用例ベース翻訳のためのパラレルコーパスからの対訳対発見

荒牧英治 †, 黒橋禎夫 ‡, 佐藤理史 †, 渡辺日出雄*

† 京都大学大学院 情報学研究科
‡ 東京大学大学院 情報理工系研究科
* 日本 IBM 東京基礎研究所
aramaki@pine.kuee.kyoto-u.ac.jp
kuro@kc.t.u-tokyo.ac.jp
sato@pine.kuee.kyoto-u.ac.jp
hiwat@jp.ibm.com

要旨

本稿は、文同士の対応がとられた日本語と英語の対訳文を入力とし、句レベルで対訳対を発見するシステムについて述べる。本システムは最初に辞書引きを行い対訳対を発見する。その後に周辺の統語情報や文全体の整合性から他の対訳対を発見する。本システムは約 70%のカバレージで 80%の精度で対訳対を発見することに成功した。文同士は非常に適切に対応しているパラレルコーパスにおいては、本手法は従来の統計的手法よりも有効であると考えられる。

Finding Translation Correspondences from Parallel Parsed Corpus for Example-based Translation

Eiji Aramaki †, Sadao Kurohashi ‡, Satoshi Sato †, Hideo Watanabe*

† Graduate School of Informatics, Kyoto University
‡ Graduate School of Information Science and Technology, the University of Tokyo
* IBM Research, Tokyo Research Laboratory
aramaki@pine.kuee.kyoto-u.ac.jp
kuro@kc.t.u-tokyo.ac.jp
sato@pine.kuee.kyoto-u.ac.jp
hiwat@jp.ibm.com

Abstract

This paper describes a system for finding translation correspondences from parsed parallel corpus that is the paired dependency structures of a source sentence and its translation in a target language. At first, the system we have developed finds correspondences by consulting a source-to-target word dictionary, and then finds the other remaining correspondences based on dependency structure and the balance of all correspondences. The system achieved about 70% coverage and 80% accuracy. For finding correspondences, the fact that the source sentences in parallel corpus correspond suitably with the target sentences, is available. This method can collect more correspondences than a statistical approach.

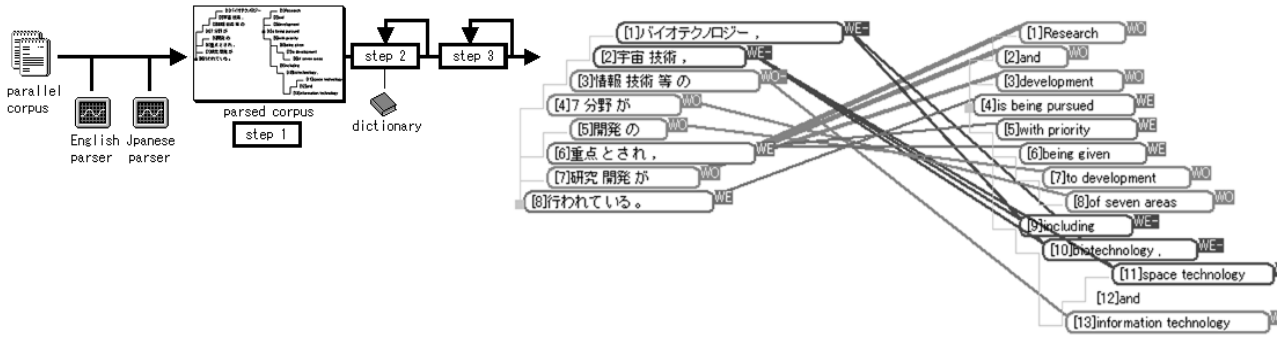

図 1：システムの処理の流れと出力結果

1. はじめに

用例ベース翻訳とは、翻訳すべき入力文に対して、それと類似した翻訳用例をコーパス中から探し、その類似用例をもとに、入力文を翻訳する方式である。用例ベース翻訳を実現するためには、適切なレベルで対応のついた数万から数十万の翻訳用例が必要だと考えられている [1]。

近年、パラレルコーパスから文同士の対応関係を求める研究や、文対応のついた状態から語レベルの対応関係を発見する研究は盛んに行われているが、翻訳用例として適切なサイズである句レベルの対応関係を発見する研究は少ない。

対応関係を求める研究においては、コーパス全体における単語の分布情報を手がかりにして統計的に対訳語や対訳句を求める手法がよく用いられる。統計的手法で高頻度で出現する対訳対を得ることは可能だが、発見された対訳対のコーパスにおけるカバレージは小さくなってしまう [2] [3]。

パラレルコーパスにおいては、その性質上、文を構成する語句同士が非常によく対応している。そこで、本稿では文中の統語情報や全体の整合性を十分に利用することにより、より多くの対訳対を発見するシステムを提案する。

本稿の構成は以下である。次の章で、システムが用いた方法の概要を述べる。3章でシステムの詳細を説明する。4 章で実験と結果を述べる。5章で結論を述べる。

2. 方法の概要

本システムは、パラレルコーパス中の日本語文と英語文の組から、その中に含まれる句の対応関係を求める (図 1)。

その手法は以下の3つのステップから構成される。

Step1: 日英両言語の文を構文解析し、句を単位とした依存構造を得る。
Step2: 辞書引きによって、日英両言語の語の対応を調べ、これをもとに句レベルの対訳対を発見する。
Step3: 対応がつかず残った句について、依存関係等の統語情報や全体の整合性から、対訳対を発見する。

辞書引きにより対応関係を求める際には、複数の対応の候補群から適切な対応を選択する必要がある。また、適切な対応先が存在しない場合も存在し、問題を複雑にする。 この問題を解決するために、本システムでは最初に句を形成し (step1)、句単位での対応付けを行う (step2)。 句単位では、同じ句内の複数の語の対応関係を利用することにより、安定した対訳対を発見できる。

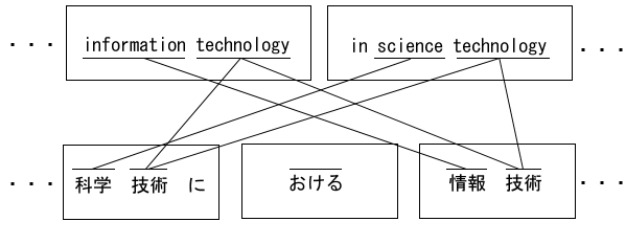

図 2: 辞書引きによる語の対応リンクと句レベルの対応

例えば図 2 では、語 ”technology” と“技術” が 2 つづつ存在するために辞書引きにより、複数の対応リンクが得られる。ここから、正しい対応リンクを決定することは、語レベルでは困難だが、句レベルでは、[information / 情報] と [science / 科学] の対応リンクにより適切に対訳対を発見できる。

このように辞書引きによってできる限りの対応を得た後に、対応がつかずに残った句について対訳対を発見する (step3)。

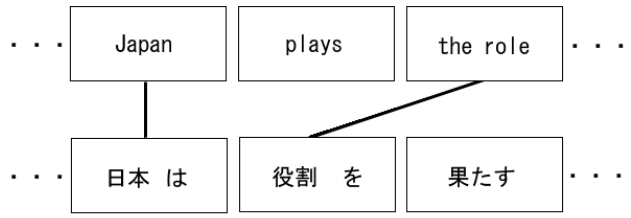

図 3: 辞書引きできない対応の発見

例えば上図(図3)では、2つの対訳対 [Japan / 日本], [role / 役割] をすでに辞書引きによって発見しており、そのまわりにある対応がついていない句は [plays] と [果たす] しかない。このような場合、Step3では、[plays / 果たす] を対訳対として発見する。

この step3 の対訳対発見の精度は step2 の辞書引きによって発見された対訳対の精度よりも低い。よって、本システムは、閾値によってその発見の度合いを調整している。

3.方法

本節では本手法の３つのステップ、(1) 句を単位とした構文解析、(2) 辞書引きによる対訳対の発見、(3) 残った句を用いての対訳対の発見、について詳細を述べる。

3.1 句を単位とした構文解析

対訳文の日本語文、英語文それぞれを構文解析して句を単位とした依存構造を得る。

日本語においては、日本語パーサーKNP（京都大学) [4]の文節の単位を句として用いた。

英語においては、ESG パーサー（IBM Watson Research Center ）[5] を用いて、語単位での依存構造を出力し、その結果を修正して、句を単位とした依存構造を得た。

修正規則は以下である。

- 機能語は後続する内容語に付加する。複数の機能語が連続する場合はそれらすべてを後続の内容語に付加する。
- 内容語を句の終端とする。
- 並列関係を示す語「and」「or」などは、(内容語でないが)単独で句とする。
- 複合名詞は１句に含める。
- be 動詞は後続する動詞を含む結合して 1 句とする。

上記の規則を適応して構文解析結果を整形し、句を単位とする依存関係を得る。

3.2 基本対訳対

まず、辞書引きによって、句同士の対応関係を可能な限り求める。この対応関係を基本対訳対と本稿では呼ぶ。基本対訳対の発見は、候補を可能な限り生成したのち、3つの評価基準を用いて優先順位をつけ優先順位の高い候補から採用していくことにより行う。

3.2.1 基本対訳対の候補生成

最初に句を構成する全ての語について辞書引きを行い、語レベルで対応リンクを張る。

また、辞書引きのアルゴリズムは以下のような例外処理を行う。

例外処理 1:辞書引きしない

両言語とも機能語の辞書引きは行わない。これは機能語が過剰にマッチするのを避けるためである。

例外処理 2:部分一致の採用

日本語の語の長さが2文字以上の場合、それらを任意の2つの語に分割して辞書引きを行う。 よって、”東西”などが一つの語としてパーサーの出力結果では扱われていても、”東”,”西”それぞれの語として辞書引きされうる。

例外処理 3:語を活用して引く

2つの方法で語を変形させた形でも辞書引きを行った。

・日本語の語では「する」を語尾に付加した形でも辞書引きを行う。

・英語の語は語尾を「ly」「d」「ed」「s」「es」「ies」など屈折させた形でも辞書引きを行う。

これらの例外は辞書の不足分を補うために設けたものであり、アルゴリズムの動作を本質的に変えるものではない。

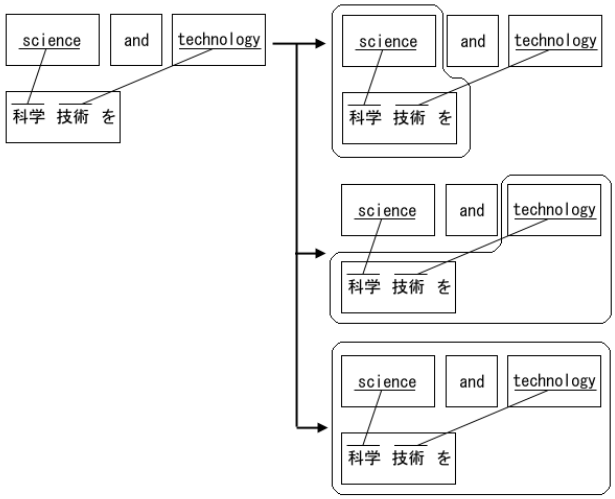

図 4: 複数の句を含む候補の生成

辞書引きの結果図4のようにある句が相手側の言語の複数の句に対応リンクが張られる場合がある。このような場合、複数の句が隣接しているか、または、機能語を挟んで隣接していれば基本対訳対の候補として考える。

図4の場合は、日本語文の１つの句と英語文の2つの句へ、語の対応リンクが張られており、かつその2句は間に機能語を挟んで隣接している。よって、機能語も含めた英語文の3つ句と日本語文の1つの句との基本対訳対候補を、1句同士の候補とともに、生成する。

3.2.2 基本対訳対の採用

生成された対訳対候補の評価は以下の3つの尺度で行い、評価の高いものから順に採用していく。 また、採用された対訳対同士が矛盾しないように、ある対訳対が採用された際には、その対訳対に含まれる句を用いた他の対訳対の候補は棄却する。このように基本対訳対の候補の採用、または棄却を繰り返し、採用する基本対訳対の候補がなくなれば、この処理は終了する。

評価基準1:充足度

句に含まれる内容語が余ることなく、より多く対応している候補を優先させるために充足度という評価基準を設けた。

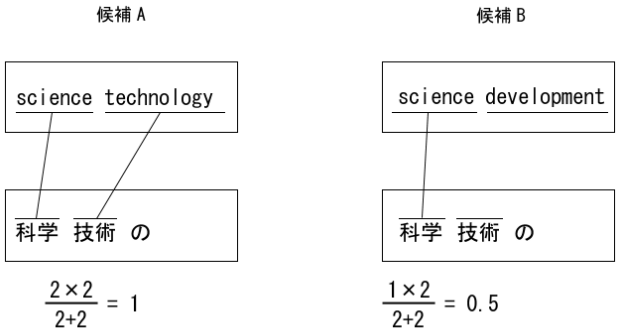

図 5: 対応スコアの定義

例えば、図 5 の2つの対訳対候補 A と B では、句内のすべての内容語が互いに対応している候補 A を優先して採用するべきである。よって、充足度は句に含まれる語の対応リンクの数と、両言語の句の内容語の数を用いて以下のように定義した。

$$\text{充足度} = \frac{\text{対応リンク数} \times 2}{\text{日本語の内容語の数} + \text{英語の内容語の数}}$$

また、この充足度が低い場合は、どちらかの言語側の内容語が対応リンクが張られずに余っている事を示す。充足度とどちらの言語側で余っている内容語があるかの2つの情報によって、対応候補を以下の4つのタイプに分類する。この分類は後の拡張対訳対の処理で利用する。

1: 充足

対訳対内のすべての内容語の対応がお互いで取れている。例えば、図5の候補 A は充足である。

2: 日本語不足

日本語の内容語のリンクが対応先に張られていない。

3: 英語不足

英語の内容語のリンクが対応先に張られていない。

4: 両言語不足

両言語で対応先へのリンクがはられていない内容語が存在する。例えば、図5の候補 B は両言語不足である。

評価基準2:句数

多くの句からなる候補を優先するために、句数という評価基準を設けた。

句数 = 候補の日本語の句数 + 候補の英語の句数

この句数は候補の英語側と日本語側に含まれている句の和で計算する。これによって、ある対訳対候補の一部もまた対訳対候補として考えられる場合は、大きな対訳対候補が優先される。

評価基準3:近傍支持度

周辺に他の対訳対候補が多く存在している候補を優先するために、近傍支持度という評価基準を設けた。

近傍支持度 = 近傍に存在する他の候補の数

上式の 「近傍」の定義としては、現在、句の距離 6 を閾値として用いている。

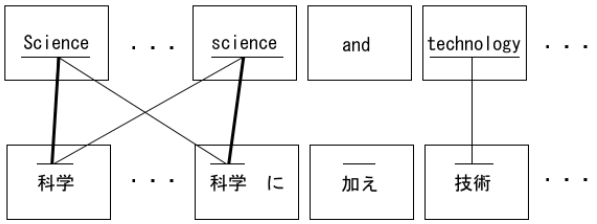

図 6: 近傍の対訳対にサポートされた候補

例えば、図 6 においては、[science] と [科学] がそれぞれ2つ存在するため、前述の評価基準では決定できない。評価基準3では、別の候補 [technology / 技術] が近くに存在するために、図の太線の基本対訳対候補が優先される。

以上の3つの評価基準によって、対応候補に優先度をつける。評価基準は記述する順のとおりの優先度を持つものとする。

よって、まず評価基準1の充足度によって優先される候補を採用する、その際、充足度が等しい基本対訳対候補があれば、それらを評価基準2によって評価し、優先されるものを採用する。

3.3 拡張対訳対

基本対訳対は辞書引きをもとにしているため、辞書引きされない語で構成される句は対訳対として出力されない。そこで、基本対訳対には含められず残った句に対して新たな対応の発見を行う。この処理で、見つかった対応を拡張対訳対と呼ぶ。

3.3.1 拡張対訳対の候補生成

残った句は実際に、適切な対応先が存在しないかもしれない。その場合は、未対応のままにしておくべきである。しかし、そうでない場合は、図 7と図 8 のような 2 通りの可能性があり、それぞれ適した処理を行うべきである。

可能性1: 残った句は基本対訳対に含めるべきである(よって、残った句を加えた対訳対を発見する処理を行う)。

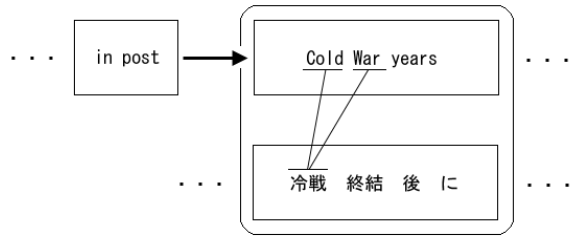

図 7: 拡張対訳対の可能性 1(基本対訳対の修正)

可能性2: 残った句は、もう一つの残った句と対訳対を作るべきである(よって、残った句同士の対訳対を新規に発見する)。

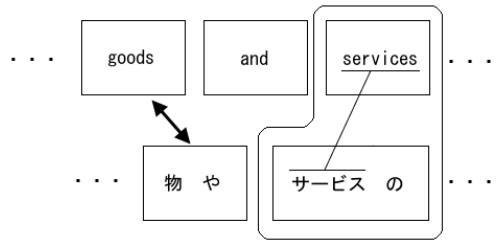

図 8: 拡張対訳対の可能性 2(新規対訳対の発見)

上記の2つ可能性を考えて、すべての残った句について、拡張対訳対の候補を考えうる限り生成する。

3.3.2 拡張対訳対の採用

生成された候補に評価スコアを定義し、高いスコアの候補から順次採用する。これはスコアが閾値を下回るまで繰り返す(閾値以下の候補は未対応句のまま出力される)。

評価スコアは以下のように定義し、依存関係などの統語情報や全体の整合性により定まるようにした。

$$\text{評価スコア} = \frac{B}{4} + \sum_{k=1}^{n} \frac{X}{J(k) + E(k)}$$

n は、拡張対訳対の候補の近傍にある基本対訳対の数である。近傍として、句の距離2を閾値として用いた。

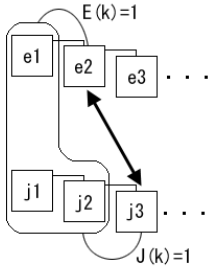

図 9: 拡張対訳対の候補[e2 / j3]のスコア

E(k) は英語側での、基本対訳対 k の英語句と拡張対訳対の英語句との距離である。J(k) は日本語側での、基本対訳対 k の日本語句と拡張対訳対の日本語句との距離である。

図 9 は、日本語、英語ともに残った句同士で対訳対を作る拡張対訳対の候補である。この場合の E(k) は、e1 と e2 との距離となり、1となる。

同様に J(k)も j2 と j3 の距離となり 1 となる。

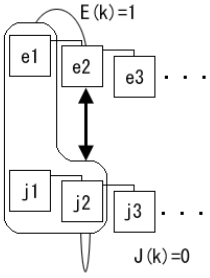

図 10: 拡張対訳対の候補[e1,e2 / j1,j2]のスコア

図 10 は、余った句 e2 を基本対訳対 [e1 / j1,j2] に含める拡張対訳対の候補である。この場合、E(k)は 1 であるが、J(k)は 0 と考えて処理した。

変数 X は、周辺の句の情報によって定義され、初期値を X=1 とし、下記表の条件が満たされる場合には、X に表記された値を乗じる。複数の条件を満たしているならば、それらすべてを掛け合わせる。

X	条件
2	**日本語不足**である基本対訳対に日本語未対応句を付け加える場合、または、 **英語不足**である基本対訳対に英語未対応句を付け加える場合
$\frac{1}{8}$	未対応句を既出の対訳対に加える場合、元となった対訳対が**充足**であった場合
$\frac{1}{2}$	未対応句を既出の対訳対に加える場合、異なる品詞である場合。 ただし、品詞は句内の形態素に動詞が一つでも存在しているか、否かで動詞句、名詞句と 2 つに分類している。
$\frac{1}{2}$	未対応句を既出の対訳対に加える場合、それらが依存関係にない場合

表 1: X の値の修正

変数 B は、文章中における基本対訳対に含まれる句の割合によって定義される。

$$B = \frac{\text{基本対訳対中の句の数}}{\text{文中の句の数}}$$

変数 B によって、評価スコアは近傍の基本対訳対によるばかりではなく、遠くの基本対訳対の影響をうけ、積極的に拡張対訳対の発見を行うことになる。

4. 実験

本節では、行った実験とその結果について述べる。

実験の前処理、実験結果、本実験における適合率−再現率の定義と結果、カバレージの定義と結果、実際の出力例、という構成である。

4.1 実験の前処理

パラレルコーパスとして以下の 2 つのコーパスを利用した。

コーパス	特徴
コーパス A 科学技術庁と経済企画庁の白書 [10]	比較的長文が多い。使用されている語句メインが限定されている。
コーパス B 学研辞書の用例	比較的短文が多い。

表 2: コーパス

これらのコーパスから次の条件で各々100 文を無作為抽出し評価セットとした。

- 日本語、英語両言語の句数が20 句以下である。
- 日本語、英語両言語において語彙数が大きく異ならない (句数の比が 0.5〜2 となる) 。

これらを日本語においては、KNP 日本語パーサー (京都大学) [4]、英語を ESG パーサー (IBM Watson Research Center)[5] を用いて依存関係を明らかにし、本システムを用いて対訳対を発見した。

また、システムが発見する対訳対の精度を調べるために、評価セット内の句について正しい対応先を人手で記述した。記述にあたっては、複数の句同士の対応も考慮しながら、可能な限り小さな対応を記述した。

実験の評価にあたっては、以下のような方法で行った。

システムが出力した対訳対と人手による対訳対とが完全に一致した場合、これを正解とした。過不足が存在した場合は半正解とした。また、そうでない場合は不正解とした。

システムが対応を出力する際には設定された閾値によって、拡張対訳対を発見する積極性が変化する。閾値は実験の設定では 3 がほぼ上限であり、閾値＝3 の時には、拡張対訳対を生成せず基本対訳対のみが出力される。逆に閾値＝0 の状態においては、拡張対訳対発見を最大限積極的に行う。

4.2 実験の結果

実験結果においては、最初閾値を最初0に設定し、0.5 づつ3に近づけていった。以下 2 つのコーパスについてその結果を載せる。

対訳対の句の数は下表のようになる。拡張対訳対の値は閾値=0 として、拡張対訳対を最大限に多く発見した際の数値である。

拡張対訳対の句数は下表に示されるように基本対訳対よりも小さいが、これは未対応な句同士で対訳対を発見する際、1句同士の対訳対となることが多いためである。

	英語	日本語
基本対訳対	2.20	3.20
拡張対訳対	1.67	2.73

表 3: 対訳対のサイズ

評価グラフ1，2は共に表の値をグラフ表示したものである。グラフ1は横軸に閾値、縦軸に発見された対訳対の数を表示した。グラフ2においては、横軸に閾値、縦軸に評価の内訳をパーセント表示した。

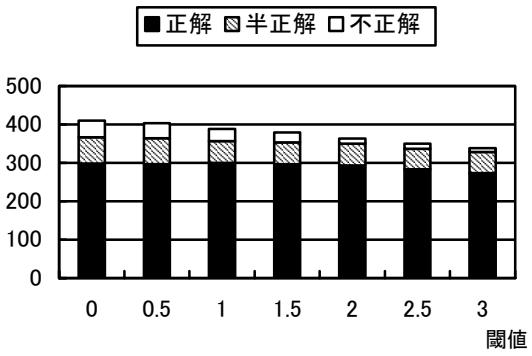

図 11: コーパス A の対訳対数と評価

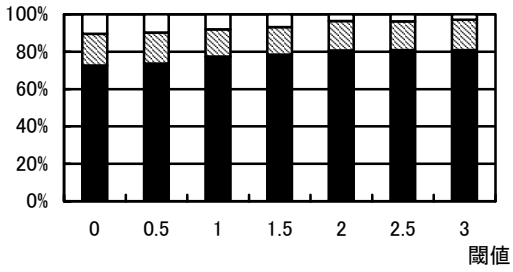

図 12: コーパス A の評価の割合

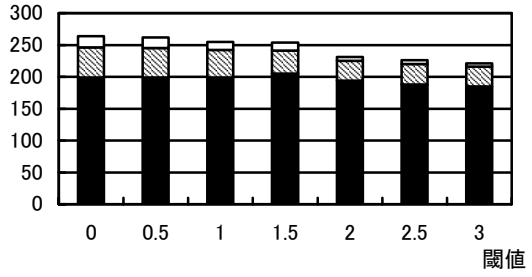

図 13: コーパス B の対訳対数と評価

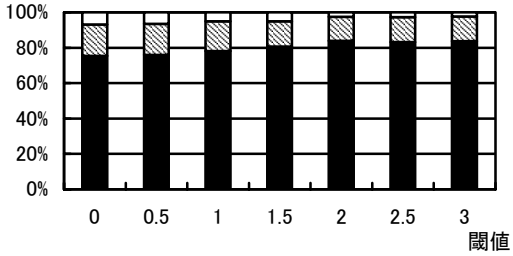

図 14: コーパス B の評価の割合

閾値を高くすれば拡張対訳対の発見される数が減るが正解の割合は増える。これは、拡張対訳対の精度が悪いからであるが、拡張対訳対による正解は辞書引きでは得られない対訳対なので、これは重要であると考える。

4.3 適合率-再現率

適合率-再現率グラフを下に載せる。適合率と再現率の定義は、本来の正解を部分一致に分解して出力してしまうことにより定義が困難である。本稿では以下のようにこれを定義した。

$$\text{適合率} = \frac{(\text{正解の対訳対の数}) + \frac{1}{2} \times (\text{半正解の対訳対の数})}{\text{発見された対訳対の数}}$$

適合率は正解を 1、半正解を 0.5、不正解を 0 とし、出力された対訳対の個数で割ることによって定義した。

$$\text{再現率} = \frac{(\text{正解の対訳対の数}) + \frac{1}{2} \times (\text{半正解の対訳対の数})}{\text{人手で発見された対訳対の数}}$$

再現率は正解を 1、半正解を 0.5、不正解を 0 とし全正解個数で割ることによって定義した。

参考に、正解のみを 1 とし、半正解を不正解と同様に 0 と考えた場合を点線で表示した。

結果のグラフは以下のようになった。

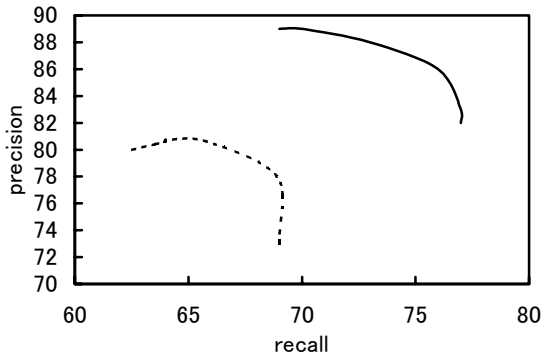

図 15: 適合率-再現率

このグラフから、最大で人手による正解の 77%を半正解もしくは正解として出力可能であることが分かる。その際の適合率も 82%と高い。

また、この再現率が 77%の点は拡張対訳対を生成する閾値を 0 にし、積極的に拡張対訳対を発見した場合である。

4.4 カバレージ

適合率及び再現率は発見された対訳対の数で計算している。ここで、カバレージとして、文中の全句数中での出力された正解の句の数を計算した。半正解においては、含まれる正解句の数のみをカウントした。

閾値=3 として基本対訳対だけが発見される場合のカバレージは 51.9%であった。

閾値=0 として拡張対訳対を積極的に発見した場合のカバレージは 68.1%であった。

4.5 実際の出力例

実際にどのような対訳対が発見されるのか、以下例示する。半正解は、削除されるべき語を()、追加されるべき語を[]表記した。

英語	日本語
In particular	特 に
among major countries	主要 国 の
with end of Cold War	冷戦 終結 とともに
in world market	世界 市場 に おける
by monthly instalments	月賦 払い で

表 4: 基本対訳対の正解例

英語	日本語
(crossing) borders	国境 を
in that area	(旧 東ドイツ) 地区 の
(like) home	我が家 に
into his suitcase	(彼 は) スーツケース に
Transnational	国 を [超えて]

表 5: 基本対訳対の半正解例

基本対訳対の発見は辞書引きされる語を必要とするため、複合名詞などの表現が多い。

英語	日本語	スコア
is being pursued	行われて いる	2.75
of G7 nations	先進 7 カ国 の	2.6
is vital	重要 だ	1.5
of TFP	全 要素 生産性 が	1.5
with priority- being given	重点 と する	0.33

表 6: 拡張対訳対の正解例

英語	日本語	スコア
tree (become)	その 木 は	1.2
went [to bed]	寝る	1.0
is (also) important	重要 で ある	1.0
She (held)	彼女 は	0.5
by companies	(低迷 する) 企業 の	0.33

表 7: 拡張対訳対の半正解例

拡張対訳対では、G7, TFP のような頭文字語や動詞句など辞書引きでは得られない対訳対を発見している。

5. おわりに

本手法はパラレルコーパスを入力とすることを前提としており、comparable なコーパスにおいては従来の確率的な手法を行うしかないと思われる。しかし、パラレルコーパスにおいては、無駄な部分が少なく本手法のように強力に対応付けが可能だと考える。本研究は高いカバレージを持ち、精度も遜色がないため、発見された対訳対は人手により不適当なものを取り除く作業を行うことにより、短時間で有用な対訳対が得られると考えられる。このような強力な対応付けはパラレルコーパスでしか行えないと思われるが、パラレルコーパスは貴重であり、最大限その資源の生かすために本手法は有効である。

ただ、本手法で得られた対訳対を翻訳システムに搭載しての評価は行っておらず、今後の課題としたい。

6. 参考文献

[1] Hideo Watanabe, Sadao Kurohashi, Eiji Aramaki, ″Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation,″ Proc. of 18th Coling, pp. 906--912, 2000.

[2] Dekai Wu, “An Algorithm for Simultaneously Bracketing Parallel Texts by Aligning Words,” ACL95

[3] Kaoru Yamamoto, Yuji Matsumoto, “Acquisition of Phrase-level Bilingual Correspondence using Dependency Structure,” COLING-2000.

[4] Sadao Kurohashi, Makoto Nagao, “A Syntactic Analysis Method of Long Japanese Sentences based on the Detection of Conjunctive Structures,” Computational Linguistics, Vol. 20, No, 4, 1994.

[5] McCord, C.M., Slot Grammars, “Computational Linguistics,” Vol. 6, pp. 31-43, 1980.

[6] Nagao. M., “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle,” Elithorn, A. and Banerji, R. (eds.): Artificial and Human Intelligence, North-Holland, pp. 173-180, 1984.

[7] Sato, S. and Nagao, M., “Toward Memory-based Translation”, Proc. of 13th Coling 90, Vol.3, pp. 247-252, 1990.

[8] Sumita, E., Iida, H., “Translating with Examples: A New Approach to Machine Translation”, Proc. of Info-Japan ’90, 1990.

[9] Hideo Watanabe, “A Similarity-Driven Transfer System,” Proc. of 14th Int. Conf. of Computational Linguistics ’92, pp.770-776, 1992.

[10] Hitoshi Isahara and Masahiko Haruno, ″Japanese-English aligned bilingual corpora″, in Parallel Text Processing: Alignment and use of translation corpora. (Text, Speech and Language Technology, Vol. 13), p. 313-334, Kluwer Academic, 2000.