質問応答システムを利用したゼロ照応の効率的なアノテーション

淺尾 仁彦　　飯田 龍　　鳥澤 健太郎
国立研究開発法人 情報通信研究機構 (NICT)
{asao, ryu.iida, torisawa}@nict.go.jp

1 はじめに

文中に明示されていない格要素を同定するゼロ照応解析の処理は、機械翻訳や情報抽出の部分問題として重要な処理であり、盛んに研究が行われている [1, 3, 9]。近年のゼロ照応解析においては、機械学習アルゴリズムの発展により、ゼロ照応がアノテーションされたコーパスを利用する手法が主流となっている。このため、いかに効率的かつ一貫したゼロ照応のアノテーションを行うかが重要な課題となる。松林ら [7] が報告しているように、ゼロ照応がアノテーションされた既存のコーパスにはアノテーションの誤りが多く含まれているため、質にも留意したデータ作成の方法が必要になると考えられる。

ゼロ照応の事例収集においては、コーパスに対し網羅的にアノテーションを行う手法が一般的である。本研究では、質問応答システムを利用することで、指定した格のゼロ照応をもつ可能性が高い事例 (述語と格要素の候補、およびその 2 つが出現している文) を自動的に収集し、それらの事例に対してのみアノテーションを行うことで、アノテーションの効率を向上させる手法を提案する。

本研究で提案するアノテーション手法はどの格にも適用できると考えられるが、今回は任意格であるデ格のゼロ照応をアノテーション対象とした。述語に対してデ格をとる名詞は場所・原因・手段などを表し、テキストからの情報抽出においては重要な役割を果たすが、既存のコーパスにおいて任意格のゼロ照応はアノテーションされていないか、あるいは必須格のゼロ照応に比べて事例が少なく、機械学習による解析手法を適用することが困難である。本研究では、デ格ゼロ照応の候補 30,000 事例についてアノテーションを行った結果、表 1 に示すようなデ格ゼロ照応の事例を得た。本手法により、従来手法に比べ約 3 倍の効率でデ格ゼロ照応の事例を収集できた。また、アノテータ間の一致率についても、κ 値 (Fleiss) [2] で 0.555 という一致率を得た。

本稿では、まず 2 節で、本研究で提案する効率的なアノテーションの手法について説明し、3 節で実際に行ったアノテーションの手順について説明する。4 節でアノテーションの結果をまとめ、5 節でアノテーションの効率について既存手法との比較を行う。6 節で本研究で構築したデータを用いた評価実験の結果を報告する。最後に 7 節でまとめと今後の課題を述べる。

表 1: 収集されたデ格ゼロ照応の事例

これは直接顔に $\underline{氷}_i$ を乗せて [$\underline{\phi_{i\,デ}}$] 顔を**冷やします**。
世界ボクシング協会 (ＷＢＡ) フライ級タイトルマッチ１２回戦 (７日・神戸ワールド記念ホール) の調印式と $\underline{記者会見}_i$ が６日、神戸市内で行われ、同級１１位で挑戦者の亀田大毅 (亀田) は「怖さは全くない。対策をするようなレベルの相手じゃない。間違いなくおれが勝てる」と [$\underline{\phi_{i\,デ}}$] 強気に**言い切った**。

2 提案手法

本研究では、ゼロ照応の事例を効率的に収集するため、質問応答システムを利用し、あらかじめゼロ照応を含む可能性の高い事例のみに対してアノテーションを行う。

2.1 質問応答システムを用いる手法の概要

本手法におけるゼロ照応のアノテーションは、質問応答システムの提示した回答を評価するという形で実施される。

例えば、「奈良で何を食べる」という質問文に対して、質問応答システムを用い、回答「豆腐」と、その根拠となるコーパス中の原文「奈良に行ったら、豆腐を食べるといいです。」を得る。アノテータは、回答「豆腐」が適切かどうか (原文に「奈良で豆腐を食べる」という内容が含まれているかどうか) を判定する。これにより、原文において「食べる」の省略されたデ格が「奈良」であるというアノテーション結果が得られる。

ここで示したデ格や、ヲ格・ニ格などのゼロ照応は、ガ格と比較して頻度が低い。そのため、コーパスに対して網羅的にアノテーションを行っていく手法では、検出に細心の注意を払わなければ、事例を見逃してしまう可能性が高い。これに対し、本手法ではアノテーションの候補を陽に提示することで、ゼロ照応を比較的容易に検出できると期待される。

質問応答システムとして、我々の研究グループで開発している WISDOM X [8] を利用する。WISDOM X は Web 40 億文書を知識源としており、それを利用することで多様なゼロ照応の事例を収集できると考えられる。

2.2 アノテーション対象となる質問・回答候補文ペアの構築方法

本節では、WISDOM X を利用した、アノテーション対象となる質問・回答候補ペアの具体的な取得方法について述べる。手法の概要を図 1 に示す。

本手法では、まず、収集したいゼロ照応の格での出現

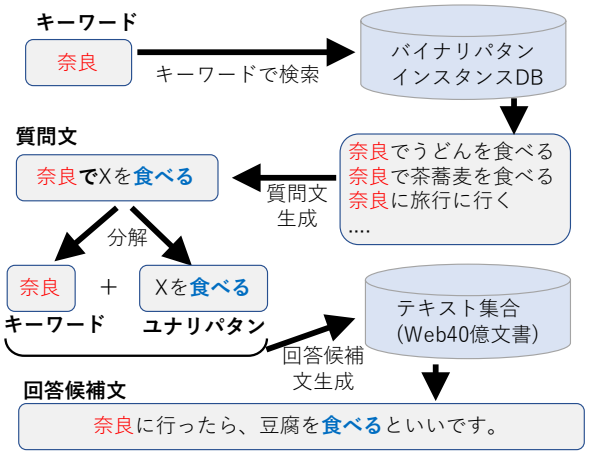

図 1: アノテーション対象の構築方法

頻度の高い名詞 (キーワード) を対象として、WISDOM X の質問サジェスト機能を用い、そのキーワードを含む質問文を自動生成する。具体的には、質問サジェスト機能は、以下のように動作する。まず、その名詞を含むバイナリパタンの事例を検索する (バイナリパタンとは、「A が B になる」「A で B を食べる」のように、述語およびそれと係り受け関係にある 2 つの格要素から成り、名詞部分が変数となっているパタンを指す)。例えばキーワードが「奈良」であれば、「奈良でうどんを食べる」「奈良に旅行に行く」のような事例が検索される。次に、このようなバイナリパタンから得た事例を名詞の意味クラスごとに集計し、頻度の高い組み合わせに基づいて「奈良で X を食べる」「奈良に X に行く」のように変数を 1 ヶ所含むパタンを生成する。これらのパタンを「奈良で何を食べる」「奈良に何に行く」という質問に対応するものと考え、質問文と呼ぶ。ここでは、質問文のうち、キーワードが収集対象の格を取っているもののみを利用する (例えばデ格ゼロ照応を収集したい場合、「奈良 で X を食べる」のような例を利用する)。

次に、WISDOM X の質問応答機能を用い、質問に対する回答候補文を得る。WISDOM X では回答候補文は複数の方法で取得されているが、ここでは以下の方法で取得されたもののみを利用する。まず、質問文をキーワードとユナリパタンに分解する (ユナリパタンは述語およびそれと係り受け関係にある 1 つの格要素から成り、名詞部分が変数となっているものを指す)。上述の例の場合、質問文はキーワード「奈良」とユナリパタン「X を食べる」に分解される。この 2 つを利用し、キーワードとユナリパタンの両方を含む文を検索する。この結果、例えば、「奈良に行ったら、豆腐を食べるといいです。」のように、キーワード「奈良」が述語「食べる」と離れた位置に出現し、かつ、「奈良」が述語「食べる」に対して収集対象とする格関係 (ここではデ格) となる可能性が高い文 (回答候補文) が得られる。この回答候補文が、「奈良で豆腐を食べる」という内容を持つかどうかを判断することで、述語「食べる」に対して、「奈良」がデ格ゼロ照応となる事例を容易に収集することができる。

3 デ格ゼロ照応を対象とした事例収集

本研究では、ゼロ照応の検出が困難な格の一例として、デ格のゼロ照応のアノテーションを実施した。

3.1 回答評価アノテーション

本研究ではまず、Web 6 億文書コーパス [10] においてデ格名詞として現れやすい上位 5,000 名詞をキーワードとし、前述の方法でそのキーワードを含む質問を自動生成した。そのうち、キーワードがデ格で現れている質問を対象に、述語の自然な頻度分布を反映するようランダムに質問を選択してシステムに与え、30,000 事例の質問と回答の組み合わせを取得した (質問によってはユナリパタンによる回答が得られない場合もあるため、30,000 事例が揃うまで繰り返し質問を追加した)。この 30,000 事例に対し、各事例につき独立に 3 名 (1,000 事例ごとに一部担当者が交替し計 16 名) のアノテータ (著者とは異なる) が、質問応答システムの返した回答が適切かどうか (回答候補文に、質問に対する回答となる情報が実際に含まれているかどうか) についての評価作業を行い、3 名の多数決を最終的な評価結果とした [1]。作業には概算で 100 人日を要した。

3.2 言い換え評価アノテーション

第二段階のアノテーションとして、言い換え評価を行った。このステップは、以下の理由で必要になる。

WISDOM X は 「X を購入する」→「X を買う」のような言い換え知識を用いて質問と回答候補文の照合を行っている [4]。例えば、「コンビニで何を買う」という質問に対し、「X を購入する」→「X を買う」という言い換え知識を適用することで、「コンビニで野菜ジュースを購入した」から「野菜ジュース」という回答を得ることが可能である。2.2 節で示したユナリパタンから回答候補文を得る際には、カバレッジの向上のため、この言い換え処理を行いながら回答候補文を取得している。つまり、例えばユナリパタンが「X を買う」の場合は「X を購入する」などの言い換え可能なパタンに関しても回答候補文を得ている。

この言い換えの適用が失敗している可能性があるため、上述の回答評価アノテーションで回答候補文が適切でないと判断されても、言い換えの誤りなのかゼロ照応の誤りなのかそのままでは判断することができない。例えば次の例では、「X を作る」→「X を醸す」という言い換えが適用されているが、言い換えがこの文脈では適切でないため、回答候補文が適切ではないと判断されてしまう。

- 質問文: 農村で X を 醸す

[1] なお、回答候補文が事実的でないモダリティをもつ場合 (例えば「奈良に行ったらおいしいお寿司は食べられますか？」のような疑問文など) は、質問応答システムの目的からいえば除外すべきとも考えられる。しかし、本研究で収集したいゼロ照応の事例としては、モダリティとは無関係に正例となるため、事実として述べられているかどうかにかかわらず言及があれば正例として分類するようにした。

- 回答候補文: 小柳さんは施設の維持費が 1 ／ 4 になるよう設計し、これまで農村地帯 8 ヵ所に汚水処理施設を 作って きた。

このようなケースを排除するため、回答評価で不適切と判断された事例のうち、言い換えが適用されたものに限定して、言い換えが適切かどうかのアノテーション作業を行った。アノテーションは各事例につき独立に 3 名 (1,000 事例ごとに一部担当者が交替し計 13 名) のアノテータ (著者とは異なる) が行い、多数決で最終的な言い換えの可否を決定した (概算で 40 人日を要した)。その結果、言い換えが適切だと判断された事例のみをゼロ照応の負例として扱い、それ以外は破棄した。回答評価、言い換え評価の手順をまとめると図 2 のようになる。

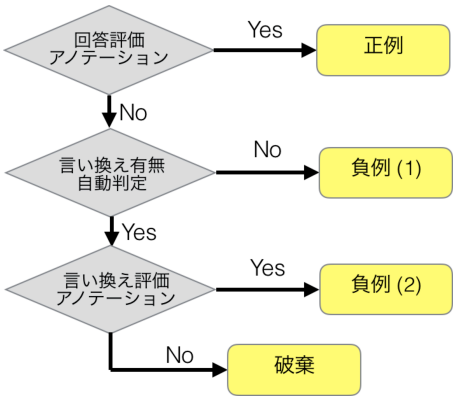

図 2: アノテーションの流れ

4 アノテーション結果

回答評価アノテーションの結果、正例 6,136 事例、負例 23,864 事例が得られた。負例 23,864 事例のうち、言い換えが行われていない 11,567 事例はそのままゼロ照応の負例として利用した。言い換えの行われている 12,297 事例について第二段階のアノテーション (言い換えの評価) を行い、言い換えの正例 4,261 事例、言い換えの負例 8,036 事例を得た。

さらに、後処理として以下のようなものを機械的に除外した。

- 除外事例 A: WISDOM X から得た回答候補文には形態素・係り受け解析が含まれていないため、再度解析を行う必要があるが、形態素解析結果や文節チャンキング結果の違いによって質問文中の述語との照合に失敗する場合 (1,850 例)
- 除外事例 B: 先行詞が合成語の非主要部であるもの。例えば「鹿児島$_i$ 実業高校で [$\phi_{i_{デ}}$] 学んだ」のようなものはゼロ照応の事例から除外した。(3,903 例)
- 除外事例 C: 先行詞と述語が係り受け関係にあるもの。例えば「この 本$_i$ を [$\phi_{i_{デ}}$] 勉強した」のようなものは除外した。(4,276 例)

除外ののち得られた事例数は表 2 の通りである。全体 (30,000 事例) のうち、ゼロ照応の正例として収集できたものは 3,070 事例 (10.2%) であった。

表 2: ゼロ照応の事例数

	除外前	除外事例			除外後
		A	B	C	
正例	6,136	492	639	1,935	3,070
負例 (1)	11,567	1,051	2,385	1,827	6,304
負例 (2)	4,261	307	879	514	2,561
破棄	8,036				
合計	30,000				

アノテーションの質を評価するため、回答評価、言い換え評価それぞれについて κ 値 (Fleiss) によってアノテータ間の一致率を調査した。回答評価の κ 値 (Fleiss) の平均は 0.555 で moderate agreement、第二段階の言い換え評価については κ 値 (Fleiss) の平均は 0.668 で substantial agreement となった (κ 値の解釈は [6] に従う)。

5 他のアノテーション手法・結果との比較

5.1 単純なアノテーション作業による比較実験

ゼロ照応の事例収集のためのより基本的な方法として、事前のアノテーション対象の絞り込みを行わず、コーパス中でゼロ照応が生じている箇所をアノテータが見つけ出し、直接アノテーションする方式が考えられる。本研究で提案する手法がゼロ照応の事例収集をどの程度効率化しているかを示すため、この基本的な方法と比較する実験を行った。

この比較実験では、コーパス内の指定された述語に対し、デ格の関係にある名詞を直接記入してもらう作業を実施した。例えば、下の例であれば、下線部分の「高まる」に対してデ格の関係にあると考えられる名詞を列挙する (この場合「しょうが汁」が該当する)。

- 「魚のたんぱく質は加熱することで吸収が高まり、しょうが汁を使うと中性脂肪減少効果も 高まる。」

6 億文書コーパスのうち約 100 万文から成るサブセットを用意し、ここから述語の自然な分布を反映するようランダムに 1,000 文抽出した。この 1,000 文について、3 名のアノテータ (著者とは異なる) が独立に作業を行い、複数名が挙げた名詞を正例として収集した。

アノテータの作業品質を見積るために作業の一致率を調査した。このアノテーション作業では、作業対象が事前に決まっていないため、κ 値による一致率の調査を行うことができない。そこで、アノテータの各ペアの作業結果に対し、一方を正解データ、もう一方をシステムの出力とみなし、F 値を計算することで作業品質を調査した。この結果、アノテータの全組み合わせの F 値の平均は 0.442 という結果を得た [2]。

係り受け関係にあるものはゼロ照応の事例とは見なさないため、これを除外すると収集できた正例は文数 34 件 (語数 36 件) にとどまった。ゼロ照応の正例が収集できたのは全体の 3.4%となった。提案手法では正例の割

[2] この数値は再現率 5 割、精度 5 割の結果得られる F 値より悪い結果となっており、このアノテーションの結果が良いとは言い難い。

合は 10.2%であったので、単純な手法に比べ、提案手法は約 3 倍の効率があることが示された。

5.2 京都大学テキストコーパス

本節では、既存の言語資源との比較を行う。京都大学テキストコーパス (以下、京大コーパス) には関係タグが付与されており [5]、デ格ゼロ照応の情報も含まれている。ここでは、本研究におけるデ格ゼロ照応の正例の頻度と、京大コーパスにおけるデ格ゼロ照応の頻度を比較する。表 3 は、京大コーパスにおけるデ格ゼロ照応の頻度をまとめたものである [3]。

表 3: 京大コーパスにおけるデ格ゼロ照応

文内デ格ゼロ照応をもつ述語	311
文間デ格ゼロ照応をもつ述語	189
デ格ゼロ照応をもつ述語 計	500
述語	14,987

アノテーション対象である述語のうちデ格ゼロ照応がアノテーションされたものは 2.1% (311/14,987) であり、本研究のアノテーション対象となっていない文間照応を含めても 3.3% (500/14,987) にとどまる。本研究で収集されたデ格ゼロ照応の事例数は、既に京大コーパスの規模を大きく上回っている。なお、同じ基準でガ格について京大コーパスを調査すると、ガ格ゼロ照応をもつ述語の数は 2,713 事例 (述語全体に占める割合は 18.1%) あり、ガ格に比べ、デ格のゼロ照応で同規模の事例数を確保するのが難しいことがわかる。

6 ゼロ照応解析手法における利用

本研究で収集されたデ格ゼロ照応の事例に対し、既存のゼロ照応解析手法 [3] を適用することで、どの程度の性能が得られるかの調査を行った。アノテーション作業の結果得られた全事例 (ただし、破棄されたものは除く) を、7,611 事例を学習事例、2,498 事例を validation 用事例、2,539 事例を評価用事例の 3 種類に分割し評価を行った [4]。この結果、再現率 0.564、精度 0.398、F 値 0.466 という結果を得た。この評価実験では、アノテーションされた述語と名詞の対のみを評価対象としているため、既存の評価結果と直接比較することは困難であるが、例えば、Ouchi ら [9] のヲ格とニ格のゼロ照応解析の結果は F 値で 0.244 と 0.048 であり、本研究で得られた学習事例を使って学習した結果で、このような低い性能を改善できる見込みがある。ただし、F 値で 0.466 という性能はガ格のゼロ照応解析は Iida ら [3] の結果 (F 値で 0.525) と比較して低いため、デ格など任意格に特化した解析手法を考える必要もあると考えられる。

[3] 京大コーパスでは合成語の構成要素間等にも関係タグが付与されているが、基準を本研究と合わせるため、述語の格要素となる場合のみカウントした。また、同一の述語に対して複数のデ格ゼロ照応がある場合があるため、ゼロ照応の数は表で示したゼロ照応をもつ述語の数より若干多く、計 525 事例である。

[4] キーワードが文中に複数回出現している場合があるため、合計は表 2 の正例と負例を合わせた数より多くなる。

7 おわりに

本研究では、質問応答システムを利用してアノテーション対象の事前の絞り込みを行うことで、ゼロ照応の事例が効率的に収集できることを示した。

ただし、本稿で提案したアノテーションの手順では、言い換えの評価などの結果、正例と負例のいずれにも利用できない事例が全体の半数以上になるなど、作業結果に無駄が多いという問題がある。この問題を解決するために、例えば、言い換えを含む文を作業対象から除外するなど、アノテーションの手順を改善した後に、より大規模なアノテーション対象に対して作業を実施する予定である。また、本研究ではアノテーションの一例としてデ格のゼロ照応のみを対象としたが、ヲ格、ニ格などのゼロ照応に対しても作業を実施し、その結果得られた事例を利用して学習することでより頑健なゼロ照応解析のシステムを研究開発する予定である。

参考文献

[1] Chen Chen and Vincent Ng. Chinese Zero Pronoun Resolution with Deep Neural Networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 778–788, 2016.

[2] Joseph L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, Vol. 76, No. 5, pp. 378–382, 1971.

[3] Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, Canasai Kruengkrai, and Julien Kloetzer. Intra-Sentential Subject Zero Anaphora Resolution using Multi-Column Convolutional Neural Network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1244–1254, 2016.

[4] 川田拓也, Kloetzer Julien, 鳥澤健太郎, 橋本力. 質問応答システムのための含意パターンペアの生成. 言語処理学会第 21 回年次大会発表論文集, pp. 159–162, 2015.

[5] 河原大輔, 黒橋禎夫, 橋田浩一. 「関係」タグ付きコーパスの作成. 言語処理学会第 8 回年次大会発表論文集, pp. 495–498, 2002.

[6] J Richard Landis and Gary G Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, Vol. 33, No. 1, pp. 159–174, 1977.

[7] 松林優一郎, 吉野幸一郎, 林部祐太, 中山周. Project NEXT 述語項構造タスク. 言語処理学会第 21 回年次大会ワークショップ: 自然言語処理におけるエラー分析, 2015.

[8] Junta Mizuno, Masahiro Tanaka, Kiyonori Ohtake, Jong-Hoon Oh, Julien Kloetzer, Chikara Hashimoto, and Kentaro Torisawa. WISDOM X, DISAANA and D-SUMM: Large-scale NLP Systems for Analyzing Textual Big Data. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016) (Demo Track)*, 2016.

[9] Hiroki Ouchi, Hiroyuki Shindo, Kevin Duh, and Yuji Matsumoto. Joint Case Argument Identification for Japanese Predicate Argument Structure Analysis. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 961–970, 2015.

[10] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology. In *Processings of the 3rd International Joint Conference on Natural Language Processing*, pp. 189–196, 2008.