様相・条件・否定表現の言語学的分析に基づく確実性判断のためのアノテーション済みコーパスの構築

川添　愛[*1]　齊藤　学[*2]　片岡　喜代子[*3]　崔　栄殊[*4]　戸次　大介[*5]

[*1] 津田塾大学　[*2] 中華大学　[*3] 九州大学　[*4] 一橋大学大学院　[*5] お茶の水女子大学

1. 目的

自然言語のテキストの中で記述される命題は、常に事実(書き手にとって真である情報)とは限らない。(1)のような事実以外にも、推量や伝聞などのように書き手にとって真偽が明らかでない情報(2)(3)や、比況や否定、反事実条件の中に現れる命題(4)(5)(6)のように、偽であることが明らかである情報がある。

(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) 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。(反事実条件)

人間にとってごく普通にできる上のような言語情報の確実性判断が、機械によって自動的にできるようになれば、様々な応用が考えられる。たとえば、感染症情報など緊急の判断が必要とされる情報の取得の際に、不要な情報を取り除いて効率的に情報を取り出すことや、情報の確実性・信憑性判断に関わる人的コストを減らせることなどが考えられる。

筆者らは現在、自然言語のテキストに現れる事実とそれ以外の情報との区別、また推量や仮定などの間に見られる確実性の差を自動的に識別するための基盤として、様相表現、条件表現、否定表現とそれらのスコープをアノテーションしたコーパスを構築している。コーパスに実用性と言語学的な裏付けの両方を持たせるため、様相表現・否定表現・条件表現を「確実性」の側面から分析・分類し、それに従ってアノテーションスキーマを作成した。本論文では、本研究の成果物であるアノテーションガイドラインの概要とコーパスの現状を述べる。

2. 先行研究

言語情報の確実性判断を目指した言語処理研究は、主に英語では医学・生物学テキストを対象に、推測や意見を事実の記述から区別するタスク(hedge classification)がある。Lignt et al. (2004)では、MEDLINEアブストラクト内の推測を表す文を人手によってアノテーションしたコーパスを構築している。様相などの表現に対するアノテーションは行っていないが、suggest, potential, likely, mayなどの14の表現をキーワードとして用いた実験の結果が、SVMによる学習結果と同等のパフォーマンスを示したと報告している。Medlock and Briscoe (2007)でもアノテーションガイドラインを構築して人手によるコーパスを作っているが、推測を表す文に特徴的な表現を認識することに重きを置いており、スコープはその表現を含む文全体と見なしている。またSzarvas et al.(2008)では、言語学者が生物学テキストに否定表現、様相表現とそのスコープをアノテーションしたBioScopeコーパスを作成している。Kilicoglu and Bergler (2008)は、言語学の成果を利用して推測を表す表現の辞書を構築し、更に各表現に重み付けをして推測の度合いの強さの計算を行っている。実験にはMedlock and BriscoeのコーパスとBioScopeを利用している。また、GENIA event corpus (Ohta et al.2007)では、生物学テキスト中の事象に対して、certain, probable, uncertainの三段階の情報をアノテーションしている。日本語では松吉ら(2010)による、テキスト中で言及される事象に対して真偽判断等の情報を付与する判断情報アノテーションの研究がある。

3. アノテーションガイドライン

上記の通り、これまでの研究で構築されたコーパスには大きく分けて1)不確実性を含む文全体をアノテーションしたもの(Light et al. 2004, Medlock and Briscoe 2007)、2)確実性に関わる表現とそのスコープをアノテーションしたもの(BioScope)事象を表す表現そのものにアノテーションしたもの(GENIA event corpus、松吉ら(2010))がある。本研究で構築するコーパスは「確実性に関わる表現とそのスコープをアノテーションしたもの」である点でSzarvas et al. (2008)の方針と同じであるが、日本語のテキストを対象とし、なおかつより一貫した信頼度の高いアノテーションを目指す上で、いくつかの点に留意している。

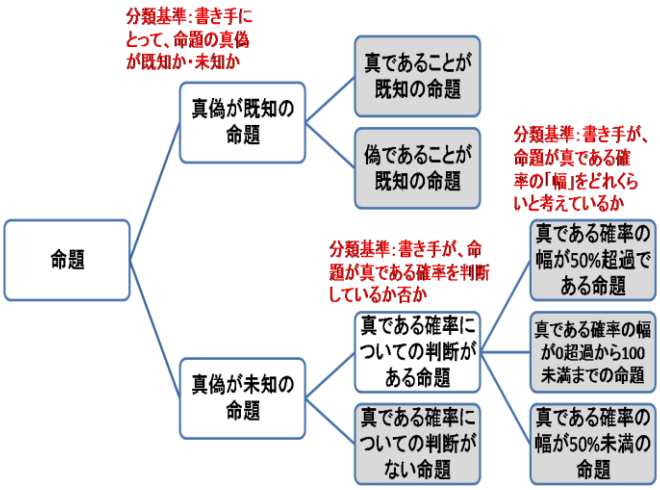
3.1 確実性に基づく命題の分類

留意点の一つは、「確実性」という用語と、それに基づいた命題の分類についてである。本論文では「確実性」という言葉を、「テキストの書き手が判断する、命題の内容が真である確率」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「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



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、分類基準を明確にする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次節で述べる。

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(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。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。

3.2 言語学的な知見の再編集、確実性による表現の分類

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。そのためには、アノテータが「なぜこんなアノテーションをしたのかがわからない」ような状況や、その理由を他人と共有できない状況を避ける必要がある。

しかしながら、命題の確実性に関わる判断は複雑である。日本語の場合は、手がかりとなる様相表現・否定表現・条件表現には多義性のあるものが多い(証拠推量、比況、婉曲の「ようだ」など)。またそれらの表現間の埋め込みも多く見られ、埋め込まれた命題の確実性に対して影響を与えることから、アノテーションスキーマの記述を詳細にするだけでは対処しきれないのではないかという懸念がある。理想としては、言語的な手がかりによる確実性の判断の際に、依って立つことのできる理論的分析があり、かつ個々の使用例に対してその分析を適用できることが望ましい。

しかし、「そのまま使える言語学的な分析」が必ずしも存在するわけではない。様相・条件・否定表現に関しては理論言語学や日本語学で盛んに研究されているが、研究の対象が特定の表現・用法に集中している場合が多く、確実性判断という実用的な用途に利用でき、なおかつテキスト中に現れるほぼすべての表現を網羅した分析を見つけるのは難しい。

よって、これまでになされた分析を言語学的な立場と応用の立場から再編集したり、足りない部分を補足したりする必要がある。このプロジェクトでは、言語学的な分析と、アノテーションガイドラインの作成を同時進行で行うというアプローチを採用している。

本研究では、様相表現、否定表現、条件表現を表のように分類した。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、「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。

表現の上位の分類の多くは、既存の言語学的分析に従っている。たとえば、様相表現中の証拠推量表現と認識的推量表現の区別は Palmer(2001)に従い、田窪(2001)で紹介されている「今ごろ」を使ったテスト等を利用している。

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ている。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し、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%)の場合を含まないものと判断した。

(8) 表現の分類

様相表現	ラベル	表現の例	命題の分類との対応	表現の数（2011年1月現在）
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	44
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	13
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	68
	epistemic_51_100（確実性50%超過～100%以下）	だろう、（～に）違いない、はずだ		16
	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、		31
	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	61
	epistemic 1_49（確実性0 »超過～50%未満）	可能性は低い、おそれは低い、（書き手が～と）思わない	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	21
	epistemic_0_49 （確実性0 »以上～50%未満）	まい、		1
	epistemic 0（確実性 0%）	可能性はない、おそれはない、		15
	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	3
他人の認識を表す表現	hearsay, other_epistemic_100, other_epistemic_51_100,...	（に）よると、（と）いう、（と）する、（書き手以外の者が）～と思う	真であることが未知―真である確率についての判断なし	108
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	12
比況表現	simile	まるで、ようだ、みたいだ	偽であることが既知	3
反叙実表現	anti-factive	騙る、勘違いする	偽であることが既知	14

否定表現	ラベル	表現の例	命題のタイプとの対応	表現の数（2011年1月現在）
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	5
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	8

条件表現	ラベル	表現の例	命題のタイプとの対応	表現の数（2011年1月現在）
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	5
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	2
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、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	17
	cond_epistemic_0_49（確実性 0 »以上～100%未満）	たら、なら（ば）、れば、時、場合、とする、仮定する（「仮に」と共起する場合）	（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	17
一般的条件表現	generic	たら、なら（ば）、れば	（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なし	13
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	11

ある表現の導入する命題の確実性が 50%の場合を含むかどう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。

条件表現の分類には、益岡(2007)および有田(2007)の分析と、認識的推量表現の分類で利用した言語テスト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」等を含むカテゴリである。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。認識的条件表現の下位分類には、先の認識的推量表現の分類手法を応用している。

本研究で提案した分類手法について、詳しくは川添・齊藤・片岡・崔・戸次(2010)を参照されたい。

3.3 アノテーション概観

表現に対するアノテーションのためのクラスは MODAL, NEG, COND である。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 3 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示す。

(9) 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)

(10) <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)

4. コーパスの現状

アノテーションガイドラインに従い、2011 年 1 月現在、500 のニュース記事(4675 文)に対してアノテーション済みである。表現の出現数の内訳は、様相表現 2667、条件表現 220、否定表現 24 である。

5. 結語

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。現在、本論文で紹介した手法を利用して、韓国語の様相表現の分類を行っている。今後は、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構築する予定である。

[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, H, Bergler, S: "Recognizing speculative language in biomedical research articles: a linguistically motivated perspective," BMC Bioinformatics, 2008;9:S10, 2008.

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 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.

Ohta, T., Kim, J.D., Tsujii, J. *Guidelines for Event Annotation.* (http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/release/GENIA_event_annotation_guidelines.pdf)

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

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.

有田節子:『日本語条件文と時制節性』、くろしお出版、2007。

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。

川添愛、齊藤学、崔栄殊、片岡喜代子、戸次大介「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」、『九州大学言語学論集 31』, pp.109-129、2010。

益岡隆志:『日本語モダリティ探究』、くろしお出版、2007。

松吉 俊, 江口 萌, 佐尾 ちとせ, 村上 浩司, 乾 健太郎, 松本 裕治「テキスト情報分析のための判断情報アノテーション」電子情報通信学会論文誌 D, Vol.J93-D, No.6, pp.705-713, 2010。

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。