情 報 学 基 礎　68－1
自 然 言 語 処 理 151－1
（２００２．９．１７）

情報検索システムを利用した日英対訳語推定

鈴木健二† 梅村恭司†
†豊橋技術科学大学情報工学系

要約

近年、非対訳コーパスを対象とした対訳語推定が研究されている。非対訳コーパスを用いた対訳語推定の問題のひとつは対訳語を含む部分を特定することである。本稿では、情報検索システムと言語横断情報検索システムの検索結果から対訳語を推定する手法について述べる。言語横断情報検索システムは小規模の基本辞書を用いる。実験として 100 件の名詞について対訳語抽出実験を行った。

Term translation estimation using information retrieval system

Kenji Suzuki† Kyoji Umemura†
†Deartment of Information Engineering, Toyohashi University of Technology

Abstract

In recent years, there are some studies about term extraction from non-parallel corpus. One problem with term extraction from non-parallel corpus is how to distinguish documents which contain right term translations. In this paper, we use information retrieval system to get documents which contain term translations. Our method uses information retrieval system, cross language information retrieval system, and basic dictionary. We have conducted term extraction experiment about 100 nouns.

1 はじめに

コーパスから対訳語を自動的に抽出するという試みはこれまでも良く行われてきた。コーパスから自動的に抽出することの利点は、辞書に新しい語彙を追加するためのコストを軽減できることである。対訳辞書は計算機言語処理における、基本的な情報であるといえる。例えば、機械翻訳や言語横断情報検索にとって、対訳辞書は必要不可欠な要素であり、その重要度はますます増加しているといえる。

これまで、人手によって対訳関係を付加された対訳コーパスからの対訳語抽出が研究されてきた [1]。しかし、対訳コーパスの得られる分野が限られていること、対訳コーパスの量が十分とはいえないことから、対訳コーパスではなく非対訳コーパスを対象とした対訳語抽出法も研究されている [2][3][4][5]。

本稿における非対訳コーパスとは、同一分野で対訳情報が付加されていないコンパラブルコーパスや対応する文書の存在が保証されないノイジーパラレルコーパス、WWW などを指す。

非対訳コーパスを用いた対訳語抽出を行う際に問題となることは、原言語と目的言語での対訳部分を推定することである。しかし、非対訳コーパスでは必ずしも文書レベルでの対訳が存在しない。本稿では情報検索システムの出力を非常に粗い対訳関係とみなし、対訳関係がないことを補うような対訳語抽出モデルを提案する。このモデルは原言語の情報検

索システムによって一種の質問文拡張を行い、原言語 - 目的言語での言語横断情報検索システムで拡張した質問文を検索し得られた結果から対訳語を推定するモデルの一つである。このモデルを実装し、コンパラブルコーパスから対訳語抽出実験を行った。また ϕ 相関係数を用いたベースラインシステムを作成し、同様の実験を行い、両者の結果を比較した。

本稿の構成は、2節で本稿における対訳語抽出問題を定義し、この問題を解決するモデルの概要を述べる。3節では、実験の条件およびベースラインシステムとの比較結果について述べる。4節では評価結果をふまえて考察を行う。

2 提案法

2.1 対訳語抽出問題の定義

本稿での対訳語抽出問題とは、原言語の用語を与えるとき、目的言語で対応する用語を得る問題とする。本稿では日英対訳語抽出を行ったので、原言語は日本語、目的言語は英語である。また、本稿における用語とは、単名詞もしくは複合名詞とする。

2.2 基本的なアイデア

本稿の基本的なアイデアは、情報検索によって得られた文書集合は、クエリーを中心概念としているはずであるという仮定である。つまり、情報検索処理や言語横断情報検索処理によって検索結果にノイズが混入しても、検索のトピックス自体は変化しないという考えに基づいているともいえる。したがって、情報検索システムと言語横断情報検索システムを用いて内容の似通った文書集合を取得し、情報検索処理で混入したノイズを除去することで、対訳語が得られると考えた。

2.3 モデルの要素

対訳語抽出問題を解決するために情報検索システムを利用したモデルの要素を述べる。モデルの要素は、質問と正解、コーパス、情報検索システム、訳語抽出システムの4つに分けることが出来る。以下でそれぞれについてのべる。

原言語 s、目的言語 t をモデルの入出力言語とすると、原言語用語 $term_s$ は抽出したい用語であり、目的言語での用語 $term_t$ は、用語 $term_s$ の目的言語で対応する用語である。対訳関係は未知であるため、$term_t$ を得ることが目的となる。対訳語候補 $term_{st}$ は対訳語抽出を行った結果として得られる。対訳語候補 $term_{st}$ 中で $term_t$ が出来るだけ高順位であるようにしたい。

コーパスは原言語および目的言語の2つのコーパスを用いる。原言語コーパス C_s は原言語で記述された文書集合である。原言語コーパスは $|C_s|$ 件の文書が含まれているとする。同様に、目的言語コーパス C_t は目的言語で記述された文書集合である。目的言語コーパスは $|C_t|$ 件の文書が含まれているとする。

情報検索システムは、単言語情報検索システム IR と言語横断情報検索システム CLIR を用いる。単情報検索システム IR は原言語検索語 $query_s$ とコーパス C_s を入力とし、C_s から $query_s$ に関連する原言語文書集合 D_s を出力する。言語横断情報検索システム CLIR は原言語検索語 $query_s$、コーパス C_t、対訳辞書 $Dict_{st}$ を入力とし、$query_s$ に関連する目的言語文書集合 D_t を出力する。ここで、対訳辞書 $Dict_{st}$ は原言語 - 目的言語の単語対応を与える対訳辞書である。

訳語抽出システム TE は、文書集合 D とコーパス C を入力すると、それぞれの統計量を用いて、D に含まれる語に対してスコアをつけ、対訳語候補を出力する。

2.4 モデルの処理手順

前述のモデルの要素を用いて、対訳語抽出モデルの処理を説明する。モデルの処理は、原言語情報検索処理、原言語 - 目的言語情報検索処理、対訳語抽出処理の3つからなる。処理の概略を図 1に示す。

2.4.1 原言語の情報検索処理

最初の処理である原言語の情報検索処理では、原言語用語 $term_s$ に関連する語を得ることが目的である。そのために、情報検索システム IR を用いてコー

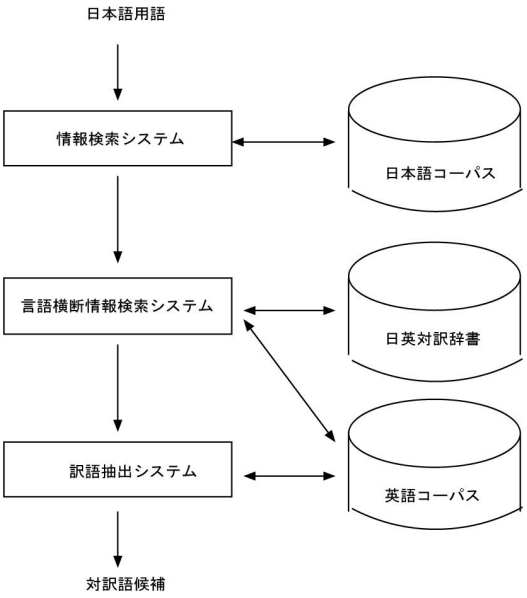

図 1: モデルの概略図

パス C_s から原言語用語 $term_s$ に関連する文書集合 D_s を取得し、文書集合 D_s 中の $term_s$ に関連する語を決定する。ここでの問題は、文書集合 D_s には情報検索処理によって関連しない文書がノイズとなって混入すること、また、コーパスの大きさは有限であるため、$term_s$ の関連語が全てコーパス中にあるとは限らない事である。

2.4.2 言語横断情報検索処理

次の処理は、原言語 - 目的言語での言語横断情報検索処理である。この処理では、目的言語用語 $term_t$ を含む文書集合を取得することが目的である。具体的には前段階で得られた関連語をもとに文書集合 D_s から検索語を生成し、言語横断情報検索を行う。言語横断情報検索処理によって得られた文書集合 D_t は、目的言語用語 $term_t$ を含んでいることが期待できる。ここでの問題は、一般に言語横断情報検索処理は同一言語での情報検索よりもノイズが混入しやすいことが知られている。

2.4.3 対訳語抽出処理

対訳語抽出処理では、言語横断情報検索の結果として得られた文書集合 D_t から、フィルタリングを行うことで対訳語を抽出する。本稿では、フィルタリングの方法として、文書集合 D_t と目的言語コーパス C_t の単語分布の違いを利用する。情報検索によってトピックス自体は変更しないと考えられるので、標本空間にのみよく現れる用語候補に高いスコアを与える。

w を訳語候補とし、$df(w, D_t)$ は文書集合 D_t に w が 1 回以上出現する文書数を表す。また、文書集合 D_t の文書数を $|D_t|$ とすると、文書集合 D_t の文書に訳語候補 w が現れる確率 $P_{D_t}(w)$ は

$$P_{D_t}(w) = \frac{df(w, D_t)}{|D_t|} \tag{1}$$

であり、同様に、目的言語コーパス Ct の文書に訳語候補 w が出現する確率 $P_{C_t}(w)$ は

$$P_{C_t}(w) = \frac{df(w, C_t)}{|C_t|} \tag{2}$$

である。$P_{D_t}(w)$、$P_{C_t}(w)$ をパラメータとして、スコア関数を式 3のように定義する。

$$score(w) = \frac{1}{1 + e^{\alpha P_{D_t}(w)}} \times \frac{1}{1 + e^{\beta P_{C_t}(w)}} \tag{3}$$

ここで、α と β は定数である。
今回は、実際のコーパスでの用語の分布から、$\alpha = -5.0$、$\beta = 100.0$ と定めた。式 3をグラフで描くと図 2のようになる。図 2から分かるように、$P_{D_t}(w)$ が高く、$P_{C_t}(w)$ が低いほど、検索結果に集中しているほど、スコアは高くなる。

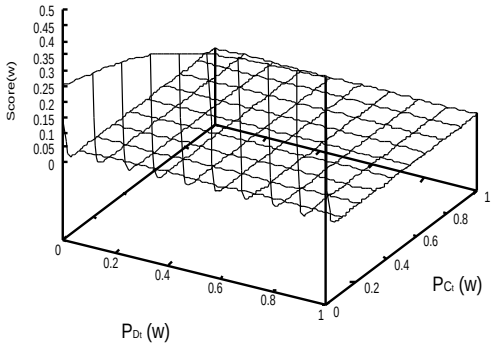

図 2: スコア関数とパラメータの関係

3 実験

原言語を日本語、目的言語を英語として日英対訳語抽出実験を行った。実験方法として、基本辞書に収録されていない用語を入力とし、どの程度の正解が得られるかを確かめた。また、ベースラインシステムとして、情報検索システムのかわりに、ϕ相関係数を用いて、文書集合D_tを取得した場合について実験した。また、基本辞書の語彙数を1/2,1/4にした辞書を作成し、精度の変化を確認した。

3.1 実験で使用したモデルの要素

3.1.1 コーパス

日本語コーパス、英語コーパスとしてNTCIR1テストコレクションからタイトルと要約以外の情報を取り除いたコーパスを機械的に生成して使用した。NTCIR1テストコレクションは日本語論文抄録約18万件と英語論文抄録約18万件の論文抄録集である。それぞれの論文抄録集を日本語コーパス、英語コーパスとして使用した。

3.1.2 情報検索システム

情報検索システムは$tf \cdot idf$尺度をベースとした情報検索システムを使用した。

3.1.3 言語横断情報検索システム

言語横断情報検索システムは$tf \cdot idf$尺度をベースとした、辞書ベースの言語横断情報検索システムを使用した。辞書ベースの言語横断情報検索システムは、日英辞書で原言語質問文中で辞書引きできた単語を目的言語質問文とし、情報検索を行うようなシステムである。

3.1.4 基本辞書

基本対訳辞書としてはEdict[6]を使用した。Edictは日本語と英語の基本的な対訳対を約7万対収録している。実験に際して、辞書に正解がない状況を想定するために、正解データ対をあらかじめ削除した。また、辞書の語彙数と精度の関係をみるために、辞書の語彙数を1/2,1/4にした辞書を生成した。削除する対訳対は、日本語コーパス、英語コーパスでの出現頻度を数え、出現頻度の少ないものから削除した。

3.2 正解データの選考基準

実験の評価を行うために、日本語と英語の用語対を正解データとして作成した。用語対には単名詞と復号名詞の両方が含まれる。正解データの作成は以下の手順で行った。

1. NTCIR1テストコレクションに付加されるキーワードから日本語と英語の対訳対になっているものを機械的に抽出する。
2. 対訳対の日本語用語と英語用語の単語頻度、文書頻度を数える。使用するコーパスでの単語頻度、文書頻度のいずれかが10以下ならその対訳対は除外する。
3. 残った対訳対からランダムに100件選択する。

3.3 スコアを計算する語の決定基準

対訳語抽出処理部でスコアを計算する語を言語横断情報検索で得られた文書集合より選択した。その基準は次のようなものである。

- $n < 5$までの単語n-gram
- 情報検索の結果として得られた文書集合内で、文書頻度が3以上の語
- 目的言語コーパスの文書に1回以上出現する確率P_{C_t}が0.5以上である語

最初の基準は、専門用語は通常5単語程度であるという経験則に基づく基準である。2番目の基準は、低頻度語の影響によるノイズを除去するための基準である。最後の基準は、ストップワードを除去するための基準である。

3.4 ベースラインシステム

比較対象として、情報検索システム、言語横断情報検索システムのかわりにϕ相関係数を基準として、文書集合D_tを以下の手順で取得するようなシステムについて同様の実験した。

1. 日本語コーパスをもとに ϕ 相関係数を用いて、日本語用語と関連する基本辞書中の対訳対を上位 20 件決定する。

2. 英語コーパス中の全文書に対して、 20 件の英語訳が含まれるかどうかを調べ、含まれるなら、その語の相関係数値をスコアとして加算する。

3. スコアの高い順に文書 100 件を取得する。ただし、スコアが相関値の上位 20 件の算術平均以下になったら足切りをおこなう。

以上の手順で取得した文書集合 D_t に対して、スコアを計算した。

3.5 実験結果

実験結果として、正しい対訳語が得られたケースについて、表 1に示す。

表 1: 対訳語抽出例

日本語用語	対訳語	順位
CASEツール	case tools	3
スケジューリング	scheduling	1
ニューラルネットワーク	neural networks	1
ファジー推論	fuzzy inference	1
周期境界	periodic boundary	3
片持梁	cantilever	1

訳語候補での順位が 1 位であるときの正解率と上位 10 位までに正解が含まれる場合の正解率を評価した。提案法とベースラインシステムの比較結果をそれぞれ表 2, 表 3に示す。表の Full は基本辞書そのままであることを表し、 Half,Quarter はそれぞれ基本辞書の 1/2,1/4 である事を表す。

表 2: 精度の比較結果 (1 位での正解率)

	Full	Half	Quarter
IR	16%	15%	15%
ϕ 相関係数	6%	6%	4%

表 3: 精度の比較結果 (10 位までの正解率)

	Full	Half	Quarter
IR	49%	46%	44%
ϕ 相関係数	43%	42%	37%

実験の結果、基本辞書の語彙数にかかわらず、全ての結果について情報検索システムを用いた方が良い結果を得られた。また、当初の予測通り、基本辞書の語彙数が減少すると精度も低下することが確認できた。

この理由としては、相関係数は語句レベルの関連性を表すが、情報検索システムでは、文書レベルの関連性に重点をおいているためであると思われる。文書レベルで見た場合、ある文書に表現されている意味の範囲は単語レベルでの意味の範囲よりもより特定されている。そのため、文書レベルの対応を近似であっても取ることは効果があると思われる。

得られた対訳語候補は、正解との関連性はあるものの、 1 位での精度は 16.3% であり、 10 位までの精度に比べて低い。これはこのフレームワークの限界を示すとも解釈できる。すなわち、トピックスから単語を一つに特定することが必ずしもも容易ではなく、一般には、トピックスから推定される単語は複数ある事実が、この結果となったと推定される。

4 関連研究

類似した研究としては、文献 [2] があげられる。ある語の前後に共起する語は、たとえ言語が異なっても類似しているという観点にもとづき、コンパラブルコーパスから独英対訳語抽出を行っている。この研究の報告では、 1 位での正解率が 70% 程度であることが報告されている。なお、この研究では複合語を対象としていない。

文献 [3][4] でも共起に基づく訳語抽出を行っている。[3] では複合語も対象にした研究をしている。この文献も基本的な考え方は同様であるが、情報検索のアプローチを用いていること、および、複合語に対する評価を行っていることが異なる。

文献 [5] では、基本辞書から対訳語候補を作成し、曖昧性解消によって対訳語らしさを決定する方法を提案している。曖昧性解消による方法は、計算量が比較的少なく、強力であるが、対訳語候補が基本辞書から生成できるものに限られてしまう。

情報検索システムを利用した対訳語抽出は、曖昧性解消による方法では抽出できない用語を抽出できる可能性があり、その点では研究する価値があると思われる。

5 今後の課題

実験の結果、本稿で提案するモデルだけで対訳辞書を作成すると考えた場合、精度が不十分であることがわかった。精度が不十分であるという問題にはいくつかの原因が考えられる。第一の原因は、対訳語抽出で使用している抽出尺度の精度である。現在は、言語横断情報検索システムから出力された文書集合と目的言語コーパスを比較して、文書集合に集中出現する語に高いスコアを与えている。しかし、現在使用している尺度が最良であるという理論的な保証はない。第二の原因として、統計的対訳抽出は目的言語コーパスおよび言語横断情報検索システムの精度に大きく依存することがあげられる。しかし、原言語 - 目的言語で、単語 - 複合語, 複合語 - 単語の対訳関係も得られたことは興味深い。

精度を向上させるための一つの方法は、複数の情報源から対訳語候補を得ることである。複数の情報源としては、新聞記事、WWW などがあげられる。複数の情報を利用することで、より大きな母集団から統計情報を得ることができる。しかし、WWW 上の情報にはノイズが多く含まれることが経験的に知られており、どのように有用な情報をフィルタリングするかが問題である。

また、精度を向上させる別の方法として、文法情報を利用することがあげられる。英語や日本語では、構文解析器や形態素解析器などが使用できる。既知の文法情報をもちいて、解析を行うことで訳語候補を絞めると期待できる。

6 むすび

本稿では、情報検索システムを用いた対訳語抽出モデルを提案し、実際に、日英対訳語抽出実験を行った。その結果、上位 1 位での正解率は 16% であり、上位 10 位までの正解率は、 49% であった。また、語の相関をとるシステムとの比較を行い、情報検索システムを利用した抽出法がより効果があることが分かった。

参考文献

[1] Dan Melamed. "Empritical Methods for Exploiting Parallel Text" *MIT Press.2001*

[2] Reinhard Rapp(1999). "Automatic Identification of Word Translations from Unrelated English and German Corpora" *Proceedings of the 37th Annual Conference of the Association ofor Computatuional Linguistics,pp.519-525*

[3] Pacalse Fung & McKeown,K. R. (1997b). "Finding terminology translations from non-parallel corpora" *Proceedings of the 5th Annual Workshop on Very Large Corpora,Hong Kong, August 1997,pp192-202 37th ACL* pp.519-526

[4] Pascale Fung(2000). "Statistical View on bilingual lexicon extraction" *Parallel Text Processing 2000, pp. 219–236*

[5] Nakagawa,H.,2001, "Disambiguation of Compound Noun Translations Extracted from Bilingual Comparable Corpora" *NLPRS2001(6th Natural Language Processing Pacific Rim Symposium), pp.67-74*

[6] EDICT Project *www.csse.monash.edu.au/ jwb/edict.html*