，自然言語処理に有用な言語知識を得ようとする技術は目覚しく進歩しており，情報社会の国際化による多言語を同時に扱う必要性から，高精度な機械翻訳システムが必要とされている．その翻訳品質は，そのシステムが用いる文法や対訳辞書の性質に依存する．質の高い翻訳辞書をユーザの対象となる分野の模範的な翻訳文例集から自動的に作成し，使用することにより，翻訳品質を向上させることができると考えられる．

本研究では，翻訳辞書を自動生成する際，焼きなまし法 (simulated annealing) を用いて英日の両コーパスから対訳単語対の抽出を目指す．本稿では日本語英語間における焼きなまし法を用いた抽出についてその手法と今後の展望について述べる．

2 基本的な Word-to-Word モデル

本研究においては，対訳関係を抽出する際，対象となる単語が 1 対 1 で対応することを前提とする翻訳モデルである Word-to-Word モデルを用いる．

本研究においてこのモデルを用いる際には，パラメータとしてこのモデルにおける対訳関係が真である確率 λ^+，同様に偽である確率 λ^-，単語の組 (u,v) が対訳関係である尤度関数 $L(u,v)$ を用いる．$L(u,v)$ は共起回数に比例し，周辺回数に反比例するような初期値が与えられる．この翻訳モデルは以下のアルゴリズムによって実現される．

1. $L(u,v)$,Linking Algorithm のより対訳テキスト中の単語間のリンクを探す．
2. 1. で求めたリンクを用いて λ^+,λ^-, $L(u,v)$ を再推定する．
3. リンクが収束するまで 1 ～ 3 を繰り返す．

3 Competitive Linking Algorithm

前述の Word-to-Word モデルを用いる際には Linking Algorithm が必要となる．　本研究においては,Competitive Linking Algorithm[1] を用いる．両言語間の単語の組である u_k,v_k について，共起する確率が偶然による期待値よりも高ければ，対訳関係にあると推測できることは自明である．しかし，u_k と u_{k+1} が同言語中でよく共起すれば，u_{k+1} と v_k も偶然による期待値よりも多く共起することが推測される．対訳関係にある共起は direct association, 対訳関係にないが期待値よりも高い確率で起こる共起は indirect association と呼ばれる．Competitive Linking Algorithm は indirect association にも有効で，indirect association が direct association よりも関係が弱いことを利用する．尤度関数 $L(u,v)$ の値を比較し，indirect association を対訳関係の候補からはずすことを可能にする．そのアルゴリズムは以下のように記述できる．

1. $L(u,v) < 1$ である組 (u,v) を破棄する．
2. $L(u,v)$ を降順に並べる．
3. 最大の $L(u,v)$ を持つ組 (u,v) を得る．
4. 対訳テキスト中の全ての組 (u,v) にリンクを張る．
5. 1 対 1 という仮定により，リンクが張られた単語をテキストから除く．
6. $L(u,v)$ が存在する組 (u,v) があれば，3 から繰り返す．

*Creating a Dictionary using Simulated Annealing . Hiroaki Kikuchi . Department Computer Science,Keio University 3-14-1 Hiyoshi,Kohoku-ku,Yokohama,Kanagawa 223-8522,Japan .

4 尤度関数 $L(u,v)$

Word-to-Word モデルにおいては, 単語間のリンクの最尤推定を行うための尤度関数が必要となる. 本研究では, パラメータ λ^+, λ^-, 尤度関数 $L(u,v)$ を用いることは既に述べた. $L(u,v)$ を決定する為には, u と v のリンクの個数 $k(u,v)$ が, u と v の共起回数 $n(u,v)$ と u と v が対訳関係である確率 $p(u,v)$ による二項分布に従うことを利用する. これは, リンクが互いに独立であることによる. この関係から λ^+, λ^- を推定し, n 回の共起から k 個の正しいリンクが観測される確率 $B(k_{(u,v)} \mid n_{(u,v)}, \lambda^+)$, n 回の共起から k 個の正しくないリンクが観測される確率 $B(k_{(u,v)} \mid n_{(u,v)}, \lambda^-)$ が導かれる. これを用いて $L(u,v)$ を次のように定める.

$$L(u,v) = \frac{B(k_{(u,v)} \mid n_{(u,v)}, \lambda^+)}{B(k_{(u,v)} \mid n_{(u,v)}, \lambda^-)}$$

5 焼きなまし法を用いた対訳単語対の抽出

本研究においては前述の Word-to-Word モデルに基づいて以下の過程を経て対訳辞書の作成を行う.

1. コーパスから名詞を抽出する.
2. 双方の各単語の組にそれぞれ尤度関数の初期値を与える.
3. Competitive Linking Algorithm を用いて, 単語間にリンクを張る.
4. 焼きなまし法を用いて尤度関数の再推定を行う.
5. 3.4 の操作をリンクが収束するまで繰り返す.
6. 生き残ったリンクを対訳関係にある単語対として抽出する.

6 本手法による抽出例

本節では本手法を用いて行った抽出の例を述べる. 本手法の効果を確かめるために実験を行った. また, 比較の対象として, 単純に共起回数の多い単語対から対訳単語対として抽出する実験を予備実験として行った.

この 2 回の実験においては, 対訳コーパスから名詞のみを形態素解析ツール [3][2] を用いて抽出して対訳関係の抽出を行った. 文同士の対訳関係の対応が前提となっている対訳コーパス (約 30000 文) 用いて対訳単語対の抽出を行い, その効果を検証した.

表 1: 抽出実験の結果

	得られた単語対	適合率
予備実験	3317	36.97%
本手法による実験	1034	72.63%

7 おわりに

本稿では, 組合せ最適化問題の解法である焼きなまし法を用いて対訳単語対を抽出する手法を提案した. 本研究の特徴を以下に示す.

- 対象となる双方の言語間での単語の対応を 1 対 1 に仮定している
- 文対文の 1 対 1 の対応は必ずしも必要でなく, コーパスがある単位同士で対訳関係にあれば, 対訳単語対は抽出できる.
- 焼きなまし法を用いることで, パラメータに基づいて対訳関係を推定する際, 局所解に陥らない最尤推定を行うことができる.

また, さらに精度の高い対訳単語対の抽出を実現するために次のような課題を解決する必要があると考えている.

- 複数の単語が連なって 1 つの意味をなすような連語表現の抽出も可能であるように改良する.
- 共起関係のみではなく最大エントロピーなどの他の統計量を評価値に用いる.
- 遺伝的アルゴリズムなどの組合せ最適化問題の解法を用いての精度の高い最尤推定を実現する.

また, さらなる目標として今後は Linking Algorithm に組合せ最適化問題の解法を直接適用した対訳関係抽出の手法の実現を目指していきたいと考えている.

参考文献

[1] I. Dan.Melamed. A Word-to-Word Model of Translational Equivalence. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997.

[2] 松本裕治, 北内啓, 山下達雄, 平野善隆, 今修一, 今村友明. 日本語形態素解析システム『茶筌』 version 1.5 使用説明書.Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学,1997.

[3] 佐藤健吾,link grammer の概念を採り入れた d-bigram 確率モデル. 慶應義塾大学, 修士論文,1997.