2C3-2

様相・否定・条件表現の言語学的分析に基づく確実性アノテーションスキーマの設計

Designing annotation schema for certainty based on linguistic analysis of modal, negative and conditional expressions

川添 愛[*1] Ai Kawazoe
齊藤 学[*2] Manabu Saito
片岡 喜代子[*3] Kiyoko Kataoka
崔 栄殊[*4] Young Soo Choi
戸次 大介[*5] Daisuke Bekki

[*1] 津田塾大学 Tsuda College
[*2] 中華大学 Chung Hua University
[*3] 九州大学 Kyushu University
[*4] 一橋大学大学院 Hitotsubashi University
[*5] お茶の水女子大学 Ochanomizu University

Texts in natural language contain not only factual assertions but also uncertain information, such as speculations, inferences and hypothetical thoughts. This paper presents a schema design that allows us to construct corpora with modal, negative, and conditional expressions being annotated. Those elements can be a clue for readers to recognize the incredibility of given information. With the aim of developing a general automatic system of certainty annotation, we provide some discussions based on linguistic/practical considerations.

1. 目的

自然言語のテキストには、以下に見られるように、事実以外に、推測、仮定、仮想現実など、テキストの書き手にとって事実であるかどうかが不明な情報も含まれる。人間は、自然言語で書かれた情報を読むとき、さまざまな知識を駆使して「この情報の信憑性はどのくらいか」「この情報の発信者はどれほどの確信を持っているか」などの判断をある程度行うことができる。機械による確実性の判断を可能にしたい場合、人間が確実性の判断をする際に意識的あるいは無意識的に利用している知識(の少なくとも一部)を、機械に利用可能な形で与えることは、自然かつ有効なアプローチであるように思われる。

(1) 県内で新型インフルエンザが発生した。(事実)
(2) 県健康推進課が県内で新型インフルエンザが発生したと報告した。(伝聞)
(3) 県内で新型インフルエンザが発生したとみられる。/県内で新型インフルエンザが発生した可能性がある。/県内で新型インフルエンザが発生した可能性は低い。(推量)
(4) まるで県内で新型インフルエンザが発生したようなパニックが起こっている。(比況)
(5) 県内で新型インフルエンザが発生したわけではない。(通常否定)/ 県内で新型インフルエンザが発生したというのは正しくない。(メタ否定)
(6) 県内で新型インフルエンザが発生したら、どう対応するべきか。(仮定) /あの時県内でインフルエンザが発生していたら、パニックになっていただろう。(反実仮想)

筆者らは、人間が言語情報の確実性判断の際に利用している、様相(文の内容に対する書き手の認識・判断)を表す表現、否定表現、条件表現などに関する知識に着目し、これらの意味的性質・統語的性質の情報をタグ付した言語データ(アノテーション済みコーパス)を作成している。このコーパスは、機械による確実性判断の基盤という実用的な用途のために構築されるものであるが、同時に、テキストのタイプや、確実性の判断を必要とするユーザーのニーズの違いなどに関係なく、さまざまな種類のテキストや用途に利用できる有効なアノテーションを行うことを目指している。筆者らは、「確実性」の標識となる言語表現の持つ特性のうち、どのような文脈においても維持されるような統語的特性および意味的特性をアノテーション仕様に組み込むことで、多様な内容の情報に対応できるようなアノテーションが可能になると考えている。そしてそれらの「統語的特性・意味的特性」を見極めるために様相表現、条件表現、否定表現に対する言語学的分析を行っている。本論文では、アノテーションスキーマの設計に関わる理論面と応用面の議論を紹介する。

連絡先:川添愛, 津田塾大学女性研究者支援センター, 〒187-8577 東京都小平市津田町 2-1-1, 042-342-5142, zoeai@tsuda.ac.jp

2. アノテーションスキーマの概要

機械による言語情報の確実性判断には、実用面でのニーズがある。たとえば、現在、感染症発生情報のソースとして Web 上のニュースなどが利用されているが、人間が短時間で情報の信憑性および緊急性を判断しなくてはならない。もし機械が確実性に影響する意味的文脈を認識できれば、不要な情報を取り除いて検索範囲を狭め、より効率的な情報抽出が行える上、情報の確実性判断に関わる人的コストを減らせると考えられる。

言語情報の確実性判断を目指した言語処理研究は、主に英語の文献を対象に、推測や意見を事実の記述から区別するタスク(hedge classification)が過去数年間に開始されている(Lignt et al (2004)、Medlock and Briscoe (2007)、Szarvas et al(2008)、Kilicoglu and Bergler (2008)等)。日本語では江口ら(2009)による判断情報アノテーションの研究がある。

先行研究のいくつか(Lignt et al (2004)、Medlock and Briscoe (2007)、Szarvas et al(2008)、江口ら(2009))は、テキストに対して意味的な情報をタグ付けしたコーパスを構築し、利用している。このようなアプローチにおいては、テキストのどの部分に対してどのような情報を、どのようなルールにしたがってアノテーションするかを定めた仕様が重要になる。アノテーション仕様の設計では、アノテーション済みコーパスの用途やテキストの種類などはもちろん考慮に入れる必要があるが、一貫性のあるアノテーションをどう実現するかということには特に配慮が必要である。特に、人手でアノテーションを行う際には、作業者がアノテーション

スキーマに従って常に適切な判断ができることが理想的であるが、そのようなスキーマを設計することは容易ではない。

筆者らは、テキストを構成する文（および従属節の一部）に対してその確実性の度合いを適切に記述し、また極力一貫性のあるアノテーション結果を得ることを目指し、以下の点に留意して仕様の設計を行った。以下の各節で順次詳しく述べる。

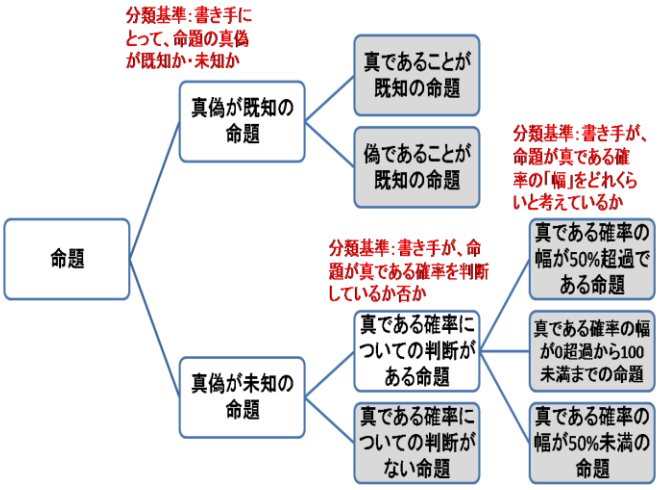
- 「確実性」に基づいて命題を分類する
- アノテーションの対象を、確実性に影響する言語表現とそのスコープとする
- アノテーション対象の言語表現を、確実性という観点から言語学的な手法によって下位分類し、命題の分類と対応させる

3. 確実性に基づく命題の分類

本論文では「確実性」という言葉を、テキストの書き手が命題の内容を「真」と考えている確率という意味で使う。これは、完全に客観的な確実性とは異なる。つまり、情報の受け取り手にとっての情報の「信頼性（credibility）」とは、深い関わりはあるものの、異なる概念であることを注意しておきたい。情報の信頼性には、加藤・黒橋・江本(2006)が指摘しているように、発信者の信頼性などさまざまな要因が関わる。本論文で扱う「確実性」とは、それらの要因の一つであるところの、書き手が情報と事実の間の距離をどれくらい近いと考えているかということである。また、「事実」という言葉は、誤解が生じない限り、「テキストの書き手にとって真であることが既知であるような情報」を指して使う。

筆者らは、確実性を基準に、以下のように命題を分類した。

(7) 命題の分類（オントロジー）



筆者らの最終的な目標は、テキスト中のあらゆる命題を、(7)のオントロジーに従って分類することである。このようなオントロジーを使うことで、分類基準を明確にし、かつ一つの命題が複数の命題のクラスに分類されるのを防ぐことを目指している。ここでは、「真であることが既知」と「真偽が未知だが、真である確率（確実性）を 100%と判断する状態」は区別することに注意されたい。「偽であることが既知」と「真偽が未知だが、真である確率（確実性）を 0%と判断する状態」についても同様である。また、「真である確率についての判断がある命題」の三つの下位クラスは、命題を、それが「真である確率」が具体的に何%かではなく、値のとりうる「幅」が何%から何%までかに基づいて分類するものである。命題の確実性を特定の数値ではなく「数値のとりうる幅」で指定する理由は、「新型インフルエンザが今年流行する確率は 53%である」のように計算によって得られた具体的な値に言及されている場合を除いて、言語情報のみから特定の数値を得るのは困難だからである。たとえば、「新型インフルエンザが今年流行する可能性が高い」という文だけからは、この文の書き手が考える「確率」をただ一つの値に特定することはできない。ただし、後述するように、「可能性が高い」という表現の性質を調べることで、「50%超過で 100%未満」という「幅」を特定することは可能である。この点について詳しくは第 5 節で述べる。

4. 様相・否定・条件表現とそのスコープに対するアノテーション

本研究では、各文を直接(7)に従って分類し、結果をアノテーションするという手法はとらない。まず、命題の確実性に影響する表現とそのスコープに対するアノテーションを行う。そののち、各スコープの確実性を計算することで、上に従った分類を得る。最初に表現に対するアノテーションを行う理由は、上述の命題の分類と言語表現との対応を明確にすることで、アノテータによる不明瞭な判断を極力防ぐためである。アノテーション対象の表現のいくつかには多義性があるため、完全に明確な対応ではないが、後述する通り、言語学的な分析に基づくテストが利用できる。

表現に対するアノテーションは、様相表現に対するもの、否定表現に対するもの、条件表現に対するものの三つがある。それぞれ、クラス名は MODAL, NEG, COND となる。属性は、わずかな例外を除いて、タグの識別番号を値にとる id 属性、表現の下位分類を示す type 属性、スコープの id を示す scope 属性、統語範疇を示す pos 属性の 4 つである。スコープに対するアノテーションのクラス名は SCOPE である。SCOPE クラスの属性は、識別番号を値にとる id 属性、スコープ内で記述される出来事の起こる時間が、書き手がテキストを書いた時間基準として未来に属するか、非未来（現在および過去）に属するかを記述するための time 属性の二つである。

以下にアノテーションの例をいくつか示す。

(8) このうち 10 人以上がゴールデンウイーク中に横手市の秋田ふるさと村を訪れており、<SCOPE id="009" time="non-future">イベントで動物に接触したことによる経口感染</SCOPE>が<MODAL id="010" type="epistemic_1_99" scope="009" pos="verb">疑われている</MODAL>。（秋田魁新聞 2006/06/18）

(9) <SCOPE id="0001" time="future">3 人の退院は早くても 17 日午後になる</SCOPE><MODAL id="0002" type="evidential" scope="0001" pos="noun">見通し</MODAL>。（読売新聞 2009/5/15）

5. 言語学的な考察に基づく表現の分類

本研究では、様相表現、否定表現、条件表現を表(10)のように分類した。様相表現の分類は、(7)の命題の分類に従っている。過去の研究における様相および様相を表す表現の分類（Palmer (2001)など）では、叙実表現や比況表現などは含まれないことが多い。しかし、これらの表現は、命題の「確実性」といった基準に照らし合わせると、「真（あるいは偽）であることが既知の命題」を導入する表現であることから、本研究ではこれらの表現も分類の対象に含めている。

(10) 表現の分類

様相表現	ラベル	表現の例	命題の分類(7)との対応
叙実表現	factive	（～ことを）知る、わけだ、あいにく、幸い、事実、ではないか	真であることが既知
証拠推量表現	evidential	ようだ、みたいだ、らしい、見込み、模様	真であることが未知―真である確率についての判断あり―真である確率の幅が 50%超過
認識的推量表現	epistemic_100 （確実性 100%）	絶対、100%、必ず、絶対に、間違いなく	真であることが未知―真である確率についての判断あり―真である確率の幅が 50%超過
	epistemic_51_100（確実性 50%超過～100%以下）	だろう、（～に）違いない、はずだ	
	epistemic_51_99 （確実性 50 »超過～100%未満）	可能性が高い、おそらく、多分、きっと、	
	epistemic_1_99（確実性 0%超過～100%未満）	かもしれない、可能性がある、のではないか、ひょっとしたら	真であることが未知―真である確率についての判断あり―真である確率の幅が 0 超過～100%未満
	epistemic 1_49（確実性 0%超過～50%未満）	可能性は低い、おそれは低い	真であることが未知―真である確率についての判断あり―真である確率の幅が 50%未満
	epistemic_0_49（確実性 0%以上～50%未満）	まい、（書き手が～と）思わない	
	epistemic 0（確実性 0%）	可能性はない、おそれはない、	
	epistemic_X（確実性 X%）	可能性は X%（だ）、確率は X%（だ）	真であることが未知―真である確率についての判断あり
他人の認識を表す表現	hearsay	（に）よると、（と）いう、（と）する	真であることが未知―真である確率についての判断なし
不定判断・疑問表現	unknown	か、どうか、かな、かしら、？	真であることが未知―真である確率についての判断なし
比況表現	simile	まるで、ようだ、みたいだ	偽であることが既知
否定表現	**ラベル**	**表現の例**	**命題の分類(7)との対応**
通常否定表現	normal	わけではない、のではない、ということはない	偽であることが既知
メタ否定表現	meta	のではない、ということはない、嘘である、間違いである、正しくない	偽であることが既知
条件表現	**ラベル**	**表現の例**	**命題の分類(7)との対応**
事実的条件表現	factual	たら、なら（ば）	（前件・後件ともに）真であることが既知
予測的条件表現	cond_epistemic_100 （確実性 100%）	たら、時	（前件・後件ともに）真であることが未知―真である確率についての判断あり―真である確率の幅が 50%超過
認識的条件表現	cond_epistemic_0_99（確実性 0%以上～100%未満）	たら、なら（ば）、れば、時、場合、とすると、仮に、もし、とする、仮定する	（前件・後件ともに）真であることが未知―真である確率についての判断あり―真である確率の幅が 0 超過～100%未満
	cond_epistemic_0_49（確実性 0%以上～100%未満）	万一	（前件・後件ともに）真であることが未知―真である確率についての判断あり―真である確率の幅が 50%未満
反事実的条件表現	counterfactual	たら、なら（ば）、れば	（前件・後件ともに）偽であることが既知

表現の上位の分類は、既存の言語学的分析に従っている。たとえば証拠推量表現と認識的推量表現の区別は、Palmer(2001)に従い、前者を「話し手（書き手）が、推量の根拠の存在を示すもの」とし、後者を「現実の可視的状況と分岐した別の状況（離れた場所、未来、仮想など）の構成に関わる言明」（田窪(2001)）を導入する表現としている。これらの間の区別には、田窪(2001)で紹介されている「今ごろ」等を使った反実仮想文脈及び「どうやら」「どうも」等を用いた「推量の証拠が存在する文脈」を利用している。

【反実仮想文脈】

(11) 彼が本当のことを言っていたら、今ごろはもう犯人がつかまっている{***ようだ /*らしい /*みたいだ / だろう /かもしれない /可能性がある**}。

【証拠が存在する文脈】

(12) {どうやら・どうも}彼はうそをついている{**ようだ /らしい / みたいだ /* だろう /*かもしれない /*可能性がある**}。

他方、認識的推量表現の下位分類には、筆者らが新たに考案したテストを用いた。認識的推量表現の分類においては、まず「絶対」「必ず」「100%」のような表現を「真である確率が 100%の命題を導入する表現」（以下、epistemic_100）、かつ「可能性がない」「確率は 0%だ」を「真である確率が 0%の命題を導入する表現」（epistemic_0）と考える。その上で、その他の表現について、これらの表現と共起できるかどうかを観察する。共起できるものはそれが導入する命題の確実性の「幅」の中に 100%（あるいは 0%）を含み、それ以外のものは 100%（あるいは 0%）を含まないと判断した。

【「必ず/絶対/100%」との共起】

(13) 必ず/絶対/100% 来る{**であろう /にちがいない /はずの**}人に招待状を出す必要はありません。

(14) *必ず/絶対/100% 来る{**可能性が高い /可能性がある / かもしれない**}人に招待状を出す必要はありません。

【「可能性はない」「確率は 0%だ」との共起】
(15) 太郎が来る{確率は 0%だ/可能性はない}。#太郎は来る{**かもしれない /可能性はある /可能性は低い**}。
(16) 太郎が来る{確率は 0%だ/可能性はない}。太郎は来る**まい**。

ある表現が導入する命題の確実性が必ず 50%超過(あるいは 50%未満)であるか否かを判定する手段としては、「同じ命題の『肯定+認識的推量表現』と『否定+認識的推量表現』が同時に主張できるかどうか」というテストを用いている。というのは、ある命題が真である確率が 50%を超えている(あるいは 50%に満たない)ということと、同じ命題が偽である確率が 50%を超えている(あるいは 50%に満たない)ことを同時に主張することはできないからである。このテストで容認性が著しく下がるもの、たとえば以下の「だろう」などは、50%を含まない(なおかつ、(13)のテストの結果と合わせると、50%超過～100%以下)と判断した。

【同じ命題の「肯定+推量表現」と「否定+推量表現」の両立】
(17) *太郎は来る**だろう**し、来ない**だろう**。
(18) *太郎は来る**はずだ**し、来ない**はずだ**。
(19) *太郎は来る**可能性が高い**し、来ない**可能性も高い**。
(20) 太郎は来る**可能性がある**し、来ない**可能性もある**。
(21) 太郎は来る**かもしれないし**、来ない**かもしれない**。
(22) *太郎は来る**可能性が低い**し、来ない**可能性も低い**。
(23) *太郎は勝つ**まい**し、また負ける**まい**。

条件表現の分類には、益岡(2006)および有田(2007)の分析と、先述の様相表現の分析を組み合わせている。まず、事実的条件表現は前件・後件ともに書き手にとって真であることが既知であるようなもので、これは益岡(2007)に挙げられている「現実(既然)の事態を表す条件文」(24)、および「現実の事態を所与の状況において非現実の事態として扱う「事態の非現実扱い」に現れるもの(25)を含むカテゴリである。これを筆者らは「真であることが既知の命題を導入する条件表現」として位置付けている。予測的条件表現、認識的条件表現、および反事実条件表現は有田(2007)の条件文の分類に従っているが、予測的条件表現と認識的条件表現はともに「真である確率が未知の命題を導入する条件表現」、反事実条件表現は「偽であることが既知の命題を導入する条件表現」と位置付けている。すなわち、予測的条件表現と認識的条件表現は、様相表現の下位カテゴリの認識的推量表現に対応している。

【事実的条件表現(真であることが既知)】
(24) さっき駅に行ったら、山田さんが友人を待っていた。
(25) あなたがそういう態度をとるなら、私にも考えがあります。(益岡 2007:211)
【予測的条件表現(真偽が未知、判断あり、確実性 100%)】
(26) 1 時間後に駅に集合したら、その足でいつもの居酒屋へ直行しよう。
【認識的条件表現(真偽が未知、判断あり、確実性 0~99)】
(27) もしもうまくいかなかったら、別の手段を考えよう。
(28) 万一太郎に知られていたら、大変なことになる。
【反事実条件表現(偽であることが既知)】
(29) 太郎が出場していたら、試合に勝てただろう。

認識的条件表現の下位分類は、先に八つのカテゴリに分類した認識的推量表現との共起を観察することで行っている。

(30) ホシは 10 分後にそこに現れるに違いない。(もし/#万一)やつが現れたらすぐ連絡しろ。
(31) ホシは 10 分後にそこに現れる可能性が高い。(ok もし/#万一)やつが現れたらすぐ連絡しろ。
(32) 確率は低いが、ホシがそこに現れる可能性がある。(ok もし/ok 万一)やつが現れたらすぐ連絡しろ。
(33) ホシはそこには現れるまい。しかし、(ok もし/ok 万一)やつが現れたらすぐ連絡しろ。

6. 結語

現在、ここで紹介した分類を元に設計したアノテーション仕様に基づき、ニュース記事からアノテーション済みコーパスを構築している。また、韓国語についても日本語との比較分析に基づいてスキーマを構築する予定である。更に、複数の様相表現の埋め込みによって起こる確実性の変化を「計算」するための論理体系を、可能世界意味論と公理論的確率論を組み合わせて定義することも予定している。

謝辞
本論文は科学研究費補助金(基盤研究(c)20500148「確実性アノテーション:『確実性判断を表す意味的文脈』を記述したコーパスの構築」(研究代表者:川添愛)平成 20 年度～22 年度)の助成を受けたものである。

参考文献

[Kilicoglu 2008] Kilicoglu, H, Bergler, S: "Recognizing speculative language in biomedical research articles: a linguistically motivated perspective," *BMC Bioinformatics*, 2008;9:S10, 2008.
[Light 2004] Light, M, Qiu, X, Srinivasan P : "The language of bioscience: facts, speculations, and statements in between," *Proceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, 2004.
[Medlock 2007] Medlock B, Briscoe T: "Weakly supervised learning for hedge classification in scientific literature," *Proceedings of 45th Meeting of the Association for Computational Linguistics 2007:992-999*, 2007.
[Palmer 2001] Palmer, F.R: *Mood and Modality second edition*, Cambridge University Press, 2001.
[Szarvas 2008] Szarvas G, Vincze V, Farkas R, Csirik J:"The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts," *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* 2008:38-45, 2008.
[有田 2007] 有田節子:『日本語条件文と時制節性』、くろしお出版、2007。
[加藤 2006] 加藤義清、黒橋禎夫、江本宏:「情報コンテンツの信頼性とその評価技術」、人工知能学会研究会資料、SIG-SWO-A602-01、2006。
[益岡 2007] 益岡隆志:『日本語モダリティ探究』、くろしお出版、2007。
[江口 2009] 江口萌、松吉俊、佐尾ちとせ、乾健太郎、松本裕治:「日本語文章の事象に対する判断情報アノテーション」、情報処理学会研究報告 2009-NL-193 No.5、pp.1-8、2009。
[田窪 2001] 田窪行則:「現代日本語における2種のモーダル助動詞類について」、『梅田博之教授古稀記念韓日語文学論叢』、太学社、2001。