

言語学的テストに基づく意味アノテーションのガイドライン設計 — 確実性判断に関わる表現を中心に —

田中 リベカ[†] 小池 恵里子[†] 戸次 大介[†] 川添 愛[‡]

[†] お茶の水女子大学

[‡] 津田塾大学／国立情報学研究所

{g0920522,g0920514,bekki}@is.ocha.ac.jp zoeai@nii.ac.jp

1 はじめに

本論文では、コーパスに対する意味アノテーションの質を向上させる方法について議論する。言うまでもなく、アノテーション済みコーパスの信頼性は、それを基盤として構築される言語処理システムの信頼性に直結する重要な要素である。コーパスの信頼性を左右する大きな要因の一つとして、アノテーションガイドラインの設計がある。特に、意味を対象とするアノテーションにおいては、言語表現の意味の多義性解消や、用法の特定などにおいてアノテータの正確かつ一貫した判断が求められるため、ガイドラインにおいて明確な判断基準や手段を提示する必要がある。それらを欠いている、あるいはそれらの提示に成功していないガイドラインは、アノテータ間での判断の不一致などの問題を引き起こす。

アノテーションの方法論は Hovy and Lavid (2008), Fort et al (2009) 等で取り上げられているが、本論文で論じるのは、アノテーションガイドラインの設計における言語学的テストの利用である。ここで言う「言語学的テスト」とは、理論言語学の理論構築および検証に用いられるテストで、文や文の一部の容認性や適切性を判定するものである。言語学的テストは、容認性や適切性を左右する条件を特定するために、検証したい部分以外の条件をほぼ同じにした二つ以上の文からなる群として提示されることが多い。本論文では以下、第一節で意味アノテーションの作業を一種の「分類タスク」と見た場合に、適切な言語学的テストの有無が判断の一貫性を左右することを、実例を交えて論じる。第二節では、アノテーションに有効な言語学的テストの設計についての一般的な方法論を提案し、第三節でその適用例を見る。

なお、本論文にて「意味アノテーション」の具体例として用いるのは、「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」(川添ら (2011))に基づくアノテーション作業である。このガイドラインは、様相表現・条件表現・否定表現など、言語情報の確実性に影響する表現とそのスコープにアノテーションを付与し、機械による確実性判断の

基盤となるコーパスを構築するために作成されたものである。

2 コーパスアノテーションにおける言語学的テスト

意味アノテーションに限らず、コーパスに対するアノテーション作業の多くは、テキストの一部に対し、スキーマの設計者があらかじめ決定したラベルを付与していくものである。これはある意味、実際のテキストを、用意されたカテゴリに分類していく作業ととらえることができる。アノテーションガイドラインにはそのような分類を行う際の指示が多かれ少なかれ含まれるわけであるが、その細やかさ・適切さのレベルはさまざまである。

例えば、テキストに現れる「(と) いう」という表現のうち、他人による認識あるいは主張を表すもののみに対してアノテーションをしたいとする。その場合、以下のようにアノテーション対象カテゴリの記述のみを提示し、テキストに現れる個々の「(と) いう」がこれに属するかどうかの判断をアノテータに委ねるのも一つの方法である。

他人の認識【(と) いう】

分類：他人の認識（他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現）

しかしこのような方法は、表現を分類する基準をアノテータが知っていることを前提としており、母語話者であっても専門的な知識を有していない場合は判定が困難である。

これよりも少し進んだ指示としては、以下のように例文を示した上で、アノテーション対象の用法と似ているかどうかを判断させるような、一種のテスト形式と呼べるものもある。

他人の認識【(と) いう】

分類：他人の認識（他人の報告する事柄や、命題の真偽

に関する他人の判断を表す表現)

例：今冬のインフルエンザの流行は全国的に遅れているという。

上のように例文を提示すると、ほぼ同じ用法に関しては専門的な知識を要さずに判定可能であるという利点が生まれる。しかし、表現の形式が変わったり、少し異なる用法と思われる例があると、途端に判断が困難になる。上の例は実際に「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」にある記述を、内容を変えずに再編集したものであるが、筆者らが実際にアノテーション作業を行ったところ、アノテーションの可否が判断できない、あるいは判断が一致しない例に直面した。そのような例の一部を以下に示す。

1. 呼吸困難に陥る可能性があるという。
2. 「彼を絶対に許さない」と言う。
3. 太郎が責任をとるべきという人はどうかしている。
4. 太郎は結婚したという話だ。
5. インスリンというホルモン。
6. 車がガタガタという。
7. お前という人間が信じられなくなった。
8. サプリというサプリは試した。

1. は、アノテーションガイドラインにある例文と似ていることから、「他人の認識を表す『（と）いう』』としてアノテーション可能であると想像がつく。他方、2.-8. は決定的な判断基準がない。特に2.3.4. は、人によってはアノテーション可、不可の判断が大きく分かれた。特に困難であったのは3.4. の判断である。これらはどちらも名詞句を修飾しており、見た目には同じ型をしているが、ガイドラインの設計者によれば、3. は他人の認識としてアノテーション可能だが、4. はアノテーション不可能である。このように、例文ベースの指示では、例文との類似度の判断が困難となる。

そこで筆者らは、言語学的テストを導入し、ガイドラインを以下のように改善した。

他人の認識【(と) いう】

分類：他人の認識（他人の報告する事柄や、命題の真偽に関する他人の判断を表す表現）

例：今冬のインフルエンザの流行は全国的に遅れているという

テスト1：「(と) 述べる (述べられる)」に置き換えても意味が変化しない。

テスト2：名詞句を修飾する場合「～との」に置き換え不可

テスト1、2は、ともに表現の「置き換え」に基づくテストである。これらのテストを上挙げた例に適用する

と、以下のようになる。

1. (テスト1) 呼吸困難に陥る可能性がある(と)述べる(述べられる)。/ (テスト2: 適用不可)
2. (テスト1) 「彼を絶対に許さない」と述べる。/ (テスト2: 適用不可)
3. (テスト1) 太郎が責任をとるべき(と)述べる人はどうかしている。/ (テスト2) *太郎が責任をとるべき(と)の人はどうかしている。
4. (テスト1) *太郎は結婚した(と)述べる(述べられる)話だ。/ (テスト2) 太郎が結婚した(と)の話だ
5. (テスト1) *インスリン(と)述べる(述べられる)ホルモン。/ (テスト2) *インスリン(と)のホルモン
6. (テスト1) *車がガタガタ(と)述べる。/ (テスト2: 適用不可)
7. (テスト1) *お前(と)述べる(述べられる)人間が信じられなくなった。/ (テスト2) *お前(と)の人間が信じられなくなった。
8. (テスト1) *サプリ(と)述べる(述べられる)サプリは試した。/ (テスト2) *サプリ(と)のサプリは試した。

上のように、別の表現と置き換えが可能かどうかを、実際に文を組み立てて考えさせる言語学的テストを導入すると、複数の作業によるYES/NO判定の一致度が向上した。実際に、テスト1、テスト2のいずれかを通らなかった4.-8. はアノテーション不可能であると判断できるようになった。元々困難であった3.と4.の区別も、両方のテストで異なる結果が出ることで明確になった。特筆すべきは、「置き換えが可能」という判断よりも、「置き換えが不可能」というネガティブな判断の方が一致度が高かったことである。このことから、「置き換え可能ならばアノテーション対象である」という指示よりも、「置き換えが不可能であればアノテーション対象ではない」という指示の方が、一貫性のある結果を得るのに有効であるとの結論を得た。

3 言語学的テスト設計の方針

前節で述べた作業は筆者らが行った具体的なアノテーション作業であるが、アノテーションガイドラインを設計するにあたっての重要な示唆を含んでいると考えられる。本節ではこれらの考察を元に、理想的なガイドラインにおいてどのようなテストを設計すべきかを論じる。

ここで、ある表現Eと A_1, \dots, A_n の n 個の分類を考え、Eが n 個の分類中どこに属するかを判定するテストを考える。

表現Eが特定の分類 A_i に入るかどうかの判断基準となるテストを、ここでは「個別テスト」と呼ぶ。また分類 A_1, \dots, A_n の個別テストを集めたものを「テスト

セット」と呼ぶこととし、個別テストとテストセットの2段階における性質を考える。

3.1 個別テストの構築

まず個別テストは、別の表現に実際に置き換えた上で「意味が変わるかどうかが」「違和感があるかどうか」をYES/NOで判定する形式や、「活用しているか」「コト節を受けているか」など外見上明らかな条件を設定することが望ましい。「この表現に似ているか」のような形式をとった場合、判断は個々人の独自の解釈に左右される。しかし置き換えた上で意味が変わるかどうかの判定は感覚に訴えるものであり、母語話者を対象とする限りは専門的な知識を要しないと考えられる。無論、テストを作成するにはどのような語に置き換えて判定させるかについても慎重にならねばならない。例えば慣用的に使われている表現を用いると、必ずしも本質的ではない点で混乱を生む可能性がある。

個別テストには、テストの判定結果が分類の

- 1) 必要十分条件になっている場合
- 2) 必要条件になっている場合
- 3) 十分条件になっている場合

の3パターンがある。

まず、個別テストが必要十分条件になっている場合は、すなわちあるテストでYESの判定が出たときかつそのときに限り当該分類に属すると言い切れる場合である。個別テスト全てがこのような性質を持っているときは、分類 A_1 から A_n まで線形にテストを実行し、YESの判定が出た時点でその分類に属すると決定することができる。ただし、必要十分条件となるようなテストを作ることは一般に困難である。

必要条件となるテストにおいては、Eがある分類 A_i に属すると仮定するとテストでYESの判定が出る。このようなテストには、表現がある分類に属するときの特徴などが該当し、比較的簡単に作成することができる。しかしこのようなテストがあくまでも補助としてしか使えないのは、必要条件であることから明らかである。つまり、テストでYESの判定が出たとしても、その表現がその分類に属するとは言い切れない。可能性としては、ある分類に対して必要条件となるテストが複数存在し、複数ある必要条件を全て満たすような表現は他の分類には属し得ないといえるような場合も考えられるが、現実的に分類を特定するのに十分な個数のテストを作成することは容易ではない。

十分条件となるようなテストは、YESの判定が出たものに関してはその分類に含めてよいというものである。し

かしNOの判定が出たものに関しても、まだその分類に属する可能性が残ることになる。

以上を踏まえ、必要条件となるテストの対偶を取った形式を効果的なテストとして提案する。つまり、必要条件となるテストの判定結果がNOである場合、その分類に属しないとするとテストである。特に、表現の置き換えを利用したテストにおいては、「置き換え可能」という判断よりも「置き換え不可能」という判断の方が一貫性が期待できることから、ある別の表現(分類 A_i に属するとしたら置き換え可能であることが性質として知られているような表現を用いる)に置き換え不可能であったならば絶対に分類 A_i に属さない、ということのみをいえるようにしたテストが最も理想的である。このようなテストを、以下ではネガティブテストと呼ぶ。ネガティブテストではその分類に属さないということしか断定できないが、置き換えが不可能であると判定された場合は、確実にその分類に入る可能性を消去することができる。

例えば、前節で挙げたテスト1は、次のように述べなおした場合にはネガティブテストであることが明確となる。

テスト1: 「(と) 述べる (述べられる)」に置き換え不可能であるような「(と) いう」は、アノテーション対象ではない。

3.2 テストセットの構成

上のような方針で個別テストを設計することを前提とした上で、テストセットとしてどのような性質を満たすべきなのかを考える。

個別テストのみを見た場合、ネガティブテストを集めただけでは必ずしも分類を決定することができないが、テストセット全体として表現の意味の分類が網羅的であれば、適切なネガティブテストを用いて、消去法により表現の分類を特定することが出来る。ここでいう適切なネガティブテストとは、その表現に置き換えられなかったら分類に属さないと言い切れるような強い条件を用いている、又は判断に迷うようなまぎらわしい例を避けているということである。ネガティブテストで構成されたテストセットにおいては、「置き換え不可能か(不可能ならば分類 A_i には属さない)」という問いに対してその分類に本来属する表現であった場合は「置き換え不可能ではない(NO)」の判断がされ、テストセットに属する全てのテストを実行したあとに「NOの判定が出ていてその分類に属する可能性が残っているもの(保留となっているもの)」を表現Eが属する分類(解)とする。言い換えると、例がまぎらわしい等の理由で判断がつかず「保留」とされた分類が複数存在すると解が定まらない。個別テストが不完全なために分類先として複数の候補があがってしまった場合には、テストを再構築する必要がある。

る。一方で分類が網羅的でなかったという理由で適切な分類がされなかった場合には分類を追加しなければならないが、その際には新しい分類に関する個別テストは単純に追加するだけで良い。この性質はネガティブテスト自体が十分条件になっていることに由来する。

この方法は、設計者が適切なネガティブテストと網羅的な分類からなるテストセットを設計すれば、後はアノテータが言語直感に頼って判定をすれば良いという点で有効であると考えられる。無論、必要条件や十分条件となるテストを補助的に利用することも重要であるが、可能な限りネガティブテストでテストセットを構成していくことを明確な指針とすることで、より効率が良く信頼性の高いテストを設計することが可能であると考えられる。

4 個別テスト・テストセット構成例

前節の考察に基づき、1) 適切なネガティブテスト、および 2) 表現の意味の網羅的な分類を含むテストセットの例を表 1 に示す。ここでは、「(と) いう」をその意味に応じて 7 つのカテゴリに分け、テキストにおけるこの表現の個々の出現例がどのカテゴリに属するかを判定するためのネガティブテストを四つ用意している。これが完全に網羅的な分類であるかは現時点では不明であるが、この分類のみに関して言えば、ネガティブテストの組み合わせによってどのカテゴリに属するかを判断することが可能である。

5 おわりに

以上では、意味アノテーションにおける言語学的テストの利用、およびその際の方法について論じた。上で述べた方法論に基づき、「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライ

ン Ver.2.4」に含まれるものを中心に、340 種類の様相表現、13 種類の否定表現、48 種類の条件表現の網羅的な分類とネガティブテストの構築を行い、ガイドラインの改良を行う予定である。

謝辞

本研究は、平成 23 年度国立情報学研究所共同研究「自然言語テキストにおける確実性判断のための意味計算体系の構築」(研究代表者: 川添愛) の助成によるものである。また、国立情報学研究所人工頭脳プロジェクト「ロボットは東大に入れるか」NLP コア定例ミーティングにて、メンバーの方々に貴重なコメントをいただいた。

参考文献

- [1] Hovy, E., Lavid, J. Corpus Annotation: Framework and Exercises. LREC2008 tutorial
<http://www.lrec-conf.org/proceedings/lrec2008/workshops/T1.pdf>
- [2] Fort, K., Ehrmann, M., Nazarenko, A. Towards a Methodology for Named Entities Annotation Proceeding of the 3rd ACL Linguistic Annotation Workshop (LAW III), 2009.
- [3] 川添愛・齊藤学・片岡喜代子・崔榮殊・戸次大介「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」, Technical Report of Department of Information Science, Ochanomizu University, OCHA-IS 10-4, 2011.

表 1: テストセット構築例

カテゴリ	例文	ネガティブテスト	アノテーションの可否
(〜と) いう 1	花子は、彼を絶対に許さないという。太郎が責任をとるべきという人は、どうにかしている。	test1: 「述べる (述べられる)」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。 test3: 過去形「(と) いった」に置き換え不可な場合は、このカテゴリではない。	アノテーションする
(〜と) いう 2	ニュースによると、インフルエンザが流行っているという。	test1: 過去形「(と) いった」に置き換え不可な場合は、このカテゴリではない。	アノテーションする
(〜と) いう 3	山田さんという人が来た。血糖値を下げるのは、インスリンというホルモンだ。	test4: 「呼ばれる」に置き換えて意味が変化する。あるいは置き換え不可の場合は、このカテゴリではない。	アノテーションしない
(〜と) いう 4	あの人はぶつぶつ(と)言っている。車がガタガタ(と)いう。	test3: 過去形「(と) いった」に置き換え不可な場合は、このカテゴリではない。擬音語的な表現を取り去ると意味をなさない。(十分条件) *あの人は言っている。(比較: あの人は言っている。) *車がいう。(比較: 車がガタガタいう。)	アノテーションしない
(〜と) いう 5	太郎が結婚したという話だ。時期尚早であるという見解を示した。	test2: 「との」に置き換え不可な場合は、このカテゴリではない。	原則としてしない*
(〜と) いう 7	お前という人間が信じられなくなった。長年住んでいるが、東京という町には親しみがわいてこない。	test3: 過去形「(と) いった」に置き換え不可な場合は、このカテゴリではない。	アノテーションしない
(〜と) いう 8	サブリというサブリは試した。医者という医者には相談したものの、どうにもならなかった。	test4: 「呼ばれる」に置き換えて意味が変化する、あるいは置き換え不可の場合は、このカテゴリではない。	アノテーションしない

* 「話」「噂」「見解」などについて。「ということだ」「とのことだ」全体にはアノテーションする。