対訳コーパスを用いたゼロ照応タグ付きコーパスの自動構築

古川 智雅† 中澤 敏明‡ 柴田 知秀§¶ 河原 大輔§ 黒橋 禎夫§¶
†京都大学工学部 §京都大学大学院情報学研究科
‡科学技術振興機構 ¶科学技術振興機構 CREST
furukawa@nlp.ist.i.kyoto-u.ac.jp nakazawa@pa.jst.jp
{shibata, dk, kuro}@i.kyoto-u.ac.jp

1 はじめに

ゼロ照応とは、述語の項が省略されており、それが文脈中のある語を指し示すという現象である。例えば以下の文では、構文構造のアノテーション基準上、「河野外相」が「帰国する。」に係るため、「訪問し」のガ格が省略されており、その先行詞は「河野外相」である。

(1) 河野外相 はサミット後、ブラジル、アルゼンチンを 訪問し、十七日に帰国する。

このような述語の項の省略はゼロ代名詞と呼ばれる。

ゼロ照応解析は、ゼロ代名詞を認識し、その先行詞を同定するタスクであり、近年活発に研究されている。ゼロ照応解析は情報抽出や機械翻訳などのアプリケーションにおいて重要であるが、最新の研究においてもその精度は 50%程度と低い [1]。ほとんどの研究が教師あり機械学習を利用しているが、その学習に用いられているタグ付きコーパスのサイズが小さいことが精度が低い原因の一つと考えられる。

本稿では、日英対訳コーパスを用いて、日本語のゼロ照応タグ付きコーパスを自動構築する手法を提案する。本手法では、対訳文ペアの日本語と英語における構文構造のずれを手がかりとし、ゼロ代名詞の先行詞を同定する。既存の大規模な対訳コーパスを利用することができるため、容易に大規模なゼロ照応タグ付きコーパスを構築できる。

新聞記事の日英対訳コーパス約 27 万文から、ガ格については精度 83%で 1.5 万文程度、ヲ格については精度 56%で 1,600 文程度の日本語ゼロ照応タグ付きコーパスを自動構築することができた。本稿では、この自動構築するタグ付きコーパスを**擬似ゼロ照応タグ付きコーパス**と呼ぶ。

2 関連研究

ゼロ照応解析の研究はこれまで多く行われてきており、格フレームに基づく手法 [2] や著者・読者表現を扱う手法 [3] などがある。また近年ではニューラルネットワークを用いてゼロ照応解析を行う研究も進められている [1, 4]。いずれの研究もモデルのパラメータ学習に用いているタグ付きコーパスが数万文程度と小規模であり、これがゼロ照応解析の精度が低い要因の一つとなっている。

本研究で提案する対訳コーパスを用いたタグ付きコーパスの自動構築手法は、中英の統計的機械翻訳において中国語で省略されている代名詞を自動的に補って翻訳する Wang らの研究 [5] と類似している。Wang らが補うのは代名詞のみであるのに対し、本研究では文中に出現する具体的な名詞を補うことで、照応解析精度向上に寄与するタグ付きコーパス構築を目標としている点で異なる。

3 対訳コーパスを用いた擬似ゼロ照応タグ付きコーパスの構築

3.1 概要

本研究では、日英対訳コーパスを用いて日本語擬似ゼロ照応タグ付きコーパスを構築する。対訳コーパスには単語アライメントを適用、また対訳コーパスの日本語文と英語文に構文解析を適用し、それらの情報に基づいて日本語文におけるゼロ代名詞の先行詞を同定する。

先行詞同定の手がかりとして、日本語文と英語文の構文構造のずれがある。構文構造のずれの例を次に示す。以下の例では、日本語文中のゼロ代名詞を含む動詞とそれに対応する英単語に下線を引き、日本語文中

の先行詞とそれに対応する英単語に二重下線を引いている。

(2) a. 河野外相 はサミット後、ブラジル、アルゼンチンを 訪問し、十七日に帰国する。
b. Kono will visit Brazil and Argentina after the summit and fly back home on July 17.

例文 (2a) では「河野外相」は「訪問し」ではなく「帰国する」に係っている一方で、(2b) では “Kono” は “visit” に係っており、構文構造がずれている。「訪問し」のガ格はゼロ代名詞であるが、「訪問し」の英語対応先 “visit” の主語が “Kono” であり、その日本語対応先が「河野外相」であるため、ゼロ代名詞の先行詞が「河野外相」であるとわかる。

英語文において代名詞が補われている場合は、その代名詞の先行詞を参照することによって、日本語側のゼロ代名詞の先行詞を推定できる可能性がある。

(3) a. 川島公使 はこれに対し、本国に韓国政府の要請を 伝える と答えた。
b. Kawashima replied he would convey the request to the government.

例文 (3a) では、「伝える」のガ格がゼロ代名詞であり、「伝える」の英語対応先は “convey” である。“convey” の主語は “he” であり、また “he” が “Kawashima” を照応しているので、“Kawashima” の日本語対応先である「川島公使」がゼロ代名詞の先行詞であるとわかる。

3.2 手順

擬似ゼロ照応タグ付きコーパスの構築手順を以下に示す (図 1 も参照)。なお、入力となる日英対訳コーパスについて、日本語文は構文・格解析し、英語文は構文解析および共参照解析を適用しておく。

Step 1. 日本語文におけるゼロ代名詞を含む動詞 v_j の同定
日英対訳コーパスの日本語文の構文・格解析結果から、ゼロ代名詞を含む動詞を同定する。ただし、「(〜に) 対して」のような複合辞を構成する動詞は対象外とする。図 1 では、動詞「訪問」のガ格の項が省略されており、「訪問」が v_j となる。

Step 2. v_j の英語対応先 v_e の同定
日英対訳コーパスの単語アライメント結果から v_j と対応している英単語 v_e を見つける。図 1 では、

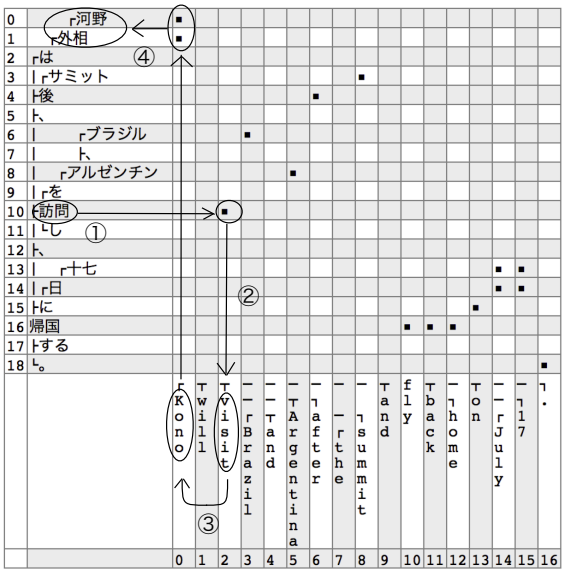

図 1: 提案手法の手順

アライメント結果から「訪問」と対応している英単語は “visit” であることがわかるので、“visit” を v_e とする。

Step 3. 英語文における v_e の主語もしくは目的語 n_e の同定
英語文の構文解析結果から、注目している日本語のゼロ代名詞がガ格の場合は v_e の主語、ヲ格の場合は v_e の目的語を同定する。同定した語が代名詞であれば、共参照解析の結果を用いて、その照応先の名詞を見つける。図 1 では、「訪問」はガ格がゼロ代名詞であるため、“visit” の主語 “Kono” を抽出する。

Step 4. n_e の日本語対応先 n_j の同定
単語アライメント結果から n_e の日本語対応先 n_j を同定する。n_j が名詞でない場合や v_j に係る場合は対象外とする。図 1 では、アライメント結果を用いると “Kono” と対応しているのは「河野外相」であることがわかるので、「訪問」のガ格ゼロ代名詞は「(河野) 外相」を照応していることがわかる。

Step 3 の v_e の主語については、「名詞主語」(nsubj) のみを用いる場合と、nsubj に加えて「制御動詞の主語」(xsubj)、「副詞節による修飾」(advcl) を介した nsubj の 3 種類を用いる場合の 2 通りを試す。これは、nsubj のみを利用した方が精度が高いが、その反面、抽出できる先行詞の数が少なくなると考えられるので、このトレードオフを検証するためである。v_e の目的語については、「直接目的語」(dobj) を用いる。

	ガ格		ヲ格
	nsubj	nsubj+xsubj+advcl	dobj
ガ格またはヲ格のゼロ代名詞を含む動詞の数 (Step 1)	251,917		40,337
取得できた v_e の数 (Step 2)	172,329		25,592
取得できた v_e の主語または目的語 (n_e) の数 (Step 3)	47,148	119,493	4,295
(共参照の関係を用いた語数)	(3,634)	(6,918)	(249)
推定できたガ格またはヲ格の先行詞 (n_j) の数 (Step 4)	15,160	40,216	1,616
(共参照の関係を用いた語数)	(1,384)	(2,601)	(96)
ランダムに選んだ 100 語の精度	83%	55%	56%

表 1: 擬似ゼロ照応タグ付きコーパスの構築結果と精度評価

4 実験

対訳コーパスとして日英新聞記事対応付けデータ 269,227 文[1] を用いて、擬似ゼロ照応タグ付きコーパスを構築した。言語解析ツールとしては、日本語には形態素解析器 JUMAN++[2] および構文・格解析器 KNP[3]、英語には Stanford CoreNLP を用いた。単語アライメントツールとしては nile[4] を用いた。

本研究では、ガ格とヲ格のゼロ代名詞を対象とした。前節で述べたとおり、ガ格のゼロ代名詞に対しては、英語の解析結果を用いる際に nsubj のみを用いる場合と、nsubj、xsubj、advcl の 3 種類を用いる場合の 2 通りの比較を行った。

4.1 構築した擬似ゼロ照応タグ付きコーパス

構築した擬似ゼロ照応タグ付きコーパスの統計を表 1 に示す。3.2 節で説明した各 Step ごとの数も示している。Step 1 から Step 2、Step 3 から Step 4 で数が減少している原因は、単語アライメントの対応先がないことである。nsubj 使用時のガ格とヲ格について、Step 2 から Step 3 で数が減少している原因は、nsubj・dobj で主語・目的語が同定できるカバレッジが低いことが挙げられる。

4.2 擬似ゼロ照応タグ付きコーパスの精度評価

次に、構築した擬似ゼロ照応タグ付きコーパスの精度を評価した。ガ格、ヲ格それぞれのゼロ代名詞を含む動詞 100 語に対して、推定した先行詞が正しいかどうかを人手で評価した。

ガ格、ヲ格それぞれの推定した先行詞の精度を表 1 の最下行に示す。ガ格ゼロ代名詞の先行詞推定精度に関しては、nsubj だけを用いた場合は 83%と高い一方で、nsubj、xsubj、advcl の 3 種類を用いた場合は 55%と低かった。逆に、獲得できた先行詞の数 (Step 4) は、nsubj だけを用いた場合には 1/3 強に留まった。ヲ格ゼロ代名詞の先行詞推定精度は 56%と高くなかった。

以下に正しく出力できた例を 2 つ示す。

(4) a. だが、人々 は今のうちに正常な生活を手に入れたいと 望み、その実現を指導者に要求しているのである。
b. But people want to be able to lead a normal life now and demand it from the government.

(4) では日本語文と英語文の構文構造のずれによって「望む」のガ格の先行詞が「人々」であると同定できた。

(5) a. ハト は二歳ぐらいのオスで、衰弱しているものの、馬場所長は「二週間ほどで元気になる」と話している。
b. The male pigeon is about two years old, and is still weak, but Baba said it should recover after about two weeks.

(5a) では「元気になる」においてガ格の項が省略されている。(5b) において「元気になる」と対応している英単語は “recover” であり、その主語は “it” である。共参照解析の結果、“it” が “pigeon” を照応しており、“pegion” と対応する「ハト」が「元気になる」のガ格の先行詞であると同定できた。

誤り分析を行ったところ、単語アライメントの誤り

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/jea/index-ja.html
[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++
[3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[4] https://github.com/jgwinnup/nile

と英語文の構文解析誤りが主な誤り原因であった。以下に誤り原因ごとの例を示す。

- 単語アライメントの誤り

(6) a. 大使が事件直後、奥 夫人 に連絡した際には、夫人は「子どもがかわいそうだ」と話したが、気丈に ふるまっていた という。
b. When Orita called Oku's wife after the incident, she remained composed, although she said it would be difficult for the children, he said.

「ふるまって」のヲ格の項は文中には存在しないが、「ふるまって」と対応する英単語を誤って “composed” ではなく “call” であると同定したために、その目的語である “wife” と対応する「夫人」が誤って出力された。

(7) a. 団体の 行動 と、個人の行為を明確に 区別した ものだが、運用の際にはさらに議論を深める必要がある。
b. The guidelines clearly distinguish between the activities conducted by a group and those by individuals.

「区別した」のガ格の項は文中には存在しないが、「区別した」と対応する英単語 “distinguish” の主語が “guidelines” であるために “guidelines” と対応する「行動」が出力された。日本語の対象文より前の文に “guidelines” と対応する日本語が存在するかもしれないが、文単位で処理しているため正しい出力が得られなかった。文章単位で処理することは今後の課題である。

- 英語文の構文解析誤り

(8) a. 四十歳以上が条件で、同財団が人材を 募集・登録し、企業の要請に応じてアドバイザーとして 派遣する。
b. The foundation recruits and registers these human resources and then dispatches them as advisers at the request of firms.

「派遣する」の正しいヲ格の項は「人材」であるが、「派遣する」と対応する英単語 “dispatches” の目的語 “them” が “recruits” を照応していると誤って解析されたため、出力が「募集」となった。

5 おわりに

本稿では、日英対訳コーパスを用いて、日本語のゼロ照応タグ付きコーパスを自動構築する手法を提案した。実験では、新聞記事の日英対訳コーパス約 27 万文から、ガ格については精度 83%で約 1.5 万文、ヲ格については精度 56%で 1,600 文程度のタグ付きコーパスを構築することができた。今後の課題として、構築したコーパスを用いて日本語ゼロ照応解析の精度を改善していく予定である。

参考文献

[1] Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Neural network-based model for Japanese predicate argument structure analysis. In *Proceedings of ACL 2016*, pp. 1235–1244, August 2016.

[2] Ryohei Sasano and Sadao Kurohashi. A discriminative approach to Japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 758–766, November 2011.

[3] Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. Japanese zero reference resolution considering exophora and author/reader mentions. In *Proceedings of EMNLP 2013*, pp. 924–934, 2013.

[4] 大内啓樹, 進藤裕之, 松本裕治. 深層リカレントニューラルネットワークを用いた述語項構造解析. 情報処理学会 第 229 回自然言語処理研究会, 2016.

[5] Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. A novel approach to dropped pronoun translation. In *Proceedings of NAACL-HLT 2016*, pp. 983–993, 2016.