言語情報の確実性アノテーションのための様相表現の分類

川添愛
（津田塾大学）
zoeai@tsuda.ac.jp

齊藤学
（中華大学）
haksa@live.jp

片岡喜代子
（九州大学）
kykk925@yahoo.co.jp

崔栄殊
（一橋大学大学院）
ld082003@g.hit-u.ac.jp

戸次大介
（お茶の水女子大学）
bekki@is.ocha.ac.jp

キーワード：確実性判断、様相表現、アノテーション

1. 目的

本論文では、言語情報処理、特に機械による言語情報の確実性認識への応用を目的とした推量表現の分類を紹介する。

自然言語で記述される情報には、以下に見られるように、事実だけでなく、筆者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」などの判断を行うことができる。機械が自動的に情報の確実性を判断できるようにしたい場合、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。
(2). 県内で新型インフルエンザが発生したようだ。
(3). 県内で新型インフルエンザが発生したとみられる。
(4). 県内で新型インフルエンザが発生したに違いない。
(5). 県内で新型インフルエンザが発生した可能性がある。
(6). 県内で新型インフルエンザが発生したのではないか。
(7). 県内で新型インフルエンザが発生したと疑われる。
(8). 県内で新型インフルエンザが発生した可能性は低い。

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、これを機械に学習させるために利用できる言語データ（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に着目し、それらの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文法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「文法的特性・意味的特性」を見極めるために「言語学的な裏付け」が必要であると考え、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た。

本論文では、この中でも様相表現、特に推量表現に焦点を当て、アノテーション仕様設計の一環として行った表現の分類と、それに伴う言語学的な議論を紹介する。

まず第 2 節では、筆者らが推進している確実性アノテーションプロジェクトの全体像について記述する。第 3 節では、当プロジェクトの重要な一部を占める推量表現の分類について述べる。

2. 確実性アノテーションプロジェクトの概要

2.1 確実性判断の必要性

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おり、短時間に情報の信憑性および緊急性を判断する必要がある。ところが単純なキーワード検索（たとえば「新型インフルエンザ　発生」のようなキーワードを利用した検索）では、先に挙げた(1)-(8)のような情報はすべて検索結果に残り、それらの間の確実性の差異は検知されないため、人間が一つ一つ見て判断するしかない。現行の自然言語処理技術を利用した情報抽出は、このようなキーワード検索よりもはるかに高度であるが、その多くはテキストの比較的表層的な特徴および用語の指示対象が属するクラスの記述（例えば「新型インフルエンザ」に対する「感染症」）などに基づいており、確実性に関わるような論理的な意味の認識は未だ課題となっている。しかし、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）の研究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。

2.2 本論文のアプローチ

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 XML などのマークアップ言語によっ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。そのようなタグ付けはアノテーションと呼ばれ、言語処理による情報抽出の研究においては、アノテーション済みのコーパスからの機械学習により、機械による言語知識の獲得や意味の認識を可能にすることを目指すというアプローチが広く用いられている。

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の設計が重要になる。アノテーション仕様の設計では、完成したアノテーション済みコーパスがどのようなタスクに使用されるか、またアノテーション対象のテキストがどのような種類のものか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、以下のように仕様の設計を行った。

1. 確実性の定義を行い、その定義に基づき命題のタイプ分けをし、最終的に得たい結果を確認する。
2. 確実性に影響する言語表現とそのスコープをアノテーションの対象に定める。
3. アノテーション対象の言語表現を確実性の度合いという点から下位分類し、１のタイプ分けと対応させる。

上に見られるように、筆者らのアプローチは、人間が情報の確実性判断の際に利用している知識のうち、言語表現についての知識をアノテーションの対象として記述しようというものである。無論、確実性判断に利用される知識は言語の知識

だけではないが、1) 口頭で伝えられる情報に比べ、書きことばによる情報伝達では、言語表現が担う役割が相対的に大きくなること、2) また言語表現に着目することで、さまざまな種類のテキストや用途に利用できるアノテーションが設計できる可能性が高いこと、3) 既に蓄積されている言語学的な知見や、言語学的な分析をアノテーションに利用できること等を考慮し、言語の知識を主な対象とすることにした。

この節の残りの部分では、上の 1、2 についての詳細を述べ、3 で述べられている言語表現の下位分類のうち、様相表現、特に推量表現の下位分類については第 3 節で重点的に述べる。

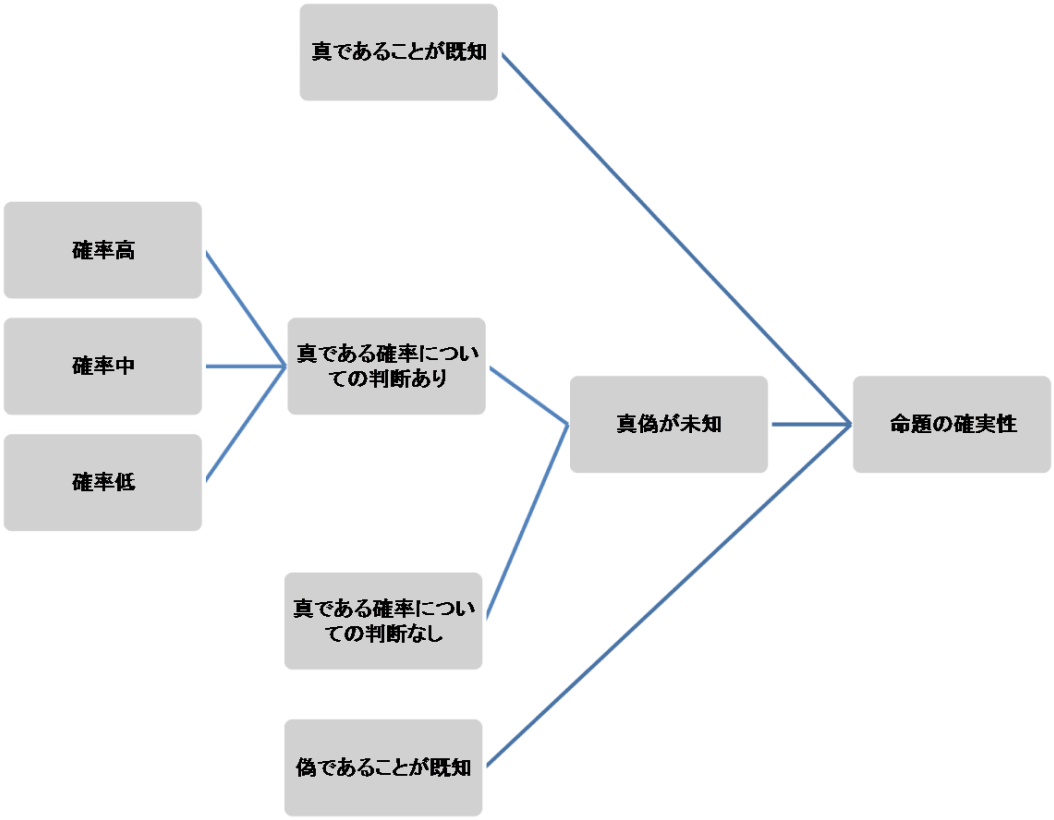
2.3 「確実性」に基づく命題の分類

筆者らのプロジェクトの最終的な目標は、テキスト中のあらゆる命題を、(9)の「確実性」を基準にした図に従って分類することである。ここで、「確実性」という言葉について定義を与えておく。本論文では、テキストの書き手が命題の内容を「真」と考えている度合いという意味でこの言葉を使う。したがって、完全に客観的な確実性とは異なる上、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。本論文で扱うのは、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う考えているかということである。また、ここでは「事実」という言葉は、誤解が生じない限りにおいて、上の表に見られる「テキストの書き手が『真であることが既知である』と考えている情報」を指して使う。

「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。

ただし、本研究では、各命題を直接(9)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（様相表現、否定表現、条件表現など）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。

(9). 命題の確実性の分類



2.4 言語表現とそのスコープに対するアノテーションの概要

本研究で行うアノテーションは、1) 確実性に影響を及ぼす表現に対するアノテーション、2)それらの表現が影響を及ぼす範囲（スコープ）に対するアノテーションの二種類に分けられる。

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 属性の三つである。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、書き手の考える「現在」を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。

(10). 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">疑われている</MODAL>。（秋田魁新聞 2006/06/18）

(11). <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">見通し</MODAL>。（読売新聞 2009/5/15）

3. 様相表現の分類

先に述べたように、本研究では、アノテーション対象の言語表現を確実性という点から下位分類し、(9)の命題の分類と対応させる。ここでは、様相表現の分類について述べる。

3.1 様相表現の分類と命題の確実性との対応

様相および様相を表す表現の分類は、過去の研究で多く試みられている。例えばPalmer (2001)は、様相（modality）を、命題の真偽値あるいは事実性に対する話者の態度に関わるPropositional modality と、潜在的・未実現の出来事に言及するEvent modality の二種に分けている。更に、Propositional modality は命題の事実性に関する判断を表現するEpistemic modality と話者の持つ命題の事実性に関わる証拠を示すEvidential modality に分けられている。Evidential modalityは証拠の種類によってさらに分類される。Willett(1988; 96)においても、Direct Evidence とIndirect Evidenceの区別がある。

(12). Palmerによるモダリティの分類

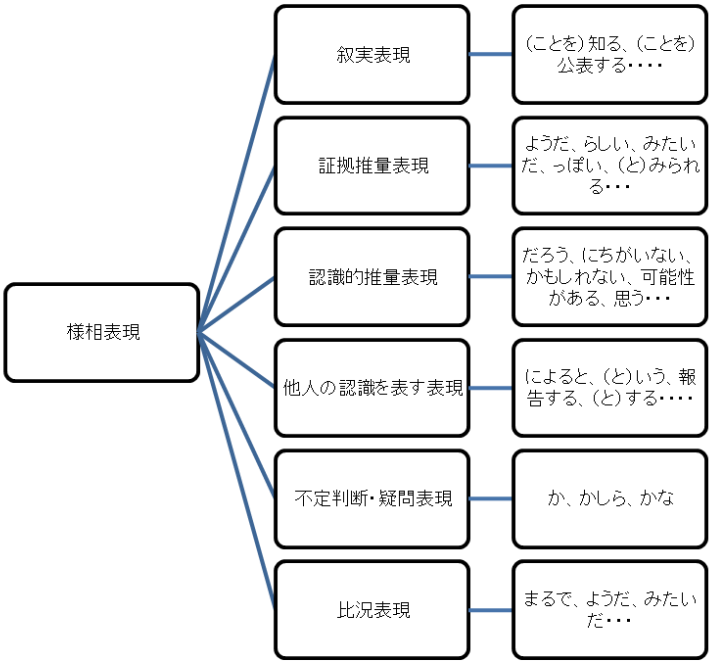
- Propositional modality
  - Epistemic
    - Speculative
    - Deductive
    - Assumptive
  - Evidential
    - Reported: Reported(2), Reported(3), Reported(Gen)
    - Sensory: Visual, non-Visual, Auditory
- Event modality
  - Deontic
    - Permissive
    - Obligative
    - Commissive
  - Dynamic
    - Abilitive
    - Volitive

（Palmer (2001,p.22)）

本研究では、命題の「確実性」の度合いに基づき命題の分類を行うため、上述の Event modality は分類の対象に含めず、Propositional modality に属する表現のみが分類の対象となる。また、先行研究のモダリティの分類では、真なる命題、もしくは偽なる命題を前提として持つ表現は対象外とされていたが、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、最も確実性が高い、もしくは最も確実性が低い表現に属させることが可能であることから、本研究ではこれらの表現も分類の対象に含めることとする。

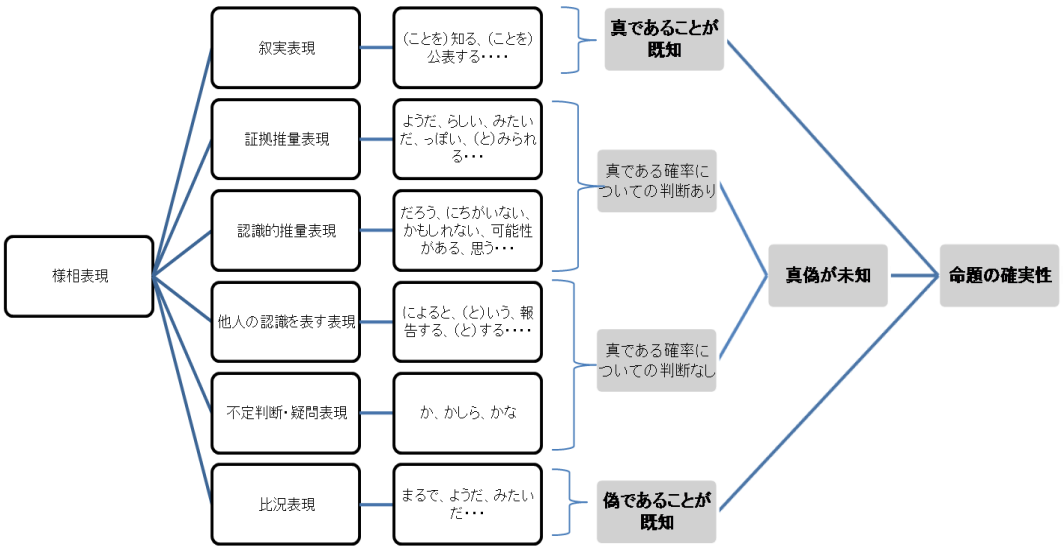
本研究では、様相表現を(13)のように分類する。

(13).　様相表現の分類



上の様相表現の分類は、確実性による命題の分類(9)と対応するように、各表現が導入する命題の確実性に基づいた分類である。対応は(14)の図のとおりである。

(14).　様相表現の分類と命題の分類との対応



証拠推量表現・認識的推量表現以外の表現について、簡単に述べておく。叙実表現には、「知る」のような動詞、「事実」「現実」などのような名詞、「さいわい」「あいにく」などの副詞（価値・評価の副詞、有田(2007：60)で文の命題内容が真であることを前提とする副詞とされているもの)、また「わけだ」「ではないか」などの表現が含まれる。他人の報告や認識を表す表現、および不定判断・疑問表現は、真偽が未知であり、かつ書き手が真である確率についての判断をしていない命題のカテゴリに対応する表現として位置付けている。他人の報告する事柄を表す表現としては、「(〜に）よると」のような表現や、「(〜と）いう」、「述べる」のような動詞表現、「報告」のような名詞をアノテーション対象とした。また、命題の真偽に関する他人の判断を表すため、書き手以外のものを主語に取る「確信している」「思っている」「思っていない」などの表現を含めた。「不定判断・疑問表現」は、益岡(2007 : 144)の真偽判断のモダリティの分類の中にある「不定判断」に由来するもので、ここでは疑問の「か」や「かどうか」、疑問符の「？」などをアノテーション対象としている。比況表現としては、比況の「ようだ」や「まるで」が含まれる。「ようだ」のように、同じ表現が複数のカテゴリにまたがる場合があるが、アノテーション仕様において各カテゴリの特徴や曖昧性解消のためのテストを詳細に述べることによって対処している。

3.2 推量表現の分類

3.2.1 証拠推量表現と認識的推量表現の区別

本論文で先に提案した(13)の分類においては、証拠推量表現と認識的推量表現は、いずれも「書き手にとって真偽が未知であるが、真である確率についての判断がある」ような命題を導入する表現として特徴づけている。証拠推量表現とは、「話し手（書き手）が、推量の根拠の存在を示すもの」(Palmer(2001)) 「ある具体的な証拠から推論によって得た知識を述べるための形式」（益岡・田窪(1992;128)）などのように特徴づけられている表現のカテゴリである。日本語においては「らしい」「ようだ」「みたいだ」のような表現（以下、「ヨウダ類」）がこれに属するとされる（寺村(1984)、Aoki(1986)、森本(1994)等）。他方、ここでいう認識的推量表現とは、「だろう」「かもしれない」「可能性がある」「はずだ」のように、「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)）を導入する表現を指している。

ここでは確実性の観点から、証拠推量表現（ヨウダ類）と認識的推量表現（ダロウ類）の区別を重視する。ダロウ類が使われた文とヨウダ類が使われた文で、どちらが確実性が高いかは一概には決められない。しかし、1) ダロウ類が事実以

外に仮想世界に対する推論を含むのに対し、ヨウダ類は事実に対する推論に限定されること、2) ダロウ類には必ずしも推量の根拠は存在しないが、ヨウダ類には必ず存在することから、より確実な情報を探す際にヨウダ類に優先順位をつけるということは、理にかなっているように思われる。

ダロウ類とヨウダ類の区別には、田窪(2001)において紹介されている「今ごろ」等を使った反実仮想文脈が利用できる。このテストは、ダロウ類が仮想世界に対する推論を含み、ヨウダ類が現実世界に対する推論に限定されることに基づいている。

【反実仮想テスト】

(15). *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｝。

(16). 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｛だろう・かもしれない・可能性がある・はずだ｝。

また、田窪(2001)では、推量の証拠の存在を示唆する「どうやら」「どうも」といった副詞を用いたテストも紹介されている。ヨウダ類はこれらと共起できるが、ダロウ類は共起できない。

【「どうやら」「どうも」テスト】

(17). ｛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｝。

(18). *｛どうやら・どうも｝彼はうそをついている｛だろう・かもしれない・可能性がある・はずだ｝。

また上のテストと関連して、「特にそう結論づける理由はないけれど」のような節と共起できるかというテストも、これら二つのカテゴリを区別する目的で使えるように思われる。

(19). 特にそう結論づける理由はないけれど、太郎は来ない｛だろう・はずだ・可能性がある・かもしれない｝。

(20). ??特にそう結論づける理由はないけれど、太郎は来ない｛ようだ・らしい・みたいだ｝。

本論文での証拠推量表現と認識的推量表現の分類は、「仮想世界に対する推論を含むか否か」「推論の証拠が存在するか否か」を示す以上のテストに従っている。

3.2.2　証拠推量表現

証拠推量については、Palmer (2001)が、証拠が視覚的か、聴覚的か、また他人の報告によるものかによって分類するという立場をとっている。ここでは、証拠の種類については区別をせず、いずれの場合においても、その証拠に基づいて書き手が自身で推量をしている場合は、証拠推量としてアノテーションを行う。ただし、単に他人の報告を記述しているのにとどまり、書き手自身が推量を行っていない場合は、前述の「他人の認識を表す表現」としてアノテーションする。

以下、アノテーション対象の証拠推量表現について詳細を述べる。

3.2.2.1　ようだ、みたいだ、らしい、っぽい、そうだ

「ようだ」「みたいだ」については、証拠推量の意味以外に、次のような「比況」（他のものにたとえる）、「婉曲」（遠回りに表現する）などの意味がある。比況の「ようだ」「まるで」は前述の通り比況表現としてアノテーションし、婉曲はアノテーション対象としない。

(21).　まるで盆と正月が一度に来た**ようだ**（**みたいだ**）。（比況）
(22).　お車の準備ができた**ようです**。（婉曲）

「らしい」「ようだ」「みたいだ」については、証拠が視覚や聴覚などを介した直接体験によるもの（例：(23)）、他人の報告に基づく間接的なもの（例：(24)）の場合がある。ただし、これらの表現については、証拠の種類に関係なく、書き手自身が推量をしていると判断できるため、いずれの場合も証拠推量としてアノテーションする。

(23).　（台所から漂う匂いから判断して）また今日もカレー**らしい**。
(24).　天気予報によると、明日は雨が降る**らしい**。

3.2.2.2　見込み、見通し、模様

「見込み」「見通し」「模様」を証拠推量表現（ヨウダ類）とする根拠は、3.2.1で紹介したテストにおいて、ヨウダ類と同じふるまいをすることである。

【「どうやら」テスト】
(25).　どうやら明日にも退院の許可が降りる**見込みだ**。
(26).　どうやら明日にも退院の許可が降りる**見通しだ**。

(27). どうやら明日にも退院の許可が降りる**模様だ**。

【反実仮想テスト】

(28). *もっと早く医者にかかっていたら、今頃はとっくに元気になっている**見込みだ**。

(29). *もっと早く医者にかかっていたら、今頃はとっくに元気になっている**見通しだ**。

(30). *もっと早く医者にかかっていたら、今頃はとっくに元気になっている**模様だ**。

ただし、「見込みがない」については、「可能性がない」などと同様、認識的推量表現の epistemic_0 類としてアノテーションを行う。この表現を認識的推量表現とする理由は、次のような反実仮想の文の容認性がそれほど低くなく、「可能性はない」と大きく変わらないからである。

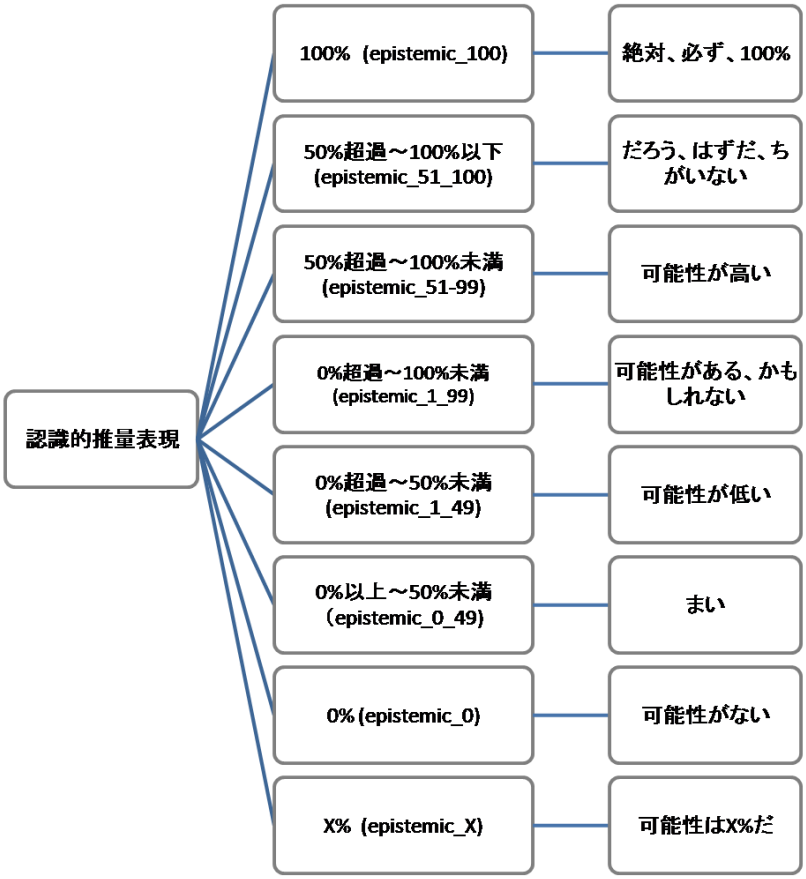
(31). あの時漁船が近くを通りかからなかったら、今頃彼が生きている**見込みはない**。

(32). あの時漁船が近くを通りかからなかったら、今頃彼が生きている**可能性はない**。

3.2.3 認識的推量表現

認識的推量表現のカテゴリには、書き手が命題の確実度を「高い」と判断している場合に使われるもの（「絶対」「必ず」「はずだ」「違いない」「可能性が高い」など）や、逆に命題の確実度が「低い」と考えている場合に使われるもの（「可能性が低い」「まい」「可能性がない」）、また可能性があることを示唆するのみで確実性の度合いに特に言及しない場合に使えるもの（「可能性がある」「かもしれない」「（と）思う」などが含まれる。ここでは、認識的推量表現を、それが導入する命題が真である確率に応じて(33)のように下位分類する。

(33). 認識的推量表現の下位分類



上に見られるように、この分類は、各カテゴリの表現について特定の数値を割り当てるものではなく、とりうる数値の「幅」を割り当てるものである。この下位分類の根拠は以下の通りである。まず、「絶対」「必ず」「100%」のような副詞的表現を、「書き手にとって命題の真偽は未知だが、真である確率が 100%であると考えているときに使う表現」（以下、epistemic_100）と考える。

(34). **必ず/絶対/100%**、太郎は来る。

次に、「真である確率が 100%の場合を含むかどうか」を判定するためのテストを考える。以下の(35)-(40)を見ると、「だろう/であろう」「ちがいない」「はずだ」は epistemic_100 の表現の「必ず/絶対/100%」と共起し、他方「可能性が高い」「可能性がある」「かもしれない」等は共起しない。

【「必ず/絶対/100%」との共起】

(35). 必ず/絶対/100% 来る**であろう**人に招待状を出す必要はありません。
(36). 必ず/絶対/100% 来る**にちがいない**人に招待状を出す必要はありません。
(37). 必ず/絶対/100% 来る**はずの**人に招待状を出す必要はありません。
(38). *必ず/絶対/100% 来る**可能性が高い**人に招待状を出す必要はありません。
(39). *必ず/絶対/100% 来る**可能性がある**人に招待状を出す必要はありません。
(40). *必ず/絶対/100% 来る**かもしれない**人に招待状を出す必要はありません。

(35)-(40)において、様相表現をすべて関係節に埋め込んでいる理由は、主節にしかないような空の要素（例えば文の断定に関わるようなもの）を「必ず/絶対/100%」等が修飾する可能性を排除するためである。実際、関係節への埋め込みを伴わない(41)-(46)においては、特に(45)と(46)において話者による判断のゆれが見られる。

(41). 必ず/絶対/100%、太郎は来る**だろう**。
(42). 必ず/絶対/100%、太郎は来る**にちがいない**。
(43). 必ず/絶対/100%、太郎は来る**はずだ**。
(44). *必ず/絶対/100%、太郎は来る**可能性が高い**。
(45). ?*必ず/絶対/100%、太郎は来る**可能性がある**。
(46). ?*必ず/絶対/100%、太郎は来る**かもしれない**。

「必ず/絶対/100%」との共起関係から、「だろう/であろう」「ちがいない」「はずだ」とその他の表現の間に非対称性があることがわかる。「だろう/であろう」「ちがいない」「はずだ」が 100%を含むかどうかは、更に「予言」の文脈で自然に使用できるかというテストで確かめることができる。通常、予言というものは書き手がその内容を 100%確信した上でなされる発言である。以下に見られるように、「だろう」「ちがいない」「はずだ」は予言の文脈で自然に使用できるが、「可能性が高い」「可能性がある」「かもしれない」はそれが使用された時点で予言とは言えなくなり、不自然である。

【予言の文脈】

(47). 私は予言する。19XX 年に大地震と大津波が起こる**だろう**。
(48). 私は予言する。19XX 年に大地震と大津波が起こる**にちがいない**。
(49). 私は予言する。19XX 年に大地震と大津波が起こる**はずだ**。
(50). 私は予言する。#19XX 年に大地震と大津波が起こる**可能性が高い**。（予言

でなくなる）

(51). 私は予言する。#19XX 年に大地震と大津波が起こる**可能性がある**。（予言でなくなる）

(52). 私は予言する。#19XX 年に大地震と大津波が起こる**かもしれない**。（予言でなくなる）

このことから、「だろう/であろう」「ちがいない」「はずだ」は 100%を含む（つまり、これらが導入する命題が真である確率は 100%以下）という結論が導ける。他方、「可能性が高い」「可能性がある」「かもしれない」等は、100%を含まない（つまり、これらが導入する命題が真である確率は 100%未満）と考えることができる。

次に、各表現の導入する命題の確実性が 50%を超えているかどうかについて見る。これを判定するためのテストは、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である。もし表現の導入する命題の確実性が必ず 50%を超過しているなら、これは不可能になると考えられる。というのは、書き手が、ある命題が真である確率を 50%より高いと考えながら、同時にその命題が偽である確率も 50%より高いと考えることは不可能だからである。以下を見ると、命題に付く表現が「だろう」「はずだ」「ちがいない」「可能性が高い」である場合は、同じ命題の肯定と否定を同時に主張することができない。他方、「可能性がある」「かもしれない」の場合は、肯定と否定を同時に主張することが可能である。

【同じ命題の「肯定+認識的推量表現」と「否定+認識的推量表現」の両立】

(53). *太郎は来る**だろう**し、来ない**だろう**。

(54). *太郎は来る**はずだ**し、来ない**はずだ**。

(55). *太郎は来るに**ちがいない**し、来ないに**ちがいない**。

(56). *太郎は来る**可能性が高い**し、来ない**可能性も高い**。

(57). 太郎は来る**可能性がある**し、来ない**可能性もある**。

(58). 太郎は来る**かもしれないし**、来ない**かもしれない**。

(53)-(56)の容認性の低さより、「だろう」等の表現群が導入する命題の確実性が必ず「50%を超過する」と考えられる。このテストの結果と、先に見た「100%を含むかどうか」を調べるテストの結果から、「だろう」「はずだ」「ちがいない」が表す確実度の範囲は「50%超過～100%以下」（以下、epistemic_51_100）、「可能性が高い」は「50%超過～100%未満」（以下、epistemic_51_99）と考えることができる。

これに対し(57)(58)が容認可能であることは、「可能性がある」「かもしれない」が確実性が 50%以下の命題を取ることも、50%以上の命題を取ることもできるような、幅広い確実度を許容する表現であると考えると説明できる。

(53)-(58)のテストは、認識的推量表現が導入する命題の確実度が「50%超過かどうか」を判定するのみならず、「50%未満かどうか」を判定するのにも使うことができる。

(59). *太郎は来る**可能性が低い**し、来ない**可能性も低い**。
(60). *太郎は勝つ**まい**し、また負ける**まい**。

最後に、「確実性が 0%の場合」を含むかどうかについてみる。ここでは、「確率は 0%だ」および「可能性はない」を、「書き手にとって命題の真偽は未知だが、偽である確率が 100%であると考えているときに使う表現」（以下、epistemic_0）と考え、これらを含む文に適切に後続できるかどうかで、0%を含むかどうか判定する。

【「可能性はない」「確率は 0%だ」との共起】
(61). 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**。[1]
(62). 太郎が来る｛確率は 0%だ・可能性はない｝。#太郎が来**る可能性はある**。
(63). 太郎が来る｛確率は 0%だ・可能性はない｝。#太郎が来る**可能性は低い**。
(64). 太郎が来る｛確率は 0%だ・可能性はない｝。太郎は来る**まい**。

(61)-(64)を見ると、「まい」以外は不適切である。よって、「まい」は0%を含む（つまり0%以上である）が、他の表現は0%を含まない（0%超過である）と結論付けられる。他のテストの結果と組み合わせて、「かもしれない」「可能性がある」は「0%超過～100%未満」（epistemic_1_99）、「可能性は低い」は「0%超過～50%未満」（epistemic_1_49）、「まい」は「0%以上～50%未満」（epistemic_0_49）と分類することができる。

以上のテストの結果に従って、アノテーション対象となる表現を分類した表が(65)である。各分類に対する判断基準は「備考」にまとめている。

[1] 文脈によっては、「勝てる確率は 0%だ。しかし、勝つかもしれない」という文が適切に発話できる場合があるが、この場合は、最初の文と次の文とで、推量を行っている主体が異なると考えられる（たとえば前者はコンピュータや一般の人間による予測、後者は話者の個人的な予測）。

(65).

表現のカテゴリ	（書き手にとって）命題が真である確率	表現の例	備考
epistemic_100	100%	絶対、100%、必ず、絶対に、間違いなく、確実に	
epistemic_51_100	50%超過～100%以下	だろう（感嘆の「だろう」以外）、であろう、ろう、でしょう、（～に）違いない、（～ことは）間違いない、（～ことは）疑いない、はずだ、はず	【判断基準】 1. epistemic_100の表現と共起すること 2. 同じ命題の「肯定+推量表現」と「否定+推量表現」の両立が不可能であること
epistemic_51_99	50%超過～100%未満	可能性が高い、おそれが強い、疑いが強い、（書き手が～と）確信する（書き手が～と）、信じる、（書き手が～と）予測する、（書き手が～と）考える、（書き手が～と）予想する、（と）する、かもしれない、おそらく、多分、きっと、	【判断基準】 1. epistemic_100の表現と共起しないこと 2. 同じ命題の「肯定+推量表現」と「否定+推量表現」の両立が不可能であること 【その他】 「十中八九」は、epistemic_80_99と考える。
epistemic_1_99	0%超過～100%未満	かもしれない、かも、かもわからない、可能性がある、おそれがある、疑いがある、可能性、おそれ、疑い、のではないか、のではないだろうか、（書き手が～と）思う、（書き手が～と）疑う、ありうる、保証はない、確信はない、確証はない	【判断基準】 1. epistemic_100の表現ともepistemic_0の表現とも共起しないこと 2. 同じ命題の「肯定+推量表現」と「否定+推量表現」の両立が可能であること 【その他】 「可能性」「おそれ」「疑い」については、ニュースの見出し等で、名詞止めで現れるものを対象とする。

epistemic_1_49	0%超過～50%未満	可能性は低い、おそれは低い、可能性はあまりない	【判断基準】 1. epistemic_0の表現と共起しないこと 2. 同じ命題の「肯定+推量表現」と「否定+推量表現」の両立が不可能であること
epistemic_0_49	0%以上～50%未満	まい、（書き手が～と）思わない、（書き手が～と）思えない、（書き手が～と）考えない、（書き手が～と）考えられない、（書き手が～と）信じない、（書き手が～と）信じられない	【判断基準】 1. epistemic_0の表現と共起すること 2. 同じ命題の「肯定+推量表現」と「否定+推量表現」の両立が不可能であること 【その他】 「考えられない」「信じられない」には「現実に起こっていることが信じられない」という気持ちを表すために使われることがあるが、この場合は従属節の内容が真であることが既知である。よってこのように使われる場合は推量表現とは考えない。
epistemic_0	0%	可能性はない、おそれはない、疑いはない、あり得ない	
epistemic_X	X%（Xに入る数字に依存）	可能性はX%（だ）、確率はX%（だ）	

4. 結語

本論文では、確実性という観点による様相表現、特に推量表現の分類について論じた。この分類が具体的な言語処理タスクにおいてどの程度有効であるかについては、今後複数のアノテータ間におけるアノテーションの一致や、機械学習の結果などによって示す必要がある。ただし、アノテーション仕様のデザインにおいて比較的明確かつ経験的な根拠があり、その一部を言語学的なテストという形でアノテータに示すことができるという点は、アノテーションの一貫性を高める上で有益であると思われる。従来の言語処理の研究においては、言語学的な考察

が「役に立つ」機会はあまりなかったように思われるが、本論文で取り上げた機械による確実性判断のように、「意味」や「文脈」に関わるタスクに対するニーズが大きくなるにつれ、言語学的な知見の利用価値が高まるのではないかと考えられる。

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

本論文に貴重なコメントを下さった二名の査読者に感謝を申し上げる。また、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

Aoki, H. (1986) “Evidentials in Japanese.,” Chafe, W. & Nichols, J.(eds) *Evidentiality. Ablex Publishing Corporation*:223-237.

Kilicoglu, H. and Bergler, S. (2008) “Recognizing speculative language in biomedical research articles: a linguistically motivated perspective.,” *BMC Bioinformatics.* 2008;9:S10.

Light, M., Qiu, X. and Srinivasan, P. (2004) “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*, Boston, May 2004.

Medlock, B. and Briscoe, T. (2007) “Weakly supervised learning for hedge classification in scientific literature,” *Proceedings of 45th Meeting of the Association for Computational Linguistics 2007*:992-999.

Palmer, F.R. (2001) *Mood and Modality second edition.* Cambridge University Press.

Szarvas, G., Vincze, V., Farkas, R. and Csirik, J. (2008) “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.

Willett, T. (1988) “A Cross-Linguistic Survey of the Grammaticization of Evidentiality,” *Studies in Language* 12:51–97.
有田節子 (2007) 『日本語条件文と時制節性』, 東京：くろしお出版.
江口萌・松吉俊・佐尾ちとせ・乾健太郎・松本裕治 (2009) 「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.5:1-8.
加藤義清・黒橋禎夫・江本宏 (2006) 「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.
寺村秀夫 (1984) 『日本語のシンタクスと意味Ⅱ』, 東京：くろしお出版.
益岡隆志 (2007) 『日本語モダリティ探究』, 東京：くろしお出版.
益岡隆志・田窪行則 (1992) 『基礎日本語文法』, 東京：くろしお出版.
森本順子 (1994) 『話し手の主観を表す副詞について』, 東京：くろしお出版.
田窪行則 (2001) 「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』:1003-1025, ソウル：太学社.

Classification of modal expressions for certainty annotation of texts

Ai Kawazoe (Tsuda College)
Manabu Saito (Chung Hua University)
Kiyoko Kataoka (Kyushu University)
Young Soo Choi (Hitotsubashi University)
Daisuke Bekki (Ochanomizu University)

Natural language texts contain pieces of information with several levels of certainty, namely, factual assertions, speculations, inferences and hypothetical thoughts. Recognition of certainty levels in textual information is crucial for efficient extraction of newly reported facts and rapid judgment of the reliability of the information. Some groups in natural language processing (e.g., Light et al. (2004), Medlock and Briscoe (2007), Szarvas et al. (2008), Kilicoglu and Bergler (2008), Eguchi et al. (2009)) have started to develop technologies for automatic certainty recognition. We are now constructing a corpus in which modal, negative, and conditional expressions and their scopes are semantically annotated, as a basis for dealing with certainty of information in Japanese texts. In order to achieve high-quality annotation, we have designed an annotation schema on the basis of the classification of key expressions according to the certainty levels of propositions they introduce. We classify modal expressions into 6 categories including factive, evidential and epistemic expressions, and sub-classify epistemic expressions into 8 categories based on the following observations: 1) co-occurrence with "100%" expressions (e.g., 'kanarazu', 'zettai'), 2) occurrence in a conjoined sentence where the same proposition is both affirmed and negated, and 3) co-occurrence with "0%" expressions (e.g., 'kanoosei-wa nai'). These empirical observations are incorporated into the annotation schema and can be used by human annotators to ensure consistent annotation.

(初稿受理日　2010 年 2 月 26 日　　最終稿受理日　2010 年 7 月 27 日)