社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93（12）
2009－ＮＬ－189（12）
2009／1／23

確実性判断に関わる意味的文脈アノテーションの試み

川添愛[*1]　齊藤学[*2]　片岡喜代子[*3]　戸次大介[*4]
津田塾大学[*1]　中華大学[*2]　国立国語研究所[*3]　お茶の水女子大学[*4]

要旨
自然言語のテキストには、事実だけでなく、推測など不確実な情報も含まれる。このような情報の確実性の識別の手がかりとなるのが、様相（文の内容に対する書き手の認識・判断）を表す表現や、否定表現などによって形成されるある種の「意味的文脈」である。本発表では、そのような意味的文脈をアノテーションした日本語コーパスを構築する試みについて述べる。

A Preliminary Study for Constructing Japanese Corpus Annotated for Certainty and Uncertainty

Ai Kawazoe[*1]　Manabu Saito[*2]　Kiyoko Kataoka[*3]　Daisuke Bekki[*4]
Tsuda College[*1]　Chung Hua University[*2]　The National Institute of Japanese Language[*3]
Ochanomizu University[*4]

Abstract
It is important in many information extraction tasks to distinguish factual assertions from uncertain ones. This paper reports a preliminary study to construct a schema to annotate key expressions (modal, negative, etc) and their scopes in Japanese texts.

１．目的

自然言語のテキストには、事実だけでなく、推測、仮定、仮想現実など様々なタイプの情報が含まれる。たとえば、以下の例の下線部のうち、「事実」（少なくとも、記事の書き手が「事実」と考えている情報）と言えるのは (1)のみで、(2) (3)は伝聞や推量の結果というように、情報の確実性に差がある。(4)−(6)に至っては表層的には当該の文を含んでいるが、現実世界についての言明ではない。

(1). 県内で新型インフルエンザが発生した。
(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). (a) 県内で新型インフルエンザが発生したようだ。/ (b)県内で新型インフルエンザが発生した可能性がある。（推量）

(4). (a) 県内で新型インフルエンザが発生しなかった。(通常否定) /(b) 県内で新型インフルエンザが発生したというのは誤報だった。(メタ否定)

(5). 県内で新型インフルエンザが発生した場合、どう対応するべきか。(仮定)

(6). 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)

これらの情報タイプを識別する主な手がかりは、様相(文の内容に対する書き手の認識・判断)を表す表現や、否定表現などによって形成されるある種の「意味的文脈」である。人間は、自然言語で書かれた情報を読むとき、これらの文脈に基づいて「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。計算機によって情報の確実性を判断し、可能な限り「確実な」情報を効率的に抽出したい場合、これらの「意味的文脈」の認識を可能にする必要がある。

筆者らは、そのような「確実性の自動的な認識」の基盤として、日本語のニュース記事に対し「確実性判断に関わる意味的文脈」をアノテーションしたコーパスの構築を開始している。本論文では、アノテーションスキーマの概要について述べる。

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事実と考えている度合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」という言葉を、誤解が生じない限りにおいて「テキストの書き手が『事実』と考えている情報」を指して使う。

本論文では以下、第二節で先行研究を概観し、第三節で本研究のアプローチの特徴的な点について解説する。

2.先行研究

先行研究については、英語の医学・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 identification)の研究があるが、現段階ではこのテーマを扱った研究はまだ少ない。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。

3. アプローチの概要

本研究で構築するコーパスは、テキストに対して 1）様相表現や否定表現その他の「確実性」に関わる言語表現のアノテーション、および 2）それらの表現のスコープのアノテーションを与えるものである。ここまでは Szarvas et al(2008)の方針とほぼ同じであるが、日本語のテキスト、しかもニュース記事や報告を対象にするということから、いくつか留意しなければならない点が生じる。その結果、本研究のアプローチは次の三つの特色を持ったものとなっている。

- アノテーション対象の拡張：伝聞、仮定、仮想現実、比況を表す表現、叙実述語などを含める。
- 表現間の相互作用を重視
- 表現のスコープが二文以上に及ぶケースに対処

以下、それぞれについて詳しく説明する。

3.1 アノテーションの対象

一般のテキストには、伝聞、仮定、仮想現実、比況などを表す表現が多く現れるため、否定や推量を表す表現に加えてこれらもアノテーションの対象とする。これらは過去の hedge identification の研究においてはほとんど扱われていない。また、「知る」「公表する」などの叙実述語のように、従属節の内容が事実であるという前提を導入する表現もアノテーションする。叙実述語は命題の確実性に強く影響するものであるにも関わらず、過去の研究では Kilicoglu and Bergler(2008)が "unhedger"として取り上げているのみである。以下では、マークアップの対象となる表現を(1)様相表現、(2)否定表現、(3)仮定表現に分けて紹介する。

様相表現

ここでは、助動詞だけでなく、話し手（書き手）の命題内容に対する態度を表すもの全般、すなわち動詞、副詞などもアノテーションすることを視野に入れている。日本語の様相表現（モダリティ表現）には、さまざまな分類が提案されている。特に、「らしい、ようだ、かもしれない、はずだ、だろう」のような、推量にかかわるとされる助動詞は、過去の研究の間でも分類が一致せず、議論の分かれるところである。ここでは、「確実性」という観点から、ヨウダ類（らしい・ようだ・みたいだ）とダロウ類（だろう・かもしれない・可能性がある・はずだ）の区別を重視する。ヨウダ類は証拠推論を表すとされる（寺村(1984)、Aoki(1986)、森本(1994)等）。証拠推量は、話し手（書き手）が、推量の根拠の存在を示すものである(Palmer(2001))とされる。これに対し、ダロウ類は田窪(2001)で「ダロウ類は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である」と特徴づけられる。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、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、理にかなったことであるように思われる。田窪においては、条件文や「今ごろ」「どうも」等の副詞を用いたテストによってこれらの間の意味的な違いが示されている。

(7). 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た｛だろう・かもしれない・可能性がある・はずだ｝。
(8). *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た｛ようだ・らしい・みたいだ｝。

これらのテストに基づくと、「思う」「信じる」などの認識動詞もダロウ類と同じ振る舞いをするので、ここではダロウ類として扱う。

ここでは、伝聞、比況、「か」「かどうか」などの疑問を表す表現、叙実述語なども様相表現の一種として扱う。伝聞情報の確実性は情報ソースの信頼性に依存するため、伝聞の場合は**SOURCE**という任意の属性で情報発信者を記述するようにする。様相表現の分類と各カテゴリへのアノテーションを表1に示す。

アノテーションの例は以下の通りである。<MODAL>および<NEG>は様相表現、否定表現をマークアップするものであり、<SCOPE>はそれらのスコープを表している。MODAL 要素内の scope 属性は、その表現のスコープを指定する。

(9). <SCOPE id="001" >県内で新型インフルエンザが発生した</SCOPE><MODAL id="003" type="evidential" scope="001">**ようだ**</MODAL>
(10). 県健康推進課が<SCOPE id="004">県内で新型インフルエンザが発生したと</SCOPE><MODAL id="002" type="hearsay" source="県健康推進課" scope="004">**報告した**</MODAL>
(11). <SCOPE id="005">県内で新型インフルエンザが発生したことが</SCOPE><MODAL id="006" type="factive" scope="005">**公表された**</MODAL>

否定表現

ここでは、通常の否定と Horn (1985) のいうメタ否定を区別する。Horn では、メタ否定は次のように特徴づけられている。

[m]etalinguistic negation must be treated not as a truth-functional or semantic operator on propositions, but rather as a device for objecting to a previous utterance on any grounds whatever -- including the conventional or conversational implicata, its morphology, its style or register, or its phonetic realization. (Horn 1985: 133).

表現の分類	アノテーション	表現の例
証拠推論（ヨウダ類）	<MODAL type ="evidential" scope="（スコープのid）">	ようだ、らしい、みたいだ
非可視的状況推論（ダロウ類）	<MODAL type ="epistemic" scope="（スコープのid）">	だろう、かもしれない、可能性がある、はずだ、思う、信じる
伝聞	<MODAL type ="hearsay" SOURCE="（情報のソース）" scope="（スコープのid）">	そうだ、らしい、によると、とのことだ、という、伝える、報じる、言う、（と）する、発表する
比況	<MODAL type =" simile" scope="（スコープのid）">	ようだ、まるで、みたいだ
疑問	<MODAL type ="question" scope="（スコープのid）">	か、かどうか
叙実述語	<MODAL type ="factive" scope="（スコープのid）">	（ことを）知る、（ことを）確認する、（ことを）公表する

表1．様相表現の分類とアノテーション

表現の分類	アノテーション	表現の例
通常否定	<NEG type ="normal" scope="（スコープのid）">	ない、V-ない、
メタ否定	<NEG type ="meta" scope="（スコープのid）">	のではない、ということはない、嘘である、誤報である、間違いである、正しくない

表2．否定表現の分類とアノテーション

例えば以下のThe king of France isn't baldの場合、二通りの解釈が可能で、一つはThe king of France is baldという命題を否定する解釈（通常否定）で、もう一つはThe king of France is baldという発話そのものが不適切であると異を唱える解釈（メタ否定）である。後者の例としては(12)のように、The king of France is baldという命題の前提が成り立たないことから、The king of France is baldという文の「発話可能性」に異を唱える解釈がある。つまり、日本語に対応させると「The king of France is baldと言うことはできない。だって、フランス王なんて存在しないのだから」という解釈を与えることが可能で、従って、 there ISN'T any king of Franceという発話が後続可能になる。

(12). The king of France isn't bald -- there ISN'T any king of France.

詳しくは次節で述べるが、否定表現が叙実述語を埋め込んでいる場合、通常の否定では叙実述語の

従属節の意味内容の確実性は変化しないが、メタ否定では変化するという違いがある。このようなことから、ここではこの二つの否定を表2のように区別してアノテーションする。

アノテーションの例は以下の通りである。

(13). <SCOPE id="007">県内では新型インフルエンザは発生してい</SCOPE><NEG id="008" type="normal" scope="007">**ない**</NEG>

(14). <SCOPE id="009">県内で新型インフルエンザが発生したというのは</SCOPE><NEG id="010" type="meta" scope="009">**正しくない**</NEG>

仮定表現

仮定表現については、「たら」「れば」「と」などの形式に加え、「とする」のような形式もCOND要素を使って以下のようにアノテーションする。

(15). <SCOPE id="001">県内で新型インフルエンザが発生し<COND id="002" scope="001">たら</COND>、パニックになるだろう</SCOPE>。

(16). <SCOPE id="003">県内で新型インフルエンザが発生した</SCOPE><COND id="004" scope="003">とする</COND>。

3.2 表現間の相互作用

日本語には様相や否定を表す表現が多く存在し、複数の表現の埋め込みも多くみられる。本研究では埋め込みによる表現間の相互作用を扱うことを視野に入れている。複数の表現間の埋め込みは、そのスコープ内にある命題の確実性に影響する。本研究ではこのような表現間の相互作用を扱うことも視野に入れている。たとえば、次の四つの例では、埋め込みが深くなるほど確実性が弱くなると判断できる。

(17). 県内で新型インフルエンザが発生したと報じられた。

(18). 県内で新型インフルエンザが発生したと報じられた可能性がある。

(19). 県内で新型インフルエンザが発生したと報じられた可能性があるとみられている。

(20). 県内で新型インフルエンザが発生したと報じられた可能性があるとみられているようだ。

ただし、単に埋め込みが深ければ深いほど、徐々に確実性が低くなっていく、というわけではない。次に見られるように、叙実述語の導入する従属節の確実性は、「ようだ」のような様相表現や「ない」のような否定表現に埋め込まれても低くならない。しかし、「ということはない」のようなメタ否定に埋め込まれると、節の内容は必ずしも事実とは言えなくなる。この現象は、前提投射(presupposition projection)として盛んに研究されている。

(21). [[県内で新型インフルエンザが発生したことが**公表された**] **ようだ**]。(下線部は「事実」のまま)

(22). [[県内で新型インフルエンザが発生したことは**公表されて] いない**]。(下線部は「事実」のまま)
(23). [[県内で新型インフルエンザが発生したことが**公表された] ということはない**]。(下線部は必ずしも「事実」ではなくなる)

また、田窪(2001)に指摘されているように、「だろう」「かもしれない」「可能性がある」などの様相表現は、仮定の後件に現れるとしばしば「反実仮想」を表す。すなわち、命題の内容が「偽である」ことが含意されることになり、これらが単独で現れる場合とは対照的である。

(24). 県内で新型インフルエンザが発生した可能性がある。
(25). もし当局の対応が実際より1時間でも遅れていたら、県内で新型インフルエンザが発生した可能性がある。(反実仮想になる)

3.3 表現のスコープ

これまでに挙げてきた例だけを見ると、様相表現や否定表現のスコープはそれらの表現の統語的な係り先(理論言語学でいう「c統御領域」)であるように見えるかもしれない。しかし、ニュース記事などには、表現のスコープが二文以上にまたがるように見える例も頻繁に見られる。たとえば次の(28)(29)で下線で示される文は、それ自体には様相表現などが含まれていないにも関わらず、前の文の様相表現のスコープに入っているように見える。

(26). 市の健康推進課**によると**、感染したのは四十代の男性。市内の料理店で食事をした後症状を訴え、感染が確認された。食事に同席した男性も入院中だという。
(27). 県内で新型インフルエンザが発生したら、パニックが起こる**可能性が高い**。まず、病院に人が殺到する。鶏肉や鶏卵の不買運動も起こるだろう。

この現象に対処するため、ここでは1)音声を持たないMODAL要素の存在を認めると同時に、2)MODAL要素にref_id属性を追加し、他のMODAL要素を参照できるようにする。すなわち、(28)(29)の下線の文には“ゼロ”の様相表現があり、前の文に明示的に現れる様相表現と参照関係を持つと考えるのである。具体的なアノテーション例は次の通りである。

(28). 県健康推進課<MODAL id="001" type="hearsay" source="県健康推進課" scope="002">によると</MODAL>、<SCOPE id="002">感染したのは四十代の男性</SCOPE>。<SCOPE id="003">市内の料理店で食事をした後症状を訴え、感染が確認された</SCOPE>**<MODAL ref_id="001"scope="003"/>**。

4. 結語

本研究では今後、ここで概要を紹介したスキーマに基づき、ニュース記事からアノテーション済みコーパスを構築する予定である。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。

更に、様相や前提が文の確実性をどのように変化させるかを「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予定である。しかし、公理論的確率論と様相・前提の組合せは必ずしも自明ではない。

情報検索や情報抽出において求められるのは、多くの場合、確実性の高い情報である。しかしながら、情報の確実性は表層的なパターン認識においては特定が難しく、現段階では人間が判断せざるを得ない。本研究では、確実性の高い情報をより効率的に取り出し、人間による情報の評価をサポートするシステムの基盤となるコーパスの構築を目指している。特に、大量の情報を短時間に評価することが求められる多くのタスク（感染症情報、災害情報、薬品の毒性報告、事故情報の監視など）において貢献できると考えている。

謝辞

本研究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

[1]. Aoki, H. (1986) “Evidentials in Japanese.” In Chafe, W & Nichols, J.(eds) E*videntiality. Ablex Publishing Corporation*:223-237.

[2]. Horn, L. R. (1985). "Metalinguistic negation and pragmatic ambiguity." *Language* 61.1, 121-74.

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

[4]. Light M, Qiu X, Srinivasan P. (2004) “The language of bioscience: facts, speculations, and statements in between.” In P*roceedings of BioLink 2004 workshop on linking biological literature, ontologies and databases: tools for users*, Boston, May 2004.

[5]. Medlock B, Briscoe T. (2007) Weakly supervised learning for hedge classification in scientific literature. P*roceedings of 45th Meeting of the Association for Computational Linguistics 2007*:992-999.

[6]. Palmer, F.R. 2001 *Mood and Modality second edition.* Cambridge University Press.

[7]. Szarvas G, Vincze V, Farkas R, Csirik J (2008) “The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts”. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* 2008:38-45.

[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.

[9]. 寺村秀夫 1984 『日本語のシンタクスと意味Ⅱ』くろしお出版

[10]. 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

[11]. 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』太学社