B2IM2032

修士論文

自然言語処理における数量表現の取り扱い

成澤 克麻

2014 年 2 月 10 日

東北大学 大学院
情報科学研究科 システム情報科学専攻

本論文は東北大学大学院情報科学研究科に
修士 (工学) 授与の要件として提出した修士論文である。

成澤 克麻

審査委員：

乾 健太郎 教授　　(主査)
木下 哲男 教授
伊藤 彰則 教授
岡崎 直観 准教授

自然言語処理における数量表現の取り扱い*

成澤 克麻

内容梗概

テキスト t が仮説 h と含意関係にある、すなわち t が h を含意するとは、t から h が推論可能であるような関係のことを指す。含意関係を認識する技術は、質問応答や情報抽出、機械翻訳など多くの言語処理アプリケーションで重要な役割を果たすことが期待されている。文書間の含意関係を認識するためには幅広い言語現象に対応する必要があるが、本研究では含意関係認識で必要になる言語現象の 1 つ、数量表現が関わる推論に焦点をあてる。本稿での寄与は主に 3 つある。

まず、既存の含意関係コーパスを分析し、数量表現を伴う文における含意関係認識にどのような課題があるのか明らかにする。現状で数量表現に対して十分に対応ができていない原因の 1 つは、既存研究において内在する問題の分析・整理が不十分である事である。分析の結果、我々は 11 のカテゴリに問題を分類した。

次に数量表現の規格化に関するアノテーション基準を提案し、そして数量表現を抽出・規格化する手法について述べる。数量表現の規格化とは、それぞれがどんな数量を示しているのか計算機に理解できる一定の形式に変換することを指す。規格化は計算機が数量表現や時間表現を理解するために必要な処理の第一歩であり、様々な言語処理のアプリケーションにおいて有用である。

最後に数量の大小を自動で判定する手法を提案する。本稿では Web から数量表現を抽出し、これを用いて大小を判定する 2 つの手法を紹介する。

キーワード

自然言語処理, 数量表現, 含意関係認識, 固有表現認識, 規格化

*東北大学 大学院 情報科学研究科システム情報科学専攻 修士論文, B2IM2032, 2014 年 2 月 10 日.

目 次

図 目 次

表 目 次

1　はじめに

The PASCAL Recognising Textual Entailment(RTE) Challenge [1] に代表されるように、**含意関係認識**が近年研究者の注目を集めている。テキスト t が仮説 h と含意関係にある、すなわち t が h を**含意**するとは、t から h が推論可能であるような関係のことを指す。この含意関係を認識する技術は、質問応答や自動要約、情報抽出、機械翻訳など多くの言語処理アプリケーションで重要な役割を果たすことが期待されている。以下に含意関係にある文章の例を挙げる。

(1)　t：インターネット広告は１５％伸びたが、ネットワークテレビの広告は３・５％しか伸びなかった。

　　h：インターネット広告はネットワークテレビより伸びている。

(2)　t：近い将来、最悪の場合 30 億人が水不足に直面する。

　　h：近い将来、世界は深刻な水不足になると懸念されている。

この文の含意関係が認識できれば、例えば質問応答においては例 1 の h がクエリであるとき、t のような記述があるテキストコーパスを持っておくことで、質問応答システムはその中から答えを導くことができる。また要約システムにおいても、t が h を含意している、すなわち h が t の内容に含まれていることがわかれば、t の文を削除してより簡潔な h の文を要約として用いるといった処理が可能になる。

文書間の含意関係を認識するためには複雑な言語現象に対応する必要がある。これに対して近年、複雑な言語現象を 1 つ 1 つの基本的な言語現象のレベルまで落とし込み、それらの基本的な言語現象を 1 つ 1 つ解決し含意関係認識システムの精度向上を図ることが提案されている [2–5]。

このような背景の元、本研究では含意関係認識で必要になる言語現象の 1 つ、**数量表現**が関わる推論に焦点をあてる。例えば上の例では「１５％」と「３・５％」を比較して「〜より伸びている」ことを導く必要がある。他にも数量表現が関わる推論を行う必要がある含意関係としては「30 億人が水不足」⇒「深刻な水

不足」だったり、「約２２０万人」⇒「８０万人以上」など実に様々な例が挙げられる。数量表現を伴う文に対して現状の含意関係認識の研究は十分な対応ができていないことは、Sammons らの分析でも指摘されている [3]。数量表現は自然言語文において広く用いられる表現であり、数量表現が関わる推論は含意関係認識を行う上で重要である。

本研究の目指すところは、含意関係認識における数量表現の問題を解決し、高精度な含意関係認識システムを構築することである。また、これを通して言語処理における数量的意味の計算方法を検討する。本稿での寄与は以下の 3 つである。

1. **数量表現を伴う文における含意関係認識の課題分析** 成澤 [9] で行った課題分析について、再度分析をし直したものを報告する。成澤 [9] では既存の含意関係コーパスを分析し、数量表現を伴う文における含意関係認識における課題を 7 つのカテゴリに分類し、今後取り組むべき課題を明らかにした。しかしこの分析では、「その他（分類ができなかったもの）」に属する問題分類が多く、十分な分析が行えていなかった。また既存コーパスの分析をもとに課題の分類を行ったものの、潜在的に問題になりそうなものを含めて分類を行っており、実際のコーパス中にはあまり含まれていない問題も 1 つのカテゴリとして分類を行っていた。本稿ではこれらの問題を解決するために行った再分析の結果を述べる。分析結果は HP 上で公開しているので参照されたい[1]。

2. **数量表現・時間表現の抽出と規格化** 数量表現の規格化に関するアノテーション基準を提案し、そして数量表現を抽出・規格化する手法について述べる。数量表現が関わる推論現象に対応するためには、まず数量表現を認識すること、またその数量表現が持つ基本的な意味を理解することが必要である。そこで本稿では数量表現の規格化について述べる。数量表現の規格化とは、それぞれがどんな数量を示しているのか計算機に理解できる一定の形式に変換することを指す。規格化は計算機が数量表現や時間表現を理解するために必要な処理の第一歩であり、様々な言語処理のアプリケーションにおいて有用である。例えば「三キロメートル」と「3000m」という表現が同

[1] http://www.cl.ecei.tohoku.ac.jp/~katsuma/resource/rte_analysis_with_numexp/

じ数量を指すことを認識することは、情報抽出などで有益であろう。本稿では文中の数量表現にタグを付与し、規格化した意味表現をタグの属性として数量表現に付与するという形で数量表現の抽出・規格化を行う。本稿では数量表現と似た性質をもつ時間表現も対象としてアノテーション基準、また抽出・規格化手法について述べる。

手法を実装したツールは HP 上で公開されている[2]。我々が開発したツールは入力された文章中の数量表現・時間表現を抽出し、規格化して出力する。規格化においては、数の表記の揺れの吸収（例えば「1989」「１，９８９」「一九八九」「千九百八十九」）や、単位の統一（例えば「キログラム」「mg」「トン」を全て「g」という単位に統一)、時間に関する表記揺れの吸収（例えば「2012 年 1 月 1 日」「2012/1/1」「2012-1-1」)、　修飾語の考慮（例えば「以上」「約」「くらい」）が行われる。ツールはルールベースで実装されており、認識に失敗した表現に対しても辞書に表現を追記することで容易に対応が可能である。

3. **数量の大小の自動判定** 数量の大小を自動で判定する手法を提案する。数量表現の規格化により、数量同士の含意関係（「30 億人」⇒「二十億人以上」など）が認識できるようになるが、冒頭例 2 の「30 億人が水不足」⇒「深刻な水不足」を導くために必要な、数量と人間の主観の間の含意関係の問題は対処できない。すなわち「数量表現が示す数量の認識」（「30 億人」という表現が、どんな数量を示すか）だけではなく、「数量が示す意味の解釈」（70 億人という数量が、どんな意味をもつか）が必要である。そこで本研究では、数量の解釈にむけた第一歩として、数量の大小を判定するという問題に取り組む。大小に着目するのは、多くの数量の解釈の根本は数量の大小の理解にあると考えられるためである。例えば、冒頭の例における尤もらしい推論の流れは以下のようになる。我々が今回取り扱うのは、1 つ目の

[2] http://www.cl.ecei.tohoku.ac.jp/~katsuma/software/normalizeNumexp/

推論である「30 億人」⇒「たくさんの人」という推論である。

*30*億人 が水不足に直面する
⊨ たくさんの人 が水不足に直面する
⊨ 深刻な水不足に直面する

本稿ではウェブから数量に関する記述を大量に抽出し、これを用いて数量の大小を判定する手法を提案する。

2、3 については手法の有効性を示すために評価実験を行った。本研究の目的を考えれば、実際の含意関係認識システムに本稿の手法を組み込み、含意関係認識の精度が向上するかどうかを調査するのが理想的な評価実験ではあるが、今回はそういった実験は行わなかった。これは既存の含意関係コーパスは数量表現が関わる推論以外にも様々な言語現象を含み、仮に提案手法により数量表現の問題が解決されたとしても他の問題により含意関係認識が正しく行われないことが考えられ、提案手法の有効性に焦点を当てた評価を行うことが難しいと考えたためである。

本稿の構成は以下の通りである。まず 2 章で関連研究について触れる。3 章では含意関係認識における課題の分析を行う。4 章ではまず数量表現・時間表現のアノテーション基準を提案し、その後 5 章で抽出・規格化の手法を述べ、提案手法の評価を行う。6 章では数量の大小の自動判定手法を提案し、この評価実験を行う。

2　関連研究

2.1　数量表現処理に関する既存研究

自然言語処理の研究において、数量表現を扱う研究は驚くほど少ない。数量表現を扱う研究としては、情報抽出における Bakalov ら [10]、情報検索における吉田ら [12] と Fontoura ら [11]、また質問応答における Moriceau ら [13] の取り組みなどが挙げられる。吉田ら [12] と Fontoura ら [11] は「200〜800 ドル」のような範囲を扱った数量をクエリとしてテキストを検索する手法を報告している。しかし、これらの手法では数量表現を単なる文字列として扱っており、数量表現が示す数量（例えば「30 万」が「300000」を示すこと）は認識できていない。情報抽出における数量表現は、製品の値段や重さであることが多い。

質問応答においては、答えの正当性をチェックするために、IBM's PIQUANT [14, 15] が Cyc [16] の情報を用いている。例えば、このシステムでは「200 マイル」という答えが「エベレストの高さ」という質問に対して候補に挙がった場合、これを除去する。なぜなら、Cyc の知識として、山の高さは山の高さは 1000〜30000 フィートでしかないことが分かっているからである。またこのシステムでは数の有効桁数の問題（500 万と 510 万と 5,200,390）や単位の変換（「平方キロメートル」と「エーカー」）などといった問題にも対処している。

2.2　含意関係認識

RTE-6 [17] では 5 つのシステムが数量表現間の対応付けに取り組んだが、その他の 13 システムの論文では数量表現に対する扱いが不明であった [18–22]。ここでの対応付けとは、テキスト t と仮説 h に含まれる数量表現の包含関係を認識することで、例えば "at least 35" は "at least 30" を含み、対応関係にあることをこれらの 5 つのシステムでは認識している。ただし、論文中で詳しい記述はあまりされておらず、具体的にどのような表現に対応しているのか、また単位の問題には対応しているのか（kg と g のような関係）など、不明な点は多い。Majumdar ら [22] は RTE-6 のデータに多くの数量表現が含まれているが、数量表現が含む

情報について記述している言語資源は存在しないため、このようなモジュールを作成したと述べている。

数量表現を伴う文における含意関係認識には多くの課題が残されている。Sammons ら [3] は RTE-5 [23] のデータを基に、含意関係の推論のために必要とされる含意現象を分析した。この分析の中で、先に述べた数量表現間での含意関係と、数に関する推論を取り上げている。RTE-5 に提出されたシステムは数の推論にほぼ未対応であり、今後はこの問題に対応していく必要があると Sammons らは述べている。また LoBue ら [6] は含意関係認識に必要となる世界知識を 20 のカテゴリに分けて論じ、その知識のカテゴリの 1 つに足し算や引き算などといった算術を行うための知識を定義している。LoBue らはこの知識はこれまで多くの研究で無視されてきた知識であると述べると同時に、含意関係認識において比較的必要とされる頻度の高い知識であると述べている。ただし、これらの研究では問題の分析には至っておらず、また現状のシステムで数量表現間での含意関係認識がどれほどの精度で行われているのかも明らかでない。

日本語含意関係認識の分野では更に研究が少ない。2011 年に行われた RITE [7] では、数量表現の問題に対処しているグループは存在せず、類似した処理として時間表現間の含意関係の問題に対応しているグループが 2 つ見られたのみであった [24, 25]。また日本語の数量表現は文中の様々な位置に表れるなど固有の性質を持ち、この扱いについて日本語形式意味論では様々な議論が行われているが [26–28]、自然言語処理においてはあまり議論がなされていない。唯一、機械翻訳において様々な位置に表れる数量表現を正しく英語に翻訳するための研究がみられる [29]。

これに対して成澤 [9] では既存の含意関係コーパスを分析し、数量表現を伴う文における含意関係認識における課題を 7 つのカテゴリに分類し、今後取り組むべき課題を明らかにした。この分析では、約半数の問題に対して必要となる推論を明確に示したが、残りの半数については上手く分析ができず、様々な推論が複合した切り分けにくい問題だと述べた。また具体的な既存コーパス中のそれぞれの推論タイプの事例数を示すことはしなかった。

2.3　数量表現、時間表現の抽出・規格化

まず初めに、時間表現に関する取り組みから紹介する。時間表現は固有表現抽出のタスクとして研究が続けられてきた。MUC-6 [30]、MUC-7 [31] では「DATES（日付）」と「TIME（時間）」を抽出するタスクが行われた。日本においても、MUC-6 を踏襲した IREX [32] が開催され、また関根の拡張固有表現 [33] においても「時間」「期間」「その他」の 3 種類の時間表現が対象とされている。MUC-7 と関根らの拡張固有表現の主な違いとして、MUC では *yesterday*, *three days ago* などの相対的な時間表現を扱っていること、関根らでは「3 時間」などの時間の量を表す時間表現を扱っていることが挙げられる。相対的な時間表現に関しては、関根らは質問応答タスクでの使用を強く意識しており、相対的な時間表現は質問応答の答えとして適していないため抽出しない、と述べている。CoNLL [34] における固有表現認識のタスクでは、時間表現も数量表現も対象とされていなかった。

TERN (Time Expression Recognition and Normalization) (DARPATIDES 2004) では、時間情報の曖昧性解消・正規化がタスクとして追加され、様々な時間表現解析器が開発された。さらに、時間表現の抽出だけでなくイベントの時系列認識をしたいという要求が高まると、時間表現とイベントとを関連づけるタグづけ基準 TimeML [35] が検討され、TimeML に基づくタグつきコーパス TimeBank [36] などが整備された。TimeML で対象とする時間表現は「日付 (DATE)」「時刻 (TIME)」「時間 (DURATION)」「頻度集合 (SET)」の 4 種類で、相対的な時間表現も対象としている。

2007 年には評価型ワークショップ SemEval-2007 におけるサブタスクとして時間表現－イベント間及び 2 つのイベント間の時系列関係を推定する TempEval [37] が開かれ、種々の時間的順序関係推定器が開発された。後継のワークショップ SemEval-2010 におけるサブタスク TempEval-2 [38] では、英語だけでなく、イタリア語、スペイン語、中国語、韓国語、フランス語を含めた 6 言語が対象となった。TempEval-2 の時間表現を認識・規格化するタスクにおいては、時間表現の認識精度は最も良かったシステムが f 値で 0.86、規格化の精度は 0.85 であった。このタスクにおいて多くのグループはルールベースの実装を行っており、最高精度のシステムもルールベースによる実装であった。Lorens ら [39] では、更に高い

精度を得るためには大量のルールが必要であり、そのために大勢の人々の手でともにルールの記述をしていくことを主張し、ルールの拡張性・再利用性やシステムの多言語化、評価のしやすさに焦点をおいた TIMEN を公開した。TempEval-2 のデータセットを用いた TIMEN の評価実験では、認識の再現率が 1.0、規格化精度が 0.90 となっている。一方、日本語においては小西らが <TIMEX3> タグに基づいた時間情報アノテーションの枠組みを提案している。含意関係認識の文脈では、Tsuboi ら [24] が簡単な時間表現の規格化の問題に対応するモジュールを作成し、この規格化モジュールがタスクの精度を向上させたと述べている。同様の処理は Watanabe ら [25] においても行われている。

数量表現も固有表現抽出の抽出対象として研究がなされてきた。MUC-6 では「金額」「割合」を抽出するタスクが開かれ、また関根の拡張固有表現においても「数値表現」というカテゴリが設けられている。しかし、イベントの時系列認識が質問応答や情報抽出などの文脈で重要視され時間表現の規格化の研究が盛んに進められてきたのに対し、数量表現の規格化にはこれまで焦点がおかれてこなかった。数少ない試みとして、含意関係認識の評価型ワークショップである RTE-6 [17] において、複数のグループが数量表現の規格化に取り組み、異なる表現の数量表現のマッチング（例：「3 人以上」と「10 人」）を試みたことが報告されている [18–22]。ただし、規格化について詳細に述べているグループは見られず、どの程度の規格化を行っているかは不明である。数量表現に関しても、時間表現と同じように、深い意味理解のためには認識・規格化・対応するエンティティーの同定などといった処理が必要となるはずだが、既存研究でこれらはほぼ行われていない。

固有表現抽出の分野では、関根らの拡張固有表現 [33] が数量表現と時間表現の 2 つに対し細かな分類を与えている。しかし前述した通り、表現を認識するだけでは含意関係認識や質問応答における問題に十分に対応できないため、含意関係を認識するためには規格化処理が必要がある。

2.4 数量に関する知識獲得

荒牧ら [40] が物体のサイズに関する知識を抽出し、それを用いて物体間の意味的関係を予測するというタスクに取り組んでいる。例えば、「本」の大きさが 20cm×25cm であること、図書館が 10m×10m であることが分かれば、図書館が本を内包する関係にあることが導けるだろう、というものである。この手法では、物体のサイズについての知識をルールベースで（例えば "*book (*cm x *cm)*" のような正規表現で）Web から抽出し、それらを用いて関係分類を行っている。

Davidov ら [41] は物体の高さや重さのような属性値をウェブから抽出、推論する手法を提案している。Davidov らの手法ではルールベースで属性値を抽出する（例えば "*Object is * [unit] tall*"）他に、抽出した属性値と物体と物体の関係の情報（「A は B より重い」「A は C より軽い」など）を用いて、未知の属性値を推論するという処理も行っている。

荒巻らの手法も Davidov らの手法も、どちらも人手で作ったルールを用いて特定の数量属性（重さ、高さ、サイズ）を抽出している。これに対して、我々の手法は全ての対象、全ての状況についての数量の知識を柔軟に抽出するものである（例えば「水不足に困るであろう人の数」）。

3　数量表現を伴う文における含意関係認識の課題分析

本章では成澤 [9] で行った課題分析について、再度分析をし直したものを報告する。成澤 [9] では既存の含意関係コーパスを分析し、数量表現を伴う文における含意関係認識における課題を 7 つのカテゴリに分類し、今後取り組むべき課題を明らかにした。しかし前回の分析では、「その他（分類ができなかったもの）」に属する問題分類が多く、十分な分析が行えていなかった。また前回は既存コーパスの分析をもとに課題の分類を行ったものの、潜在的に問題になりそうなものを含めて分類を行っており、実際のコーパス中にはあまり含まれていない問題も 1 つのカテゴリとして分類を行っていた。本稿ではこれらの問題を解決するために行った再分析の結果を述べる。

成澤 [9] で用いた含意関係コーパスは、RITE [7] の開発データとテストデータ (BC,MC の計 940 文対× 2、日本語) と小谷ら [42] の評価セット (2471 文対）である。RITE は NTCIR による含意、言い換え、矛盾の認識を目的とした評価型ワークショップであり、日本語、中国語、英語の各言語に対してコーパスが提供されている。RITE が新聞や wikipedia などの実際の文を用いて作成された比較的含意関係認識が難しいとされるコーパスであるのに対して、小谷らのコーパスは人手で作られた比較的易しい評価セットで、ほとんどの問題において表現のずれは 1 箇所程度である。またどのような推論が必要なのかも簡単に述べられている。

成澤 [9] ではコーパス中から文対のどちらか、または両方に数量表現が含まれる事例を抽出した。ここで数量表現は「数と単位によって、個数や量、順番を表現するテキスト中の文字列」とする。例えば「3 人」は数 (3) と単位 (人) で人数という個数を示しているため、数量表現と言える。前回の分析では細かい定義を行わなかったが、今回の分析ではより厳密に数量表現を定義して分析を行った。数量表現のより詳しい定義は次章にて紹介する。

同じコーパスを用いて再度分析をし直した結果、数量表現が含まれる事例は 4351 文対中 371 文対であった。この中で含意（もしくは矛盾）の関係を導くために数量表現の関わる推論が必要となる事例のみを抽出し、必要な推論の種類で 11 つのカテゴリに分類した。含意関係を導くために数量表現の関わる推論が必要となる事例とは、冒頭であげた事例のようなものである。文中に数量表現を含むも

のの、含意関係を導くために数量表現が推論に関わらない事例とは、例えば以下のような例が挙げられる。以下の例では数量表現（「一つ」「五百人」）が含意関係を導くための推論に深く関わっているとは言い辛い。

(3) t：無知の知では詐欺一つ起こせないにもかかわらず、無知の自覚（知）を促して歩く哲学者だったソクラテスは、裁判にかけられ、五百人からなる裁判員（陪審員）の表決で死刑になった。

h：ソクラテスは死刑になった。

最終的に我々は 114 文対を抽出し、これを 11 つのカテゴリに分類した（ただし重複を許す)。表 1 はその概要である。

前回の分析では「その他」に属する問題分類が多かったが、今回の分析では全体の 90%を分類することができた。分類の仕方の主な変更点は、前回は文の構造に着目して分類を行っていたが、より推論そのものの種類として近いものという着眼点で分類を行ったことが挙げられる。また前回は既存コーパスの分析をもとに課題の分類を行ったものの、潜在的に問題になりそうなものを含めて分類を行っており、実際のコーパス中にはあまり含まれていない問題も 1 つのカテゴリとして分類を行っていた。今回は実際のコーパスに含まれる問題に即してカテゴリ分けを行い、前回は明示しなかった出現頻度についても述べている。

頻度が高いカテゴリについて簡単に触れる。

- 「数量表現間の含意」は「約 220 万人」と「80 万人以上」のような、数量表現間の含意を認識する必要がある問題である。この例では数量の範囲を認識する必要が有るが、他には「52.4 パーセント」なら「約五割」など単位の違いを考慮する必要がある問題も挙げられる。数量表現が示す数量を理解する必要がある問題であり、最も基本的な問題と言える。本稿では数量表現をある形式の意味表現に変換し（これを規格化と呼ぶ）この問題の解決に取り組む。

- 「数量の解釈」は数量の大小を解釈する必要がある問題である。例えば表の例では「70 億人が水不足」というのは「深刻な水不足」なのだと解釈する

表 1: 分析におけるそれぞれのカテゴリの事例数と簡単な定義

カテゴリ名	定義	例	#
数量表現間の含意	単位の違いや数量の範囲などを考慮して数量表現を紐づける必要がある事例	*t*: 全国のアルコール依存症者は、約２２０万人といわれる。 *h*: アルコール依存症は８０万人以上いると推計されている。	32
数量の解釈	数量表現が示す数量を解釈する必要がある事例	*t*: ２１世紀半ばには最悪の場合、全人口の７割以上にあたる７０億人が水不足に直面する。 *h*: 近い将来、世界は深刻な水不足になると懸念されている。	12
語彙知識	単語の数量的な側面から推論する必要がある事例	*t*: 佐藤夫妻は 25 回目の結婚記念日を迎えた。 *h*: 佐藤夫妻は銀婚式を迎えた。	12
算術処理	加算や減算などの処理を行う必要がある事例	*t*: ２０００円札の流通枚数が５０００円札の流通枚数を１０００万枚超えたことがわかった。 *h*: ４億４０００万枚だった５０００円札に対し、２０００円札の流通枚数が４億５０００万枚となった。	11
動詞による範囲表現	数量の範囲が数量表現ではなく動詞により表される事例	*t*: ８３年５月の日本海中部地震（M７・７）による津波は、最大波高が１３・８メートルに達したとの記録がある。 *h*: 日本海中部地震による津波は１０メートルを超えていた。	9
序数、時間表現	序数や時間表現に対応する数量表現を理解し推論する必要がある事例	*t*: 第三次世界大戦では、多くの貴い人命が犠牲になった。 *h*: これまでに三度の世界大戦があった。	9
単純な書き換え	単純な書き換えを含む事例	*t*: 太郎は、握力がクラスで一番だった。 *h*: 太郎は、クラスの誰よりも握力が強い。	7
状態の変化	割合や倍数で数の変化を表している事例	*t*: 梅干しの消費量は、２０年前の１、５倍だ。 *h*: 梅干しの消費量は増えた。	6
数え上げ	文中のある概念を数え上げる必要がある事例	*t*: 日本にはアジアカブトエビ、アメリカカブトエビ、ヨーロッパカブトエビというカブトエビが生息している。 *h*: 日本には３種類のカブトエビが生息している。	3
その他			15
計			116

必要がある。他の例としては「貴重な展示品十数万点が略奪された」ならば「略奪で壊滅的被害を受けた」と解釈する必要がある例などがある。後に述べるように、この推論のベースとなるのは、「水不足に陥る人数が 70 億人」「略奪された展示品の数が十数万点」が多いのか少ないのかという認識である。本稿ではこの大小の判定のみに焦点をおき、この問題の解決を図る。

- 「語彙知識」は「銀婚式」は「25 回目の結婚記念日」や「還暦」が「60 歳」、「バイリンガル」ならば「二カ国語を話せる」といった数量が関連する語彙の知識が必要になる問題である。

- 「算術処理」は例のように簡単な計算を行い文中の数量の間の関係性の正し

さを証明する必要がある例である。成澤 [9] ではこれをテンプレートベースの情報抽出としてアプローチできる可能性を示した。例えば、この例では「A の数量」「B の数量」「A と B の差」という 3 つのスロットを用意し、それぞれのスロットとして「4 億 4000 万」「4 億 5000 万」「+1000 万」であることを抽出できれば、スロット間の関係が正しいかどうか（「A の数量」ー「B の数量」＝「A と B の差」になっているか）を検証することで、数量の関係の正しさを証明できる。類似の手法は阿部ら [?] によって提案されている。阿部らは数量表現に関わる情報を正規表現により抽出し、抽出した情報を用いて「変化前」「変化量」「変化後」のいずれかの枠に数量表現を格納し、それらを用いて計算を行っている。小学校算数文章題を対象とした評価実験では 72 %の正解率を得ている。ただし算数文章題は語彙が限られているため、これがそのまま今回の事例のような複雑な文に対応はできないと考えられる。

本稿で解決を目指すのは「数量表現間の含意」「数量の解釈」の問題である。これは事例数で考えると 1,2 番目に大きいカテゴリとなり、全体の問題の 37.9 %にあたる。この数字を小さいと思われる読者もいるかもしれないが、含意関係認識においては実に様々な種類の言語現象に対応する必要があり、1 つの手法でその様々な言語現象に対応するのは非常に難しいということを述べておく。我々のこの研究は数量表現処理の第一歩となると我々は考えている。

4　数量表現・時間表現のアノテーション仕様

課題分析により明らかになった「数量表現間の含意」の問題を解決するため、本章では数量表現を抽出・規格化する手法について述べる。「数量表現間の含意」における問題は、例えば「約220万人」ならば「80万人以上」である、ということを認識する必要がある問題である。また他にも数の表記の揺れの吸収（例えば「1989」「１，９８９」「一九八九」「千九百八十九」）や、単位の統一（例えば「キログラム」「mg」「トン」を全て「g」という単位に統一）なども行う必要がある。我々は数量表現を計算機に理解できる一定の形式に変換することでこれを解決する。我々はこれを規格化と呼ぶ。本稿では数量表現と似た性質をもつ時間表現についても規格化の対象とする。これは含意関係認識における数量表現の課題とは関係ないが、非常に有益な処理であるということは分かって頂けるであろう。例えば「2012年1月1日」「2012/1/1」が同じ時刻を表していることを認識すること、そもそもこれが時間表現だと認識することは、非常に重要である。

本章では抽出・規格化の手法について述べる前に、どのような形式で規格化するのか、またそもそも数量表現・時間表現の定義とは何なのかについて述べる。人間が数量表現・時間表現に対して規格化された情報をアノテーションをする際の仕様を明らかにするという形で、これを説明する。

4.1　数量表現のアノテーション仕様

本節では日本語数量表現に対するアノテーション仕様の概略を示す。数量表現のアノテーション仕様はこれまでに存在しないため、我々はTimeML (J. Pustejovsky et al. 2003b) における <TIMEX3> タグの仕様を参考に、<NUMEX> タグを提案する。<TIMEX3> は言語資源管理に関する国際標準ISO/TC 37/SC 43 において2009年に採用されたISO 24617-1(SemAF/Time)の基になっている。

4.1.1　タグ付けされる数量表現

本稿の研究対象である数量表現は「数と単位によって、個数や量、順番を表現するテキスト中の文字列」とする。例えば「3人」は数（3）と単位（人）で人数

という個数を示しているため、数量表現と言える。また「還暦」のように数と単位で示される数量表現（60 歳）と意味が等しい表現も、数量表現として扱う。

以下では関根らが定める「数値表現」との違いについて述べる。我々が定義する数量表現は、関根らが定義する数値表現にかなり近いが、関根らが数値表現を「数を含む表現」と定義し比較的広い範囲の表現を扱うのに対し、我々は数量表現を「単位と数によって、個数や量、順番を示す表現」と定義し、やや狭い範囲の表現を数量表現とみなしている。以下で数量表現と数値表現の違いを示す（「」で関根の定義による数値表現の分類を示す）。

数量表現に含まれるもの

- 「個数」　　10 個、3 人
- 「寸法表現」　120kcal、50km/h
- 「序数」　　第一回、一人目
- 「年齢」「頻度」「倍数表現」「割合表現」「ポイント」「金額表現」「順位表現」

数量表現に含まれないもの

- 「震度」　　震度 4、震度 5 弱
- 「学齢」　　一年生、中学三年
- 「株指標」　　２６、5/6
- 「緯度経度」　　北緯 30 度
- 「寸法表現　その他」の一部　　A4
- 「数値表現　その他」　　2LDK

例えば「震度3」は表現の中に数を含むが、この数はある単位に対しての値ではなく、定められた階級を表すための数である。「震度14」「震度15」のような使い方はできないということからもこれは明らかである。また「A4」についても震度とほぼ同じことが言える。「A4」における「4」という数は、定められた規格を表すための数である。階級や規格が数を用いて表されている表現は、「3km」のような単位と数による表現と性質が異なるため、我々は「震度」「A4」を数量表現としてみなさない[3]。このように、我々は関根らの定義よりも表現の意味により踏み込んだ厳密な定義を用いて数量表現を定めた。「震度3」「A4」以外の例については付録にて詳しく述べているので参照されたい。

また数量表現に「約」「およそ」「以上」「弱」といった数の範囲を変化させる語句がともに現れた場合、TIMEX3の仕様に沿って、これらもひとまとめにして数量表現とみなす。MUC6では基本的にこのような語句は無視しているが、数量表現が示す数量を変化させるこれらの語句を無視するのは不自然であると考える。関根らは一部の語句のみを含めており、含める表現と含めない表現の差は不明確であった。

4.1.2 <NUMEX>タグ

アノテーションには<NUMEX>タグを用いる。<NUMEX>タグは@nid, @value, @counter, @mod, @rangeStart, @rangeEnd, @ordinalという属性をもつ。以下にタグ付け例を示す。

(4) a. <NUMEX nid="n0" value="1000" counter="g">1キログラム</NUMEX>

b. <NUMEX nid="n0" value="100" counter="冊" mod="APPROX">約百冊</NUMEX>

[3]ただし、震度と違い「A4」は「210mm x 297mm」という、具体的な量（紙の寸法）を示しており、単位と数による表現を示していると言えなくもない。すなわち「還暦」の例と同じ意味で、「A4」は数量表現とみなせなくもない。しかし「210mm x 297mm」という2つの値と単位を持つ表現を1つの数量表現としてみなしていいのかは微妙な問題である。よって、ひとまず我々は「A4」を数量表現とはみなさない。

表 2: @mod 属性に対する値（NUMEX タグ、持続時間表現）

値	定義	例
@mod=APPROX	近似表現	「くらい」「約」
@mod=KYOU	近似表現	「強」
@mod=JAKU	近似表現	「弱」
@mod=JUST	完全一致	「ちょうど」
@mod=EQUAL_OR_LESS	数量表現の範囲以下	「以内」
@mod=EQUAL_OR_MORE	数量表現の範囲以上	「以上」
@mod=LESS_THAN	数量表現の範囲未満	「未満」「近く」
@mod=MORE_THAN	数量表現の範囲超過	「余り」「過ぎ」

@nid 属性は 1 文書中の各数量表現に付与される識別子である。

@value は規格化を行った値を、@counter は規格化された単位や助数詞を付与する。規格化された単位とは、SI 単位系の場合は SI 接頭辞を除いた SI 単位のみを、それ以外の助数詞ではより一般的と考えられる助数詞とする。例えば「頁」は「ページ」という助数詞に規格化される。また「トン」のような単位は「グラム」で規格化する。何が規格化された単位になるかは厳密には定められおらず、これを定めるのは今後の課題である。

@mod 属性は数量表現のモダリティを表す。例えば「20 人以下」をタグ付けする際には@mod 属性に EQUAL_OR_LESS という値を割り当てる。属性に割り当てる値は TIMEX3 タグにほぼ等しい。値の詳細を表 2 に示す。NUMEX 用に新規で追加した値は JUST, KYOU, JAKU のみである。

@rangeStart, @rangeEnd 属性は範囲表現を表すためのタグである。TIMEX3 では範囲表現を扱っていないが、我々は範囲表現を扱う。例えば以下のように @rangeStart, @rangeEnd を付与する。

(5) 10〜15 個

<NUMEX value=”10” counter=”個” rangeStart=”true”>10</NUMEX>
〜<NUMEX value=”15” counter=”個” rangeEnd=”true”>15 個 <NUMEX>

@ordinal 属性は序数を示す。序数についての詳細は付録に記す。

4.2 時間表現のアノテーション仕様

日本語時間表現に対するタグ付け基準は小西らが TIMEX3 に基づいた基準を提案している。本稿でも一部の違いを除き、これに沿ったタグ付けを行う。始めに小西らが提案するタグ付け基準について触れた後、仕様が異なる部分を述べる。

4.2.1 タグ付けされる時間表現

本稿の研究対象である時間表現は時間軸上の時点もしくは時間の量を表現するテキスト中の文字列とする。時間表現は以下の 4 つの分類に分けられる。日付表現 (“DATE”, 相当)・時刻表現 (“TIME”相当) は「2008 年 4 月」「昨日」「今朝 9 時」といった、時点及び時区間の時間軸上の位置を定義することを目的として用いられる表現である。日付表現と時刻表源の違いは時間軸上の粒度の違いのみである。持続時間表現 (“DURATION”相当) は「1 時間」といった、時間軸上の位置に焦点をあてずに時間の量を定義することを目的として用いられる表現である[4]。持続時間表現は時間の量を表す数量表現と言える（ただし、我々はこれを数量表現ではなく時間表現として扱う)。頻度集合表現 (“SET”相当) は「週に 3 回」といった、時間軸上複数の時区間を定義することを目的として用いられる表現である。

4.2.2 <TIMEX3> タグ

アノテーションには <TIMEX3> タグを用いる。<TIMEX3> タグは@tid, @type, @value, @valueFromSurface, @temporalFunction, @freq, @quant, @mod, @rangeStart, @rangeEnd, @ordinal を持つ。以下にタグ付け例を示す。

(6) a. <TIMEX3 tid=”t1” type=”DATE” value=”2003-10-20” valueFromSurface=”2003-10-20”>2003 年 10 月 20 日 </TIMEX3> <TIMEX3 tid=”t2” type=”DATE” value=”2003-W43-1” valueFromSurface=”XXXX-WXXX1”> 月曜日 </TIMEX3>

[4] 小西らはこれを「時間表現」と呼び、「時区間幅を定義することを目的として用いられる表現」として定義した。我々はこれを「持続時間表現」と呼び、「時間の量を表す表現」として定義する。

表 3: 日付・時刻表現に対する@value

単位	記号	表現例	@value
年月日	XXXX-XX-XX	1980 年 7 月 7 日	1980-07-07
曜日	XXXX-WXX-X	水曜日	XXXX-WXX-3
季節	XXXX-SP, SU, FA, WI	冬	XXXX-WI
四半期	XXXX-QX	第一四半期	XXXX-Q1
年度	FYXXXX	1998 年度	FY1998
世紀	XXXX	11 世紀	10XX
紀元前	BCXXXX	紀元前 202 年	BC0202
時刻	XXXX-XX-XXTXX:XX:XX	２００６年８月８日午前８時４５分３０秒	2006-08-08T08:45:30
時刻（略記）	TXX:XX:XX	午前８時４５分３０秒	T08:45:30

表 4: 持続時間表現に対する@value

単位	記号	表現例	@value
年	PnY	3 年間	P3Y
月	PnM	2ヶ月	P2M
日	PnD	5 日	P5D
時間	PTnH	3 時間	PT3H
分	PTnM	30 分	PT30M
秒	PTnS	9 秒 80	PT9.80S
週	PnW	1 週間	P1W

@tid 属性は 1 文書中の各時間表現に付与される識別子である。

@type 属性は DATE, TIME, DURATION, SET の 4 つの値を持つ。それぞれ日付表現・時刻表現・持続時間表現・頻度集合表現を意味する。

@value 及び @valueFromSurface 属性は時間表現が含意する日付･時刻･含意の値を表す。値として ISO-8601 形式を自然言語表現向けに拡張したものを用いる。このうち@value は文脈情報を用いて正規化を行った値を付与し、@valueFromSurface 属性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。文脈情報を用いた正規化とは、例えば「昨日」に対して、文脈情報から今日が2012 年 8 月 2 日であることが分かる際に value=”2012-08-01” を付与することを指す。@valueFromSurface は小西らがオリジナルに定めた属性で、これに付与される値の形式について、小西らの論文では明示されていなかった。これについて次の節で詳しく述べる。@value 属性にわりあてる値の詳細を表 3, 4 に示す。

表 5: @mod 属性に対する値（日付・時刻表現）

値	定義	例
@mod=START	日付時刻表現の初期	「始め」「初頭」
@mod=MID	日付時刻表現の中期	「半ば」「中ごろ」
@mod=END	日付時刻表現の後期	「末」「暮れ」
@mod=APPROX	近似表現	「ごろ」
@mod=BEFORE	日付時刻表現より前	「前」
@mod=AFTER	日付時刻表現より後	「過ぎ」
@mod=ON_OR_BEFORE	日付時刻表現以前	「以前」
@mod=ON_OR_AFTER	日付時刻表現以後	「以降」「以来」

@temporalFunction 属性は true, false のいずれかの値を持ち、@valueFromSurface が文脈情報により曖昧性解消可能か否かを表す。定時間情報が得られる不定時間情報表現は true の値を持ち、その他の時間情報表現は false の値を持つ。

@mod 属性は時間情報表現のモダリティを表す。例えば「2000 年以前」をタグづけするために@mod 属性に ON_OR_BEFORE という値をわりあてることにより「以前」というモダリティを表現する。属性にわりあてる値の詳細を表 5 に示す（持続時間表現に付与される値は <NUMEX> タグの mod 属性と同じなので省略してある）。

@freq, @quant 属性は頻度集合表現に付与される頻度情報及び量化子情報である。頻度集合表現は @value, @freq, @quant 属性を組み合わせることにより複雑な時間情報を表現する。頻度情報を表すためには、期間を表す@value 属性とともに、@freq 属性に nX をわりあてることにより、焦点をあてている期間中に事象が n 回起こることを示す。例えば「週に２回」を表現する際には <TIMEX3 type=”SET” value=”P1W” freq=”2X”> 週に２回 </TIMEX3> のようにタグづけする。@quant 属性には「毎日」「毎週」「毎１０月」といった表現に EACH をわりあて、「１０日おき」「３日毎」といった表現に EVERY をわりあてる。この際@value 属性には期間を表す値だけでなく、日付・時刻を表す値が入ることがある。以下に例を示す。

<TIMEX3 type=”SET” value=”P1D” quant=”EACH”> 毎日 </TIMEX3>

<TIMEX3 type=”SET” value=”XXXX-10” quant=”EACH”> 毎１０月
</TIMEX3>
<TIMEX3 type=”SET” value=”P10D” quant=”EVERY”> １０日おき
</TIMEX3>

@rangeStart, @rangeEnd, @ordinal については <NUMEX> タグに等しい。オリジナルの <TIMEX3> タグでは@beginPoint, @endPoint という属性を付与していたが、<NUMEX> タグの属性と統一するため、これを@rangeStart, @rangeEndとした。

4.2.3 小西らとの仕様の差異

- @valueFromSurface 属性について

  日付・時刻表現において、表層の情報だけで正規化ができる表現と、文脈の情報を用いなければ正規化ができない表現がある。前者を定時間情報表現 (fully-specified temporal expression) と呼び、後者を不定時間情報表現 (underspecified temporalexpression) と呼ぶ。定時間表現において @value 属性の値と @valueFromSurface の値は等しい。不定時間表現に valueFromSurface を付与する際、小西らは「3 年後」に対しては valueFromSurface=”P3Y”、「先月 3 日」に対しては valueFromSurface=”XXXX-XX-03” を付与していた。前者では「後」という意味が失われ単なる時間の量のみを示しており、後者では「先月」という意味が丸々失われている。我々はこれらの情報を valueFromSurface 属性に含めるため、 valueFromSurface を「日付/時刻, 時間量」という形式で記述する。例えば「3 年後」「先月 3 日」はそれぞれ valueFromSurface=”XXXX-XX-XX,P3Y”、valueFromSurface=”XXXX-XX-03,P-3M” を付与する。ただし時間量が 0 の場合（「8 月 3 日」のように年の単位が不明で不定時間情報表現となる場合）は時間量は記述されず、日付/時刻が記述される。

  我々がこのような形式を提案するのは、表層から得られる情報の正規化と、文脈情報を用いた正規化は明確に分けられるべき処理であり、文脈情報を用

いて正規化する場合は正規化された表層から得られる情報（すなわち@valueFromSurface 属性）のみを参照して正規化を行うべきだと考えるためである。我々が提案する形式では、文脈から与えられる時刻/日付さえ分かれば、@valueFromSurface 属性を用いて容易に@value 属性を求めることができる。

上のような例以外に小西らは「ゲーム参加者が多い時間は 11 時。夜のね」の「11 時」の@valueFromSurface に T23:XX:XX を付与していたが、我々は T11:XX:XX を付与する。「11 時」が「午後 11 時」を示していることがわかるのは、周囲の文脈によるものである。

- 「今日 3 月 3 日」「3 月 3 日 (水)」について

  小西らは「今日 3 月 3 日」に対して、「今日」と「3 月 3 日」を別々の時間表現としてタグ付けしていた。これは 1 つの時間表現として考えるのが自然だと考える。「3 月 3 日 (水)」についても同様である。

- 年号について

  小西らは@valueFromSurface 属性を「文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する」と定義しているが、年号のような文字列の表層表現のみから十分に判定できる表現に対して特殊なタグを付与していた。例えば「大正 9 年」には valueFromSurface=”T09”が付与される。年号は多々あるため、このようなアノテートの仕方では全ての年号に対応できないと考え我々は直接 valueFromSurface=”1920”を付与する。

- その他の差異

  - 「ゲーム参加者が多い時間は 11 時。」の「11 時」に valueFromSurface=”T23:XX:XX” が付与されていたが、文脈を考慮しなければ valueFromSurface=”T11:XX:XX” が正しい。
  - 「3 歳」に valueFromSurface=”P3Y” が付与されていたが、「歳」は年齢を示す数量表現であり、時間表現ではない。

– 「3 周年」に valueFromSurface=”P3Y” が付与されていたが、「周年」は年の経過を示す数量表現であり、時間表現ではない。

– 「3 年ぶり」の「ぶり」は、3 年の期間の間にあるイベントが起きなかったという意味を付与する表現であり、時間表現ではない。「3 年」のみをタグ付けする。

– mod 属性に JUST など何種類か新しい値を定義した。また、rangeStart, rangeEnd という範囲表現を扱う属性を定義した。

– 「1 万年」「1000 万年」を表現するのに”KA10”, ”MA10” といった表記は用いない。

5　数量表現・時間表現の抽出と規格化

本章では、文中の数量表現・時間表現を抽出、規格化する手法を提案する。本稿では 4 章で定義した数量表現と時間表現のうち一部の表現のみを対象とした規格化手法を提案する。本稿で対象としない表現は、表記に数を含まない表現である。例えば「20 歳」「1 月 1 日」は対象とするが「ハタチ」「元旦」は対象としない。これを対象としなかった理由は、数を含む表現の抽出・規格化と、数を含まない表現の抽出・規格化は処理が大きく異なるタスクであると考えたためである。後者も重要ではあるものの、本稿では自然言語中に表れる数の理解をすることに焦点をあて、数を含む表現のみを対象とした。数を含まない表現への対応は今後の課題である。また時間表現に関して、我々は@value 属性を付与せず、@valueFromSurface 属性のみを付与する。すなわち、表層から読み取れる分の正規化のみを行い、文脈を考慮して時間表現を正規化することは行わない。

提案手法の概要を述べる。我々は以下の 4 ステップで、抽出・規格化を行った。

1. 数の抽出、規格化：漢数字など様々な表記で表される数を、表記の揺れを吸収して認識し、数値型の変数に格納する。

2. 数量表現・時間表現の抽出、規格化：ルールと一致した文字列を文中から抽出し、抽出した文字列を規格化する。

3. 修飾表現の抽出、規格化： 抽出した数量表現・時間表現の前後が修飾表現であった場合、これを抽出し、規格化表現を修正する。

4. 入力文にタグを付与して出力する。

図 1 は提案手法の流れを表したものである。以下でそれぞれの処理について詳しく説明する。

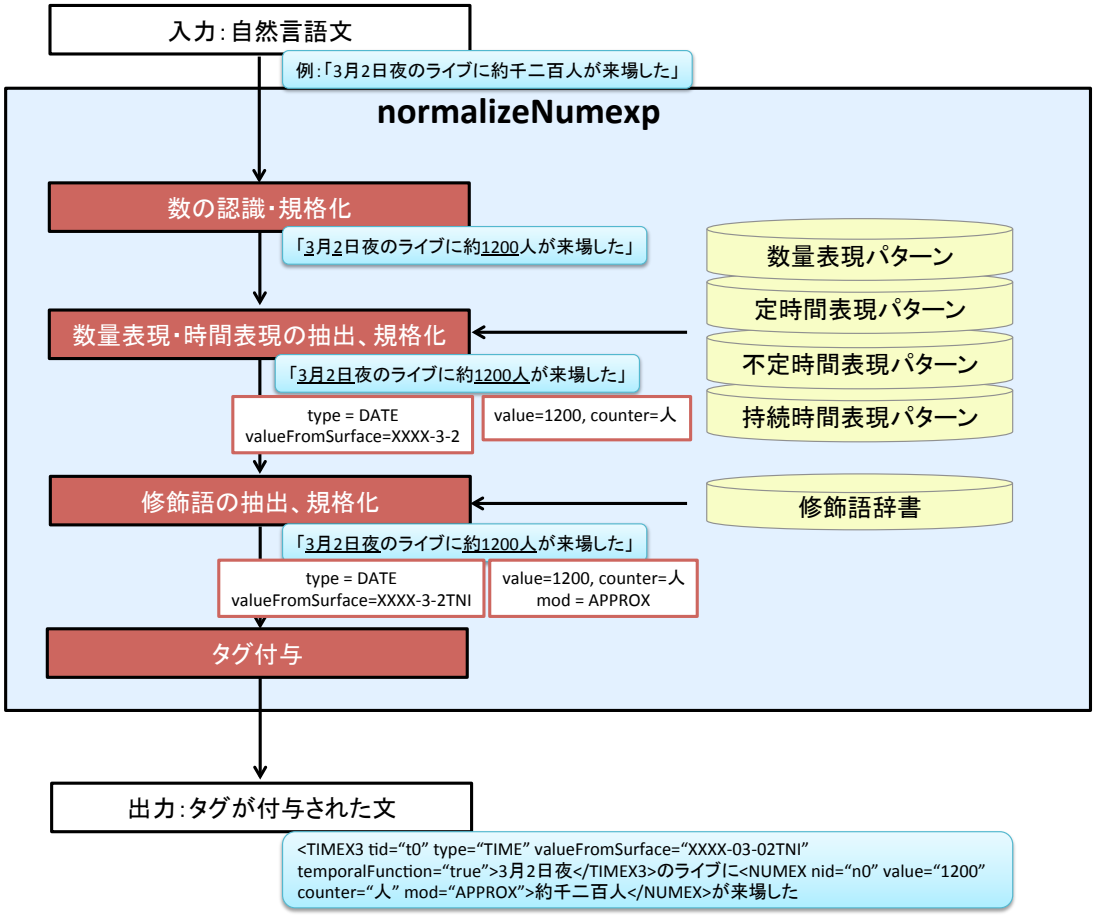

図 1: normalizeNumexp の構成

5.1　数の規格化

日本語では数が様々な表記で表される。例えば「１９８９」「1，９８９」「一九八九」「千九百八十九」などの様々な表記で「1989」という数値を表すことができる。規格化の前処理として、数の規格化では数を表す文字列を認識し表記の揺れを吸収して数値型の変数に格納する。具体的には、以下の処理を行う。

1. 数を表す表記の認識：入力文を先頭から順に見ていき、連続する数字を数を表す表記として認識する。ここでの数字とは、アラビア数字と漢数字のことである。

2. 規格化：認識した数を表記の方法に応じて規格化処理を行う。例えば、「1万5000」は半角数字と位を表す漢数字による表記法であり、これに適した処理は「万」以上のアラビア数字を10000倍し、それ以下のアラビア数字そのまま加算するという処理である。我々は以下の4つの表記法で表される数を処理する。

   - アラビア数字（位取り記数法）（例：12345）
   - 漢数字（位取り記数法）（例：一二三四五）
   - アラビア数字と位を表す漢数字（例：1万2345）
   - 漢数字（例：一万二千三百四十五）

   処理の際には全角数字も対象とする。カンマは基本的にアラビア数字の位取り記数法において、3桁毎に数を区切るために使われることが多いが、他の用法についても使われる例があったため（例えば「1万2,345」）これにも対応している。カンマの処理は次の「記号の処理」で行われる。

3. 記号の処理：規格化した数の前後の文字（列)、また規格化した2つの数の間の文字（列）が以下の条件を満たす場合、対応する処理を行う。

   - 数の前に「—」「マイナス」：数量の正負を反転させる。
   - 数の前に「＋」「プラス」：何もしない

- 数と数の間に「.（ピリオド）」「・」：これを小数点とみなし、前後2つの数を統合する。
- 数と数の間に「,（カンマ）」：カンマ区切り表記（例：1,234）とみなし、前後2つの数を統合する。カンマ区切り表記と考えられない場合（カンマ以降の数が3桁区切りでない場合。例えば、「2,3個のリンゴ」）は、範囲の表現とみなす。カンマ表記とも範囲の表現とも考えられない場合は処理を行わない。
- 数の前後、間に範囲を表す文字列（「〜」「-」「から」など）：範囲を表す文字列が数と数の間にあった場合（例：「500〜600」）、2つの数を統合してある範囲をもった1つの数とみなす。
- 数の前、もしくは数と数の間に「数」：「数十万」「十数万」といった表記へ対応するための処理を行う。例えば「数千」は「X000」に変換される。

記号の処理について、日付・時刻表現の場合は「2009.8.11」や「2009-8-11」のように、「.（ピリオド）」や「-（ハイフン）」などの記号が上記とは違った意味を持ってくる。このため、日付・時刻表現を規格化する際のみ、数の規格化処理で一部の記号の処理を行わない。

5.2 数量表現・時間表現の抽出、規格化

抽出の際は「数に該当する部分」「数以外の文字列」からなるパターンが記述された辞書を参照し、このパターンと最長一致する文字列が文章中から抽出される。辞書にはパターンとともに、そのパターンに最長一致した際に行う規格化処理の情報も記述されており、規格化の際にはこの情報を用いる。辞書はタイプ別に4つに分かれているため、分類毎にそれぞれの辞書の中身を述べる。

5.2.1 数量表現

数量表現についての辞書は、以下のような形式で記述される（全ての辞書はjson形式で記述する）。

```
{"pattern":"*キロメートル",
 "counter":"m",
 "SI_prefix":3,
 "optional_power_of_ten":0,
 "ordinal":false,
 "option":""}
```

patternと一致する文字列が文章中から抽出される。pattern中のアスタリスクは数に該当する部分を示している。例えば文中に「30キロメートル」という表現が出てきた際には、辞書中のパターンの中から上の「*キロメートル」というパターンが最長一致し表現が抽出される。パターンが一致すると、他の要素を参照して以下の処理が行われる。

- 抽出した文字列中の数を、そのまま@value属性の値に付与する。
- counterに記述された値を @counter属性の値に付与する。
- SI接頭辞による補正として、10のSI_prefixに記述された値乗を@valueに乗算する（例：「キロ」であれば、値に10の3乗を乗算する）。
- その他の補正として、10のoptional_power_of_tenに記述された値乗を@valueに乗算する（例：「*トン」というパターンでは、10の6乗を乗算する）。
- ordinalがtrueだった場合、 @ordinal属性の値をtrueとする。
- その他、特殊な処理が必要な場合は、optionの記述に従う。

5.2.2 定時間情報表現

定時間表現についての辞書は、以下のような形式で記述される。

```
{"pattern":"*年*月*日",
 "corresponding_time_position":["y", "m", "d"],
 "option":""}
```

規格化の際に行われる処理は以下である。

- 抽出した文字列中の数を、指定された時間単位にセットする。時間単位の指定は corresponding_time_position を参照して行われる。上の例では、初めの数字が年、次の数字が月、最後の数字が日として認識される。数を時間単位にセットした後、適切な値（DATE, TIME のどちらか）が @type に付与され、適切な ISO 形式の時間表現が @value に付与される。
- その他、特殊な処理が必要な場合は、option の記述に従う。

5.2.3 不定時間情報表現

不定時間表現についての辞書はほとんど定時間表現と同じだが、起点からの時間量を示すために corresponding_time_position に符号がついている。正の符号はその時刻より後ろであることを、負の符号はその時刻よりも前であることを示す。また「前」「後」といった文字列の前に「くらい」などの修飾表現がきてしまうため、この処理も行う。

```
{"pattern":"か月ぐらい後",
 "corresponding_time_position":["+m"],
 "process_type":["MOD_APPROX"],
 "option":""}
```

- 抽出した文字列中の数を、指定された通りの時間単位として設定する。ほぼ絶対時間表現の場合と同じだが、相対時間表現の場合は時間単位の指定に符号がつけられている。例えば「-y」「+y」のような形で指定され、「*年前」「*年後」といったパターンが該当する。

- prodess_type に記述された値に応じた処理を行う。詳しくは修飾表現の処理の節において述べる。

5.2.4 持続時間表現

```
{"pattern":"*か月半",
 "corresponding_time_position":["m"],
 "process_type":["han"],
 "option":""}
```

- 抽出した文字列中の数を、指定された通りの時間単位として設定する。
- prodess_type に記述された値に応じた処理を行う。詳しくは修飾表現の処理の節において述べる。

process_type や option によって、柔軟な処理の記述が行えるようになっている。現在の処理で対応できないような表現が表れた場合でも、process_type や option にある値が代入されていた場合のみ呼び出される処理をライブラリに記述することで、容易にシステムを拡張できる。現在、process_type に記述されいる値は各種の mod 属性を付与する各種の値と、han のみである。

5.3 修飾表現の処理

抽出した数量表現・時間表現の前後の文字列が修飾表現であった場合、これを数量表現・時間表現の抽出結果に加え、追加で規格化処理を行う。前節の処理に同じく、修飾表現となる文字列と一致した場合の処理を辞書に記述し、処理を行う。処理の大半は@mod 属性に値を付与する処理である。主な処理を表 6 に示す。

表 6: 修飾表現の処理一覧

処理名	定義	例
MOD_***	@mod 属性に値を付与する。MOD_APPROX や MOD_ON_OR_AFTER などが該当。	約、およそ、以降
SP, SU, FA, WI	@value に季節の情報を付与	春、夏、秋、冬
DN, MO, MI, DT など	@value に時刻の情報を付与	未明、朝、昼、日中
1J, 2J, 3J	@value に日付の情報を付与	上旬、中旬、下旬
kara_suffix	範囲表現の処理を行う。この表現と直後の表現が連続しており、タグ名、@type、 @counter の値が等しい場合は @rangeStart を付与し、直後の表現にも@rangeEnd を付与する。情報の伝搬も行う（「3 月 5 日〜6 日」だと、後者の「6 日」という日付表現の@value に月の情報が入っていないため、これを伝搬する）。	〜、 、-
dai	@value の下 n 桁を X にする（n はもともとの値の 0 の数で決まる。「1990 年代」の@value は”199X”となる）	台、代
han	数量表現の場合は 0.5 を@value の値に追加する。時間表現の場合は適切な値を@value の値に追加する（「12 時半」であれば、30 を「分」にセットする）	半

5.4 その他の処理

最後に不適切な文字列の削除を行った後、入力文にタグを付与して出力する。提案手法では文脈を考慮せず表現を抽出してしまうため、例えば「一体どうしたんだ」の「一体」も数量表現として認識してしまう。これを防ぐためこのような表現は辞書に記しておき、辞書の表現と一致した表現はこのステップで削除する。これにより「一体倒した」という文における「一体」を抽出できなくなるという問題は起こるが、「一体」が「一体倒した」の文脈で表れるのは稀であり、前述のような例で表れることの方が圧倒的に多い。

5.5 評価実験

小西らが作成した時間表現コーパスと、我々が作成した数量表現コーパスを用いて提案システムの評価を行った。提案システムの仕様から、以下の条件下で評価実験を行う。

1. 数を含まない数量表現・時間表現は評価対象としない

2. 時間表現においてvalue属性は評価対象とせず、valueFromSurface属性を評価対象とする（文脈を考慮した規格化を行わない）。

また4.3節で述べた通り小西らが作成した時間表現コーパスのタグ付け仕様と我々が提案するタグ付け仕様が若干異なるため、小西らが作成した時間表現コーパスを用いて評価を行う際は、以下の変換処理を行った。

- コーパス中のタグの変換処理
  - 「歳」「周年」「ぶり」を含む表現のタグは削除する
  - valueFromSurfaceがH,S,Tから始まる年表記の場合（「平成」「昭和」「大正」の年号が含まれる表現の際）、valueの値をvalueFromSurfaceに代入する。
  - 「01年（value = XX01)」などの表現の場合、valueの値をvalueFromSurfaceに代入する。
  - KA,MAは通常の年数表記に変換する（valueFromSurface = ”KA100”の場合、valueFromSurface = ”P100000Y”に変換する）。
  - 「value = ”XXXX-XX-XXT10” 」（10時、など）の際は、通常の時刻表記（value = ”T10:XX:XX”）に変換する
- システム出力のタグの変換処理
  - NUMEXタグは考慮しない
  - 「10年後」「明日3時」はそれぞれ「type=”DATE ” valueFromSurface=”XXXX,P10Y”」「type=”TIME” valueFromSurface=”T03:XX:XX,P1D”」のように「時刻,時間」の形で出力される。時刻が空の場合は時刻を削除しDURATIONとし[5]、時刻が空でない場合は時間を削除しDATE/TIMEとする。

[5]小西らのコーパスでは「10年後」のtypeがDURATIONであったりDATEであったり一定していなかった。

以上の処理を行ってもなお仕様上の違いの問題は残るが、それらは単に負例として扱う。例えば「ゲーム参加者が多い時間は 11 時。(11 時が午後 11 時を示すことが分かる文脈において)」の「11 時」の@valueFromSurface に小西らは T23:XX:XX を付与していたが、我々は T11:XX:XX を付与する。

また我々は新聞記事を対象とした NAIST テキストコーパスからランダムに 1000 文を抽出し、これに対して数量表現のアノテーションを行った。ただし、1 文の単位は読点を 1 つ含む文とし、少なくとも 1 つ以上の数（半角数字、全角数字、漢数字）を含む文のみを抽出した。これは、今回の対象とする数量表現は必ず数を含むためである。

評価は 1 つ 1 つの数量・時間表現を単位として、抽出と規格化それぞれで評価を行った。規格化の評価は抽出が成功したもののみで行い、全ての属性が完全に一致した場合のみ正解とした。ただし我々が属性として定めていない@definite, 小西らが定めていない@rangeStart, @rangeEnd は一致を問わない。評価実験の結果は以下のようになった。

Test set	表現の総数	P（抽出）	R（抽出）	F1（抽出）	Acc（規格化）
時間表現	3214	0.69(2002/2898)	0.62(2002/3214)	0.66	0.77 (1550/2002)
数量表現	769	0.92(713/777)	0.93(713/769)	0.92	0.99 (706/713)

数量表現抽出における誤りのうち半数程度が、文脈を考慮する必要がある事例であった（全体の 48%)。我々の提案手法は局所的にしか文をみないため、例えば「五輪に向けて調整を行う」における「五輪」を数量表現として認識してしまう。「六本木」「九段下」「八戸」といった地名を誤って認識する例も多かった。この問題に対応するためには、固有表現抽出における一般的な手法を用いるなどして、周囲の文脈を認識する必要がある。提案手法でも著しく精度を下げる文字列については、そもそも抽出しないという処理を行いこの問題に対処していたが、これは根本的な解決策ではないため、改善が必要である。残りの誤りは、主に単位表現辞書の不足によるものだった（全体の 36%)。今回用いた辞書は人手で整備したものであるが、数えられるほとんどの名詞は数量表現になりうる（「10 支店」「10 案件」「10 政党」など、数字 + 名詞で数量表現となる）ため、こういっ

た単位を自動獲得する必要があると考える。補足として、このように数量+名詞で構成される数量表現は、数量+助数詞で構成される数量表現（例えば「10 人」「10 個」）とは性質が異なる事を述べておく。例えば数量+助数詞の場合は名詞や動詞を修飾することができるが、数量+名詞の場合は修飾することができない。

数量表現の規格化はほとんどが適切におこなわれた。提案手法は一致した文字列が持つ json 形式の規格化情報を参照して規格化を行うため、文字列が一致さえすれば規格化はほとんど成功する。文字列が一致しているのにも関わらず誤った例は、「10 キロ」のように文脈の考慮が必要な事例である。例えば「10 キロ太った」と「この道は 10 キロ続く」「10 キロで走った」はそれぞれ単位が異なる。文脈を考慮した規格化が必要となる。

時間表現抽出における主な誤りは、辞書知識の不足によるものであった（全体の 40%）。特に小西らのコーパスでは「８７（昭和６２）年」といった括弧付きの表現が頻出し、これに対応するパターンが辞書中になかったため精度が大幅に下がった。また小西らのコーパスには「（イベント企画のページなどにおける）国内５　２０」「２００３」といった、時間表現であることを示す手がかりとなる文字列が存在しない、ただの数字のみからなる時間表現も含まれていた（誤り全体の 19%）。こういった時間表現に対応するためには、文脈を認識して抽出を行う必要がある。小西らと我々の仕様の差異や、小西らのコーパスのアノテーションのミスにより負例と扱われた事例は誤り全体の 15%であった。

時間表現の規格化における主な誤りは、数量表現と同じくやはり文脈の考慮が必要な事例であった。数量表現と比べて、時間表現は文脈を見る必要がある事例が多い。例えば「一日に会おう」の「一日」はある日付を示していると考えられるが、「一日を無為に過ごした」の「一日」は時間の量を示している。また負例全体の 31%は小西らと我々の仕様の差異や、小西らのコーパスのアノテーションのミスにより負例と扱われた事例であった。例えば、小西らのコーパスでは不定時間表現の TYPE を DATE にするのか DURATION にするのか一定していなかった。

6　数量の大小の自動判定

本章では、3 節における「数量の解釈」の問題に焦点をあてる。これは以下のような文対の含意関係を導くために必要な推論であった。

(7)　t：近い将来、最悪の場合 30 億人が水不足に直面する。

　　h：近い将来、世界は深刻な水不足になると懸念されている。

以上の例では、「30 億人が水不足」⇒「深刻な水不足」を導く必要がある。すなわち「30 億人」という数量を解釈する必要がある例である。含意関係を認識するための手法は様々考えられるが、この例に対して尤もらしい 推論の流れは以下であると我々は考える。

30 億人 が水不足に直面する
$\models$ たくさんの人 が水不足に直面する
$\models$ 深刻な水不足に直面する

我々が今回取り扱うのは、1 つ目の推論である「30 億人」⇒「たくさんの人」という推論である。すなわち、我々は数量の大小を判定するというタスクに取り組む。我々がこの推論に注目するのは、既に 3 節で述べたように、数量の解釈の根本は数量の大小の理解にあると考えられるためである。2 つ目の「たくさんの人」⇒「深刻」という推論を行うことは、今後の課題である。

より具体的にここで扱う数量の大小判定タスクについて説明する。大小判定タスクの入力と出力は以下のようになる。

- 入力：数量表現を含む 1 文と、ターゲットとなる数量表現
  - 例：「30 億人が水不足に直面する」※下線部の数量表現をターゲットとする
- 出力：ターゲットとなった数量表現の、その文中の文脈（またその文から推測される文脈）での大小。大きい、小さい、普通の 3 値。
  - 例：「大きい」

本章では数量の大小を自動判定する手法を提案する。本章では 2 つの手法を提案するが、どちらも Web から抽出した数量表現をもとに判定を行うため、まずは Web からの数量表現の抽出手法を述べた後、大小判定の手法を紹介する。最後に 2 つの提案手法の評価実験を行い、手法の有効性を論じる。

6.1 Web からの数量表現の抽出

本章では数量の大小を判定するために、Web 文書中に存在する数量表現とその周囲の文脈の情報を用いる。手法の詳細は次節で行うが、本稿で紹介する 2 つの大小判定手法は、入力文中の数量表現と同じ単位と文脈を持った数量表現を Web 文書から抽出し、それらの情報を用いて大小の判定を行う。

例えば「彼は身長が 190cm ある」というクエリに対しては、Web 文書から抽出された「cm」という単位と「彼の身長」という文脈をもつ数量表現（例えば「まだ中学生なのに彼は身長が 180cm もあって羨ましい」の「180cm」）を用いて大小の判定を行う。Web 上の分布に基づく手法では、数量の分布から判断して大小の判定を行い、大小の手がかり表現に基づく手法では「30 人も」「10 人しか」の「も」「しか」のような大小判定に関する手がかりを持つ数量表現を用いて判定を行う。大小判定手法の詳細は 6.2.1, 6.2.2 にて述べる。

本節では大小判定の手法について述べる前にそこで必要となる数量表現と文脈、そして手がかりを抽出する手法について述べる。

6.1.1 数量表現の抽出と規格化

最初のステップは、数量表現を抽出・規格化することである。これは前章で述べた抽出・規格化手法を用いて行われる。抽出・規格化手法については、既に述べたのでここでは省略する。ただし、今回の手法で利用する属性は value 属性と counter 属性のみとなる。

6.1.2 手がかりの抽出

次に大小の手がかり表現に基づく手法において非常に重要となる「その文章での数量の大小に関する話者の捉え方」の情報を抽出する。話者がある数量の大小に関してどんな捉え方をしているかという情報は、もし抽出できれば今回の大小判定タスクでは非常に有用な手がかりとなる。しかし、文全体の意味を考慮してこれを判断するのは非常に難しい。例えば「30 人の学生が来てくれて、あまり学生が来ないと思っていた私は大喜びだった」では、話者は 30 人は多いと思っていると推測できるが、これを認識するのは難しい。

本稿では、取り立て助詞などの表現を手がかりとして、文全体ではなく数量表現の周囲の表現だけを見て数量に対する話者の捉え方を抽出する手法を提案する。例えば以下のような手がかりがある。

- 取り立て助詞「も」　　例：「3 人も来た」
- 取り立て助詞「しか」　例：「3 人しか来ない」
- 「も」の名詞修飾形　　例：「3 人もの学生」
- 形容動詞「わずか」　　例：「わずか 3 人の学生」

数量表現が「も」「もの」を伴えば話者はその数量を「大きい」と捉えている、「しか」「わずか」を伴えば「小さい」と捉えていると考えられる。文全体を見るのに比べ、「も」「しか」といった表現を認識することは非常に簡単に行える。かつ、この大小に関する話者の捉え方の情報は非常に強力なものとなる。例えば「僕の部屋に友達が 10 人来た」の「10 人」の判定は、「僕の部屋に友達が 10 人**も**来た」という文を Web 文書中から見つけられれば、これをそのまま使って「大きい」と出力することが可能そうである。大小の手がかり表現に基づく手法ではこのような考えをもとに大小判定を行う。

本稿では、大小の手がかり表現に基づく手法において「も」「しか」の手がかりを用いて、大小の判定を行った。「もの」「わずか」を用いなかった理由は「も」「しか」に比べ、「もの」「わずか」という表現がほとんど文中に表れず、大小を判定するのに十分な数の数量表現を抽出できなかったためである。

また数量表現に「も」「しか」が伴うのは数量表現が副詞的に動詞を修飾しているとき（例「三人来た」）のみで、数量表現が名詞を修飾したり（例「三人の学生」）数量表現が主語になるとき（例「三人が来た」）はこの手がかりは使えない。このような理由から、本稿では大小判定を行う数量表現は動詞を修飾しているもののみとし、他の数量表現については本稿で対象としない。他の数量表現についても同じように大小判定を行うのは今後の課題である[6]。

抽出の際には、「も」を伴っている数量表現には large、「しか」には small というラベルを付与し、大小の手がかり表現に基づく手法ではこの 2 つのラベルどちらかを持った数量表現のみを用いて大小の判定を行う。

6.1.3 文脈の抽出

抽出した数量表現の文脈を抽出する。提案する 2 つの大小判定手法は、入力文中の数量表現と同じ単位と文脈を持った数量表現を Web 文書から抽出し、それらの情報を用いて大小の判定を行う。そのため、ここで抽出する文脈は数量表現が示している数量の対象を必要十分に表しているようなものが望ましい。例えば「そういえば彼は学校で後輩に 30000 円渡したと言っていたが本当だろうか」の文における「30000 円」の文脈としては、「そういえば」や「言っていたが〜」のようなあまり関係のない表現は無視して「ある人が学校で後輩に渡したお金の金額」などと抽出できれば望ましい。しかし、文脈を適切に設計することは決して容易な問題ではない。

本稿ではごくシンプルな要素で文脈を設計し、これを抽出する。我々が文脈として用いたのは数量表現が係る動詞の原形とその項である。動詞とその項だけで十分に文脈を表せるかどうかは疑問ではあるが、数量表現の文脈の抽出の第一歩として、このシンプルなルールを用いて、文脈を抽出する。前述の例では「動詞：渡す　ガ格：彼　二格：後輩　デ格：学校」を抽出する。我々は係り受け解析器

[6]Web 上の分布に基づく手法では「も」「しか」といった手がかりを必要としないため動詞修飾型でない数量表現についても判定を行えるが、本稿ではどちらの手法も動詞修飾型のみを対象とする。

CaboCha[7]と述語項構造解析器 KNP[8]を用いてこれを抽出した。前項で述べた通り、数量表現が動詞を修飾しない場合は対象としない。

6.2 数量の大小判定

本節では数量の大小判定を行うための 2 つの手法を提案する。まず初めに、2 つの手法に共通する箇所について説明する。どちらの手法も、前節で Web から抽出した数量表現と文脈の情報を用いて大小の判定を行う。入力が与えられると、まず以下の処理が行われる。

1. Web からの数量表現抽出と同じく、クエリの数量表現の規格化と文脈の抽出を行う[9]
2. クエリの数量表現の counter 属性と文脈が完全に一致するものを、Web から抽出しておいた数量表現から抽出する。

これらの処理を行った後、2 つの手法で異なる処理が行われ、最後に「大きい」「小さい」「普通」の 3 値のいずれかが出力される。

2 のステップにおいて一致するものがない場合、もしくは少ない場合は条件を緩和し、部分的に一致するものを抽出する。十分な量の数量表現が抽出されるまで、条件の緩和は繰り返される。条件の緩和は以下の順番で項の情報を無視していくことで行われる。ガ格まで無視しても一致する数量表現が見つからない場合は、動詞の情報も無視して、単位が一致する数量表現を抽出する。それでも抽出できなかった場合は「普通」を出力する。

- ヘ格、カラ格、ハ格、ヨリ格、マデ格、ニテ格、デ格、二格、ヲ格、ガ格

6.2.1 Web 上の分布に基づく手法

この手法は、Web 中で見られるある対象に関する数量の分布は、実際の分布に近いのではないかという仮定をもとに成り立っている。表 7 はこの仮定を説明す

[7] https://code.google.com/p/cabocha/
[8] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[9] 対象となる数量表現は既にマーキングされているので、抽出の必要はない。

表 7: 「単位：cm　動詞：ある　ガ格：身長」と一致する数量表現を含む文集合の例

文	cm	ラベル
身長が１４０センチしかありません	140	small
中３男子です。身長が１５４ｃｍしかありません	154	small
現在、高校一年生の男です身長は１６０ｃｍあります。	160	
不思議だったのは身長が１６１ｃｍもあった事だ	161	large
僕は今高２なのですが身長が１６２ｃｍしかありません	162	small
身長が１６３センチあって体重３９キロってありえますか	163	
私は身長が１６３センチしかありませんから少し大きな小学生でもいけそうです＾＾	163	small
私高校三年の１８歳男で、身長が１６５ｃｍあります	165	
彼女は身長が１６７センチもあり、私が「背が高くなりたいな〜と（略）	167	large
ヒールがなくても身長が１６８センチあるのに、このサンダルはいたら…	168	
身長が１７０ｃｍあってきれいな人でした。でもグランプリの人には負けてたかな	170	
また、身長が１８３ｃｍもあり、１６歳の女性キャラの中では１番背が高い。	183	large

るための表である。この表は実際の抽出例ではないが、実際もこの表のように「身長が 160cm〜170cm ある」のような表現は Web 中に頻繁に現れ、「身長が 140cm ある」「身長が 180cm ある」のような表現はあまり見られず、「身長が 100cm ある」「身長が 300cm ある」といった表現は存在しないと我々は予想した。Web 中に身長についての記述が十分に多ければ、身長についての分布は、およそ 160cm〜170cm(日本人の平均身長くらい）を中心とした正規分布のような形になると仮定し、この正規分布を用いて数量の大小判定を行う。

具体的な処理としては、前述の 1,2 に続き以下の処理が行われる。

3. 2 で抽出した数量表現の value 属性（規格化された値）の分布と対象の数量表現の value 属性の値を比較し、対象の値が分布の上位 5%にあれば「大きい」、下位 5%にあれば「小さい」、そうでなければ「普通」を出力する。

6.2.2　大小の手がかり表現に基づく手法

6.1.2 項で述べたように、「ある文章中での数量の大小に関する話者の捉え方」という情報を用いて大小の判定を行う。本手法では 6.1.2 項で付与した large/small ラベルよりも大きい/小さい値を、大きい/小さいとして出力する。例えば Web 文書中に「身長が 180cm もある」と書いてあった場合、「身長 180cm ならば高い」と言うことが言えるため、この手法では「190cm」というクエリに対して「大きい」を出力する。

ただしラベルは複数の数量表現に付与されており、表 7 にもあるように「身長が 161cm もあったことだ」「身長が 162cm しかありません」という矛盾した表現が表れることも考えられる。よって、ある値について大小の判定をしたいとき、その値について大部分の話者が「大きい（もしくは小さい)」と言っていれば「大きい（小さい)」とみなすことにする。これを定式化したのが以下である。

$$L(x) = \frac{p_l(x)}{p_s(x) + p_l(x)}, \tag{1}$$

$$p_l(x) = \frac{\left|\{r | r_v < x \wedge r_m \ni large\}\right|}{\left|\{r | r_m \ni large\}\right|}, \tag{2}$$

$$p_s(x) = \frac{\left|\{r | r_v > x \wedge r_m \ni small\}\right|}{\left|\{r | r_m \ni small\}\right|}. \tag{3}$$

$L(x)$ はある値 x（クエリの値）の大きさを示し、一定値以上のとき「大きい」一定値以下のとき「小さい」それ以外のときに「普通」を出力する。r は入力の数量表現の単位と文脈に一致する抽出された数量表現を、r_v と r_m は r の規格化された値と手がかりの有無をそれぞれ示す。式 2 の分子は、x という値を「大きい」と捉えている話者の数である。ここでは x 以下の値について「大きい」と捉えている話者は、x についても「大きい」と捉えるであろうという仮定をおき、x 以下の値のうち large ラベルを伴っている数量表現の数を、x という値を「大きい」と捉えている話者の数とみなしている。また分母は抽出された数量表現のうち large ラベルを伴うものの総数を示す。すなわち、$p_l(x)$ は何らかの値について「大きい」と捉えている話者のうち、x を「大きい」と捉えるだろう話者の割合を示している。逆に $p_l(x)$ は、何らかの値について「小さい」と捉えている話者のうち、x を「小さい」と捉えるだろう話者の割合を示している。以上より、$L(x)$ は Web

文書中で「も」「しか」を伴う数量表現のみを抽出したとして、全ての話者が「x は大きい」と捉えているとき 1、全ての話者が「x は小さい」と捉えているとき 0 を出力する。本手法では、$L(x) > 0.95$ 以上のとき「大きい」$L(x) < 0.05$ のとき「小さい」それ以外のときに「普通」を出力する。

6.3 評価実験

6.3.1 数量表現の抽出・規格化の精度について

大小判定手法の評価について述べる前に、ここでは抽出・規格化の精度が本手法に与える影響について議論しておく。本手法では Web から数量表現を抽出する際に抽出と規格化を、入力の大小判定を行う際に規格化を行う。数量表現の抽出・規格化の精度は既に述べた通り、それぞれ抽出の適合率が 0.92、再現率が 0.93、規格化の精度が 0.99 であった。

抽出については、適合率は高いにこしたことはないが低くてもあまり問題にはならないと考えられる。これは、仮に false positive があっても、その数量表現の文脈が他の true positive なものとは全く違うものになり、等しい文脈をもつ数量表現の情報を用いて大小判定を行う本手法では特に影響を及ぼさないためである。例えば「メアリーはその花を五輪買った」では、文脈は「動詞：買う　ヲ格：花を　ガ格：メアリーは」となるが、オリンピックの意味での「五輪」が同じ文脈で出てくるとは考えにくい。また、再現率が低いということは、Web から抽出される数量表現の総数が減るということであり望ましいことではないが、現在の精度であればこれが問題になることはないと考えられる。規格化の精度も 0.99 と十分に高かった。以上より、抽出・規格化ツールは、本手法において十分な精度があると考える。

6.3.2 評価用コーパスの作成

我々は Web 文書 [43] から動詞を修飾する数量表現を含む文をランダムに 2000 文抽出した。動詞を修飾しているかの解析には係り受け解析器 CaboCha を、抽出には既に述べた抽出・規格化手法を用いた。この 2000 文に対して、3 人のアノ

テーターがラベルの付与を行った。付与されるラベルは、この 2000 文を入力と考えたときの大小判定タスクの出力である。すなわち、以下の作業を行った。

- アノテート対象：数量表現（その数量表現を含む 1 文とともにアノテーターに渡される）
  - 例：「30 億人が水不足に直面する」
- 付与されるラベル：数量表現の、その文中の文脈（またその文から推測される文脈）での大小。大きい、小さい、普通、**やや大きい**、**やや小さい**、**判定不能**の **6** 値。
  - 例：アノテーター A「大きい」
  - アノテーター B「やや大きい」
  - アノテーター C「大きい」

もともとの大小判定タスクでは出力が 3 値であったが、人手でアノテートする際には 6 値でアノテートすることに注意されたい。「やや大きい」「やや小さい」というラベルを追加した理由は、「大きい（小さい）」とも「普通」ともどちらでも言える数量が文書中に頻繁に登場したためである。また「判定不能」は以下のような文にアノテートされる。

- 数量表現抽出、または係り受け解析のミス
- 数量の大小がその文の出てくる文脈に強く依存し判定しかねる場合。例えば「2000 円払った」という文は「何に払ったのか」という文中に書かれていない情報に「2000 円」の大小が強く依存するため、評価ができない。
- 意味不明な文章（対象とした文書は Web 文書なため、意味がわからない文章が多数存在する）

本来であればシステムもこの 6 値を出力するのが理想的ではあったが、大小判定に向けた研究の一歩目として 6 値の出力を行うのは難しいと考えたため、まずは 3 値から始めることにした。今回出力対象としなかった 3 つのラベルの評価実

一致した人数	数量表現の数
3 人	735 (36.7%)
2 人	963 (48.2%)
一致せず	302 (15.1%)
計	2000 (100.0%)

表 8: 被験者間一致率

験での扱いは、「判定不能」については対象とせず、「やや大きい」「やや小さい」については使用しない評価基準と工夫して使用する評価基準を設けた。

表 8 は被験者間一致率を示している。我々は少なくとも 2 人以上が同じラベルを付与した文のみを用いて評価用コーパスを作成した。ただし、2 人以上が「判定不能」をアノテートしたものは除いた。評価用コーパスは 640 文（小さい：20 文、やや小さい：35 文、普通：152 文、やや大きい：263 文、大きい：170 文）となった。

6.3.3 実験結果

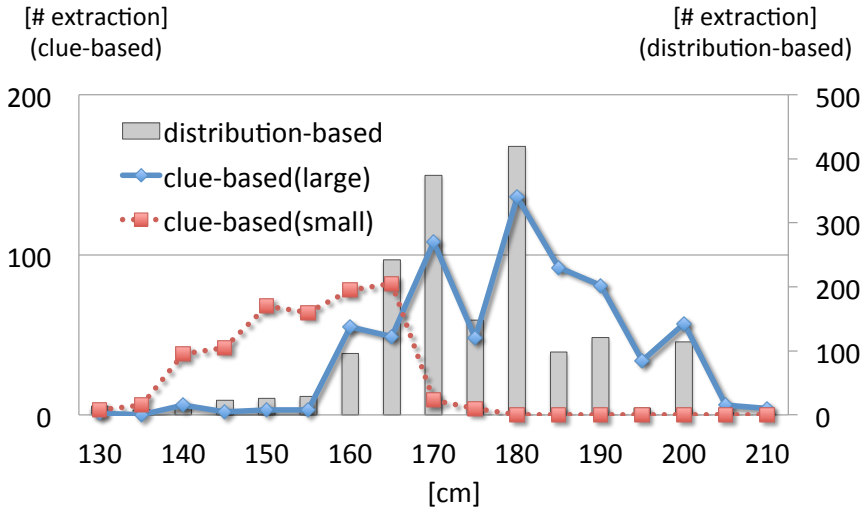

図 2: 「ガ格：身長　動詞：ある」という文脈での数量の分布

手法	ラベル	P	R	F1	Acc
Distribution	大きい+	0.892	0.498	0.695	0.760
	普通+	0.753	0.935	0.844	
	小さい+	0.273	0.250	0.262	
Distribution	大きい	0.861	0.365	0.613	0.590
	普通	0.529	0.908	0.719	
	小さい	0.222	0.100	0.161	
Clue	大きい+	0.923	0.778	0.851	0.770
	普通+	0.814	0.765	0.790	
	小さい+	0.228	0.700	0.464	
Clue	大きい	0.896	0.659	0.778	0.620
	普通	0.593	0.586	0.590	
	小さい	0.164	0.550	0.357	

表 9: 提案手法の適合率 (Precision, P), 再現率 (Recall, R), F 値 (F1), 精度 (Acc)

表 10: 出力例と誤り分析

No.	システム	正解	文	分析
1	小さい	小さい	こんなの作れるのは世界中で 三人 いるかいないかでしょう。	正例
2	普通	普通	2 匹 猫を飼っています。	正例
3	大きい	大きい	今日は ３２℃以上 あった !?	正例
4	大きい	大きい	競馬で 1 億円 儲けた	正例
5	小さい	大きい	十数人 来たために８畳程の部屋はいっぱいになった。	文脈抽出の失敗。特に述語項構造解析で失敗した例。対象を「来る人数」と解析してしまった。「来る」のニ格として「8 畳程の部屋に」を抽出し、「8 畳程の部屋に来る人数」と解析する必要がある。
6	小さい	普通	しかし、来週の新作が「X－ＭＥＮ」とか、「バック・ダンサーズ」たらいうシロモノらしいので、ここはまとめて 2 本 観ておいた方がよかろう。	文脈抽出の失敗。特に述語項構造解析で失敗した例。対象を「観る本数」と解析してしまった。「観る」のヲ格として「映画を」を抽出し、「映画を見る本数」と捉える必要がある（ヲ格の抽出には前半部で映画の話をしていることを推論する必要があり難しい例)。
7	小さい	普通	ちょっと前に彼氏と別れた友達が 2 人 いるんだけど、今、そのコたちの恋愛進行がとても楽しみです。	文脈抽出の失敗。特に項情報の詳細が欠落して失敗した例。対象を「友達の人数」と解析してしまった。「友達」を修飾している節を考慮し「ちょっと前に彼氏と別れた友達の人数」として捉える必要。
8	小さい	普通	カルビを 一口 食べさせてもらって料理長Ｙさんに感謝しました (笑)	文脈抽出の失敗。特に状況推定の必要がある事例。 対象を「カルビを口にした回数」と解析し、知識中に「この間の焼き肉ではカルビを一口しか食べれなかった」という文章から獲得した知識（一口＝小さい）があったため、判定を誤った。すなわち「(味見するような状況での）カルビを口にした回数」と「(普通は複数回食べる状況での）カルビを口にした回数」を区別する必要がある。
9	小さい	大きい	この亀は頭が 2 つ あります。	知識不足。「亀の頭の数」についての知識を持っていなかった。「動物の頭の数」のように一般化できれば、正しく判定できる可能性がある。
10	普通	大きい	この会社は面接が ４回 もあったので、何度も面接の練習をしていただきました。	知識不足。対象を「この会社の面接の回数」と解析したが、知識中に一致するものが 1 つもなかった。しかし「面接の回数」の知識はあったため、対象をやや粗く捉えることで、正しく判定できる可能性がある。

約 1 億 Web 文書 [43]、80 億文から 23000 万数量表現を抽出し、これを用いて数量の大小判定を行った。約 9%の数量表現が「も」を、約 6%が「しか」を伴っていた。図 2 は「ガ格：身長　動詞：ある　単位：m」という文脈をもつ数量表現の分布を示している。この分布から、我々の仮定の妥当性がわかる。例えばこの図から、およそ 150cm 以下の身長の人が小さい、180cm 以上ならば大きいとみなしてよさそうだということが分かる[10]。

評価用データを使って提案手法を評価するに辺り、我々は厳しい評価尺度（strict と呼ぶ）と緩い評価尺度（lenient）の 2 つの評価基準を設けた。strict ではシステムの出力と評価用データのラベルが完全に一致した時のみを正解とし、システムの出力に含まれない「やや大きい」「やや小さい」というラベルはデータから除外する。lenient では評価用データの「やや大きい」というラベルに対しては、システムは「大きい」「普通」のどちらかを出力できていれば正解とする。「やや小さい」に対しては「小さい」「普通」のどちらかを出力できていれば正解とする。

表 9 が実験結果である。+が lenient、無印が strict である。lenient における大小の手がかり表現に基づく手法の F 値は「大きい」に対して 0.851, 「普通」に 0.790、「小さい」に対して 0.464 となった。大小判定の難しさを考慮すれば、非常に良い結果だったと言えよう。大小の手がかり表現に基づく手法は Web 上の分布に基づく手法よりもやや良い結果となった。特に「小さい」に対する判定が大小の手がかり表現に基づく手法は優れていた。一方 Web 上の分布に基づく手法は「普通」に対する判定が優れていた。

6.3.4　誤り分析

表 10 に大小の手がかり表現に基づく手法の出力例と分析をまとめた。この表の例と分析は大小の手がかり表現に基づく手法のものではあるが、Web 上の分布に基づく手法でも共通して同じことが言える。2 つの手法に共通して主な誤り事例は、文脈を上手く捉えられなかったか、知識が足りなかったかのどちらか（もしくはどちらも）によるものだった。特に文脈を適当に捉えるのはやはり重要で

[10]単位は規格化されるため実際の単位は m だが、説明のわかりやすさのため cm で表示している

あり難しい。例 7 では関係節を抽出する必要があった事例ではあったが、これを文脈として抽出すべきなのかどうかは自明ではなく、否定や量化のスコープの曖昧性などにも通じる問題である。この 2 つの問題は、仮に文脈を細かく捉えられても知識が足りなくなるという関係にあり、2 つ両方を解決する必要がある難しい問題である。知識不足として挙げた例の中には、一致する文脈が Web 文書中になかったために、条件の緩和を行い、その結果情報量が落ちて正しい判定ができなかったものも存在している。共通な誤りのうちで今後の課題をまとめると、まずは項構造解析を自体の精度を上げ（例 5,6)、また項を修飾している節も考慮し（例 7)、更になんとも言えない情報も文脈としてなんとか捉え (例 8)、文脈を十分に捉えた上で、知識不足に対応するため捉えた文脈を今度は一般化して使う（例 9,10）という処理が必要になる。

Web 上の分布に基づく手法に固有な問題について述べる。Web 上の分布に基づく手法では、Web 文書中の数量の分布が実際の数量の分布に沿っているという仮定をおいて大小判定を行っていた。前述の身長の例のように、実際の分布によく沿っている対象も多々あったものの、そうでない例もみられた。例えば「本店では◯◯を 200 種類以上取り揃えています」という文を考える。これは明らかに宣伝を目的とした文であり、同じ文脈で「3 種類取り揃えています」などとわざわざ少ない値を書くことはないだろう。よって Web 文書中には大きな値である「200 種類」のような数量表現が多数表れる。こうなってしまうと、Web 上の分布に基づく手法では「200 種類」を普通の値として捉えてしまい、判定を誤ってしまう。

7　おわりに

本稿では含意関係認識における数量表現の問題を解決することを目指し、3つの課題に取り組んだ。

まず初めに成澤 [9] で行った課題分析について、再度分析をし直したものを報告した。本稿の分析ではより推論の本質に近い部分で分類を行い、より多くの事例について推論カテゴリを付与することができた。またそれぞれの推論カテゴリに含まれる事例の数も明示し、問題の大きさを明らかにした。本稿における他の取り組みは、この分析をもとに大きな問題から順番に解決を目指したものである。

次に数量表現の規格化に関するアノテーション仕様を提案し、数量表現･時間表現を規格化する手法について述べた。数量表現のアノテーション仕様は <TIMEX3> に基づき <NUMEX> タグを提案した。規格化のための手法としては数の規格化、数量表現・時間表現の規格化、修飾語の規格化と言う3ステップからなるルールベースによる手法を提案した。今後の課題は主に3つある。1つ目は精度を向上させるための試みである。評価実験より主なエラーは文脈を考慮する必要がある事例であることが明らかになったため、今後はこれを考慮できるような辞書の記述方法や、または固有表現抽出における手法の適用を考えている。辞書知識の不足も大きな問題であるが、この解決のためには人手で地道に1つ1つパターンを追加していく必要があるため、パターンの追加をより効率的に行えるようしていきたい。2つ目は今回対象としなかった不定時間表現に対する規格化処理と、数を含まない表現の抽出・規格化を行うことである。不定時間表現に対する規格化処理は、単に文書作成時間を抽出し規格化する程度の処理は簡単に行えるが、文脈を深く捉えて規格化を行う必要がある場合は非常に難しい。3つ目は TimeML で行われているイベントと時間表現間の関係の付与や、数量表現が修飾する対象の同定といった更に高度な情報の付与を行って行く事を予定している。今後、自然言語処理技術が自然言語のより深い理解を目指して発展していく際に、数量表現・時間表現を正しく理解できるかどうかが、言語の深い理解に必要な基礎的なタスクとなるのは明らかである。また単純に数量表現・時間表現を認識する処理だけでも、様々な言語処理アプリケーションに有用であると考えられる。

最後に、数量の大小を判定する2つの手法を提案した。Web 上の分布に基づく

手法ではクエリ文中の数量表現と文脈が等しい Web 中の数量表現の数量の分布を見ることで大小の判定を行った。大小の手がかり表現に基づく手法では「も」「しか」のような話者の態度を表す手がかりを用いて判定を行った。評価実験の結果、提案手法は数量の大小をよく判定できていることが分かった。今後の課題は主に 3 つある。1 つ目はより良い文脈の抽出法を考えることである。2 つ目は今回提案した 2 つの手法を上手く組み合わせて大小判定を行うことである。最後に、数量の大小の解釈のみに留まらず、更に深い数量の解釈に取り組んでいくことである。

謝辞

本研究を進めるにあたり、多くの方々にご協力をいただきました。心より感謝の意を表します。

主指導教官である乾健太郎教授と岡崎直観准教授には、お忙しい中、研究活動全般にわたり温かいご指導、ご助言をいただきました。お二方の助言は常に本質をついたもので、難しい研究内容に惑わされ、問題の本質を見失ってしまった私に対して、何度も正しい方向を示して頂きました。自分一人で問題について考える時は、お二方なら今の自分の考えになんと言うか？を自問自答することが、大学院の二年間で習慣づきました。お二方から頂いた助言の数々が、自分が大学院の二年間で得た一番の財産であると考えています。心より感謝を申し上げます。またお二方には研究生活以外の面でも温かいご支援を頂きました。特に岡崎直観准教授には留学中の生活について親身に相談に乗って頂きました。本当に感謝しております。

渡邉陽太郎助教には日常的に研究の相談にのって頂くとともに、特に論文を添削して頂く際に大変お世話になりました。毎回丁寧に目を通して頂き、細かい表現まで添削して頂き、大変助かりました。深く感謝しております。また本研究を進めるにあたり、適切なご助言をくださいました松林優一郎研究特任助教、水野淳太研究員に感謝致します。福原裕一研究員、菅野美和技術補佐員には、コーパスの作成時にお世話になるとともに、それ以外にも日常生活中に日本語の難しさ、面白さを感じるきっかけを頂きました。山口健史研究員にはツールの公開時に手助けして頂きました。感謝致します。また、研究活動および大学生活を暖かく支えてくださいました、八巻智子秘書に深く感謝致します。

研究に関して貴重なアドバイスをくださり、研究生活を暖かく支えてくださった研究室の皆様、そして大学生活において貴重かつ有意義な時間を共に過ごしてくださった皆様に心より感謝致します。先輩方には研究に関して議論させて頂き、またそれ以外にも学生生活に関する多くの助言を頂きました。優秀な後輩達には多くの刺激を貰いました。同期とは多くの苦労を共にしました。皆様、本当にありがとうございました。

ご多忙の中、審査委員をお引受けくださいました、木下哲男教授、伊藤彰則教

授に深く感謝致します。本研究に直接関わりがあるわけではありませんが、修士1年時の夏のインターンシップでお世話になった株式会社 Preferred Infrastructure の方々には多くの刺激と助言を頂き、その後の研究生活の励みとなりました。感謝致します。また Manchester University の National Centre for Text Mining (NaCTeM) の方々には修士2年時の夏のインターンシップでお世話になりました。また、本研究では NTCIR-9 のデータと黒橋・河原研究室のデータを使わせて頂きました。深く感謝致します。

最後に、私の研究生活を様々な面で支えてくれた数多くの先輩、友人、知人、そして両親に心より感謝致します。

参考文献

[1] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 177–190, 2006.

[2] Luisa Bentivogli, Elena Cabrio, Ido Dagan, Danilo Giampiccolo, Medea Lo Leggio, and Bernardo Magnini. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 3542–3549, 2010.

[3] Mark Sammons, Vinod V.G. Vydiswaran, and Dan Roth. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1199–1208, 2010.

[4] Elena Cabrio and Bernardo Magnini. Towards component-based textual entailment. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pp. 320–324, 2011.

[5] Assaf Toledo, Sophia Katrenko, Stavroula Alexandropoulou, Heidi Klockmann, Asher Stern, Ido Dagan, and Yoad Winter. Semantic annotation for textual entailment recognition. In *Proceedings of the 11th Mexican International Conference on Artificial Intelligence*, MICAI '12, 2012.

[6] Peter LoBue and Alexander. Yates. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pp. 329–334, 2011.

[7] Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. Overview of

ntcir-9 rite: Recognizing inference in text. In *Proceeding of NTCIR-9 Workshop Meeting*, pp. 291–301, 2011.

[8] 小谷通隆, 柴田知秀, 中田貴之, 黒橋禎夫. 日本語 textual entailment のデータ構築と自動獲得した類義表現に基づく推論関係の認識. 言語処理学会 第 14 回年次大会 発表論文集, pp. 1140–1143, 2008.

[9] 成澤克麻. 数量表現を伴う文における含意関係認識の課題分析. 2011 年度卒業論文, 東北大学, 2012.

[10] Anton Bakalov, Ariel Fuxman, Partha Pratim Talukdar, and Soumen Chakrabarti. SCAD: collective discovery of attribute values. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 447–456, 2011.

[11] Marcus Fontoura, Ronny Lempel, Runping Qi, and Jason Zien. Inverted index support for numeric search. *Internet Mathematics*, Vol. 3, No. 2, pp. 153–185, 2006.

[12] Minoru Yoshida, Issei Sato, Hiroshi Nakagawa, and Akira Terada. Mining numbers in text using suffix arrays and clustering based on dirichlet process mixture models. *Advances in Knowledge Discovery and Data Mining*, pp. 230–237, 2010.

[13] Véronique Moriceau. Generating intelligent numerical answers in a question-answering system. In *Proceedings of the Fourth International Natural Language Generation Conference*, INLG '06, pp. 103–110, 2006.

[14] John M. Prager, Jennifer Chu-Carroll, Krzysztof Czuba, Christopher A. Welty, Abraham Ittycheriah, and Ruchi Mahindru. IBM's PIQUANT in TREC2003. In *TREC*, pp. 283–292, 2003.

[15] Jennifer Chu-Carroll, David A. Ferrucci, John M. Prager, and Christopher A. Welty. Hybridization in question answering systems. In *New Directions in Question Answering'03*, pp. 116–121, 2003.

[16] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, Vol. 38, No. 11, pp. 33–38, 1995.

[17] H. Ji, R. Grishman, H.T. Dang, K. Griffitt, and J. Ellis. The sixth pascal recognizing textual entailment challenge. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[18] C. Blake, W. Zheng, K. Painter, and W. Weyerhaeuser. The role of semantics in recognizing textual entailment. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[19] Adrian Iftene and Mihai-Alex Moruz. UAIC participation at RTE-6. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[20] H. Jia, X. Huang, T. Ma, X. Wan, and J. Xiao. Pkutm participation at tac 2010 rte and summarization track. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[21] P. Pakray, S. Pal, S. Poria, S. Bandyopadhyay, and A. Gelbukh. Ju_cse_tac: Textual entailment recognition system at tac rte-6. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[22] D. Majumdar and P. Bhattacharyya. Lexical based text entailment system for main task of rte6. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[23] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. The fifth pascal recognizing textual entailment challenge. In *Proceedings of TAC 2009 Workshop*, 2009.

[24] Yuta Tsuboi, Hiroshi Kanayama, Masaki Ohno, and Yuya Unno. Syntactic difference based approach for ntcir-9 rite task. In *Proceeding of NTCIR-9 Workshop Meeting*, pp. 404–411, 2011.

[25] Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa, Keita Nabeshima, and Kentaro Inui. Tu group at ntcir9-rite: Leveraging diverse lexical resources for recognizing textual entailment. In *Proceeding of NTCIR-9 Workshop Meeting*, pp. 418–421, 2011.

[26] 戸次大介. (日本語研究叢書 24) 日本語文法の形式理論 - 活用体系・統語構造・意味合成. くろしお出版, 3 2010.

[27] 飯田隆. 日本語形式意味論の試み——名詞句の意味論——. 科学研究費補助金研究成果報告書 『日本語と論理学』 所収, 2000.

[28] Sumiyo Nishiguchi. Quantifiers in japanese. *Logic, Language, and Computation*, pp. 153–164, 2009.

[29] Francis Bond. Determiners and number in english, contrasted with japanese, as exemplified in machine translation. *Unpublished doctoral dissertation, University of Brisbane, Queensland, Australia*, 2001.

[30] Grishman Ralph and Sundheim Beth. Message understanding conference-6: A brief history. In *Proceedings of COLING*, Vol. 96, pp. 466–471, 1996.

[31] Nancy Chinchor and Patricia Robinson. Muc-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*, 1997.

[32] Sekine Sekine and Isahara Hitoshi. Irex: Ir and ie evaluation project in japanese. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 1475–1480, 2000.

[33] Satoshi Sekine and Chikashi Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pp. 1977–1980, 2004.

[34] Tjong Kim Sang Erik F. and De Meulder Fien. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 142–147, 2003.

[35] Pustejovsky James, Castano Jose, Ingria Robert, Sauri Roser, Gaizauskas Robert., Setzer Andrea., Katz Graham, and Radev Dragomir. Timeml: Robust specification of event and temporal expressions in text. *New Directions in Question Answering*, Vol. 3, pp. 28–34, 2003.

[36] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, Vol. 2003, p. 40, 2003.

[37] Verhagen Marc, Gaizauskas Robert, Schilder Frank, Hepple Mark, Katz Graham, and Pustejovsky James. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 75–80, 2007.

[38] Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 57–62. Association for Computational Linguistics, 2010.

[39] Hector Llorens, Leon Derczynski, Robert Gaizauskas, and Estela Saquete. Timen: An open temporal expression normalisation resource. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 2012.

[40] Eiji Aramaki, Takeshi Imai, Kengo Miyo, and Kazuhiko Ohe. Uth: Svm-based semantic relation classification using physical sizes. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 464–467, 2007.

[41] Dmitry Davidov and Ari Rappoport. Extraction and approximation of numerical attributes from the web. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1308–1317, 2010.

[42] Michitaka Odani, Tomohide Shibata, Sadao Kurohashi, and Takayuki Nakata. Building data of japanese text entailment and recognition of inferencing relation based on automatic achieved similar expression. *In Proceeding of 14th Annual Meeting of the Association for "atural Language Processing*, pp. 1140–1143, 2008.

[43] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of Information Processing*, Vol. 20, No. 1, pp. 216–227, 2012.

付録

A　数量表現のアノテーション事例集

数量表現のアノテーション仕様についてより詳しく説明するため、説明とともにいくつかの事例を紹介する。時間表現に関しては、仕様は既存研究で示されているため、ここで新たに事例を紹介することは行わない。

A.1　一般的な数量表現

以下に数量表現とそのアノテーション例の表を記す。

表 11: 数量表現とそのアノテーション例

表現	アノテーション例
一万人	<NUMEX nid=”n0” value=”10000” counter=”人”> 一万人 </NUMEX>
３０企業	<NUMEX nid=”n0” value=”30” counter=”企業”>30 企業 </NUMEX>
千数百機関	<NUMEX nid=”n0” value=”1XXX” counter=”機関”> 千数百機関 </NUMEX>
二十三選挙区	<NUMEX nid=”n0” value=”23” counter=”選挙区”> 二十三選挙区 </NUMEX>
6 カ国	<NUMEX nid=”n0” value=”6” counter=”カ国”>6 カ国 </NUMEX>
二十六カ国	<NUMEX nid=”n0” value=”26” counter=”カ国”> 二十六カ国 </NUMEX>
約百冊	<NUMEX nid=”n0” value=”100” counter=”冊” mod=”APPROX”> 約百冊 </NUMEX>
３６０ ml	<NUMEX nid=”n0” value=”0.36” counter=”l”> ３６０ ml</NUMEX>
l キログラム	<NUMEX nid=”n0” value=”1000” counter=”g”> l キログラム </NUMEX>
6 トン	<NUMEX nid=”n0” value=”6e+06” counter=”g”>6 トン </NUMEX>
５０ km/h	<NUMEX nid=”n0” value=”50000” counter=”m/h”> ５０ km/h</NUMEX>
時速１００キロ	<NUMEX nid=”n0” value=”100000” counter=”m/h”> 時速１００キロ </NUMEX>
5 キログラム毎立方メートル	<NUMEX nid=”n0” value=”5000” counter=”g/m3”> 5 キログラム毎立方メートル </NUMEX>
摂氏 6 度	<NUMEX nid=”n0” value=”6” counter=”℃”> 摂氏 6 度 </NUMEX>
120kcal	¡NUMEX nid=”n0” value=”120000” counter=”cal”¿120kcal¡/NUMEX¿
￥１００	<NUMEX nid=”n0” value=”100” counter=”円”>￥１００ </NUMEX>
１００ポンド	<NUMEX nid=”n0” value=”100” counter=”ポンド”> １００ポンド </NUMEX>
七億五〇〇〇万ドル	<NUMEX nid=”n0” value=”7.5e+08” counter=”ドル”> 七億五〇〇〇万ドル </NUMEX>
3 割 5 分	<NUMEX nid=”n0” value=”35” counter=”%”> 3 割 5 分 </NUMEX>
０．５倍	<NUMEX nid=”n0” value=”0.5” counter=”倍”> ０．５倍 </NUMEX>
３０歳	<NUMEX nid=”n0” value=”30” counter=”歳”> ３０歳 </NUMEX>
3 ヶ月	<TIMEX3 tid=”t0” type=”DURATION” valueFromSurface=”P3M”> 3 ヶ月 </TIMEX3>
還暦	¡NUMEX nid=”n0” value=”61” counter=”歳”¿還暦¡/NUMEX¿

三人目	<NUMEX nid=”n0” value=”3” counter=”人” ordinal=”true”> 三 人 目 </NUMEX>
三度目	<NUMEX nid=”n0” value=”3” counter=”度” ordinal=”true”> 三 度 目 </NUMEX>
第三回	<NUMEX nid=”n0” value=”3” counter=”回” ordinal=”true”> 第 三 回 </NUMEX>
初代	<NUMEX nid=”n0” value=”1” counter=”代” ordinal=”true”> 初代 </NUMEX>
第三豊栄丸	第三豊栄丸
長女	長女
第二次世界大戦	<NUMEX nid=”n0” value=”2” counter=”次” ordinal=”true”> 第 二 次 </NUMEX> 世界大戦
受付は二階にある。	受付は <NUMEX nid=”n0” value=”2” counter=”階”> 二階 </NUMEX> にある。
地上 8 階地下 1 階建ての建物。	地 上 <NUMEX nid=”n0” value=”8” counter=”階”>8 階 </NUMEX> 地 下 <NUMEX nid=”n1” value=”1” counter=”階”>1 階 </NUMEX> 建ての建物。
成澤さん (20 歳) が	成澤さん (<NUMEX nid=”n0” value=”20” counter=”歳”>20 歳 </NUMEX>) が
水温 20 ℃	水温 <NUMEX nid=”n0” value=”20” counter=”℃”>20 ℃ </NUMEX>
風速 10 メートル	風速 <NUMEX nid=”n0” value=”10” counter=”m”>10 メートル </NUMEX>
小さじ 5	<NUMEX nid=”n0” value=”5” counter=”小さじ”> 小さじ 5</NUMEX>
大さじ 3 杯	<NUMEX nid=”n0” value=”3” counter=”大さじ”> 大さじ 3 杯 </NUMEX>
月刊マガジン 12 月号	月刊マガジン <TIMEX3 tid=”t0” type=”DATE” valueFromSurface=”XXXX-12” temporalFunction=”true”>12 月 </TIMEX3> 号
PC-98、iPhone4、PS2	PC-98、iPhone4、PS2
1 アンダー	1 アンダー
5 バック	5 バック
2 大ゲリラ組織	2 大ゲリラ組織
そんな一幕があった	そんな一幕があった
高齢者	高齢者
47 都道府県	47 都道府県
音速	音速
半分	半分
半額	半額
Vol. 1.2	Vol. 1.2

A.2 序数について

序数とは物の順序を表す数である。日本語では英語における「first」のように数詞そのものが順番を表す単独の序数詞が存在せず、基本的には「第-」「-目」などを用いることで順序を表す。我々は一部の序数と順位表現のみを数量表現とみなす。規格化の際には、@ordinal = ”true”とする。以下は数量表現とみなす序数である。

- 個数、量を示す数量表現に「目」「第」などがついたもの（※）

- 例：「三人目」「三度目」「第三回」

- ※に意味が等しい表現が存在する、数を含まない序数
  - 例：「初代」（＝「1 代目」）「初回」（＝「第一回」）

以下は数量表現とみなさない序数である。

- 個数、量を示さない数を含んだ表現に「目」「第」などがついたもの
  - 例：「第三豊栄丸（船の名前）」「第二号機」
- ※に意味が等しい表現が存在しない、数を含まない序数
  - 例：「長女」（＝「1 番目の娘」だが、「娘」を数える単位は存在しない）
- 「第二次○○」「第二回○○」など
  - 「第二次」「第二回」のみを序数として扱う。

A.3 特殊な数量表現

「階」について、辞書には「（名詞）多層の建物の一つの層。　例：受付は二―にある」「（接尾詞、助数詞）建築物の層を数えるのに用いる。　例：三五―建てのビル」という 2 つの意味が述べられている。辞書の定義に従うと、前者の用法において「階」は助数詞ではなく、場所を表す表現であり個数や量を示さないが、例外的に数量表現として扱う。後者の用法は建物の階数を示す一般的な数量表現である。規格化の際には、どちらも同じ「階」という単位を用いる。「地上」や「地下」などは考慮しない。

(8)　受付は二階にある。

受付は<NUMEX nid=”n0” value=”2” counter=”階”>二階</NUMEX>にある

(9)　地上８階地下１階建ての建物。

地上 <NUMEX nid=”n0” value=”8” counter=”階”>8 階 </NUMEX>
地下 <NUMEX nid=”n1” value=”1” counter=”階”>1 階 </NUMEX>
建ての建物。

単位が省略されていても、その表現が示す数と単位が明らかである場合は、数量表現として扱う。

(10)　成澤さん (20) が

成澤さん (<NUMEX nid=”n0” value=”20” counter=”歳”>20</NUMEX>)
が

「水温」「風速」はタグ付け対象とするが、規格化の際に「水の温度であること」「風の速度であること」といった情報は考慮しない。

(11)　水温 20 ℃、風速 10 メートル。

<NUMEX nid=”n0” value=”20” counter=”℃”＞水温 20 ℃ </NUMEX>、
<NUMEX nid=”n0” value=”10” counter=”m”＞風速 10 メートル </NUMEX>

その他の特殊な数量表現を以下に列挙する。

- 「小さじ５」「大さじ 3 杯」　単位は「小さじ」「大さじ」とする。
- 「47 都道府県」「二府四県　前者の単位は「都道府県」とする。後者は 2 つの数量表現とみなし、単位は「府」「県」とする。。
- 「99 連発」　「連発」は単位としてみなしにくいが、例外的に数量表現として扱う。
- 「三連続トライ」　「連続」は単位としてみなしにくいが、例外的に数量表現として扱う。「トライ」は含めない。

A.4　タグ付けしない表現

震度は数を用いて定められた階級を表す表現であり、量や個数を表す表現でないためタグ付けしない。「A4」などの表現も数を用いて定められた規格を表す表現であるため、タグ付けしない。

(12)　震度 5 弱

震度 5 弱

(13)　A4 サイズの紙が欲しい

A4 サイズの紙が欲しい

「北緯 30 度」「西経 145 度」は場所を表す表現であるため、タグ付けしない(「度」のみタグ付けする)。

(14)　北緯 30 度

北緯 <NUMEX nid=”n0” value=”30” counter=”度”>30 度</NUMEX>

学年は入学年度で区別された学生の集団を示し、個数や量を表さない。

(15)　家の前の道路で小学三年生くらいの女の子が派手にこけた。

家の前の道路で小学三年生くらいの女の子が派手にこけた。

(16)　津賀田中学校 3 回生同窓の広場

津賀田中学校 3 回生同窓の広場

株指標は単位が不明なため対象としない。

(17)　日経平均株価：12,861.56

日経平均株価：12,861.56

その他、単位が不明な表現や、単に数を含むだけの表現はタグ付けしない。

(18)　2LDK の部屋、「Ｖｅｒｓｉｏｎ６．０．５」「３極」「二けた」「５合目」「二頭立て」「Ｍ５（野球のマジック）」「二重」「3 周」

その他のタグ付けしない表現を以下に列挙する。

- バージョン情報　　例：「月刊マガジン 12 月号」
- 製品名中の数字　　例：「PC-98、iPhone4、PS2」
- 試合のポイント：もしも特定のポイントなどを表す表現なのであれば数量表現とみなす。以下の例はみなさない。　例：「1 アンダー」
- 「５バック」：サッカーにおける表現である。「DF の数」を表しているとも言えるが、あくまでサッカーのフォーメーションをさすと考え、数量表現とはみなさない。
- 「2 大ゲリラ組織」「2 大政党」：「組織数が 2」と示しているのではなく、「ある 2 つの組織」を示している。数量表現とはみなさない。
- 「そんな一幕があった」「一時は避難を考えた」「第一歩を踏み出した」「一審の判決が出た」：数量表現とはみなさない。
- 「高齢者」：「高齢の人」を指し年齢を指してはいない。数量表現とはみなさない。
- 「半額」「半数」：それぞれ「額が 5 割であること」「数が 5 割であること」を示す。割合により示されるものが示されており、通常の数量表現とは異なる。数量表現としてみなさない。
- 「Vol. 1」：「第一巻」と等しいようにもみえるが、「Vol. 1.2.1」などの表現だとよくわからない。数量表現とみなさない。

発表文献一覧

学術論文誌

- Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa , Keita Nabeshima, Naoaki Okazaki and Kentaro Inui. Leveraging Diverse Lexical Resources for Textual Entailment Recognition. The Special Issue of ACM TALIP on RITE (Recognizing Inference in TExt), December 2012.

国際会議論文

- Katsuma Narisawa , Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki and Kentaro Inui. Is a 204 cm Man Tall or Small ? Acquisition of Numerical Common Sense from the Web. In Proceedings of ACL 2013, August 2013.
- Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa , Keita Nabeshima and Kentaro Inui TU Group at NTCIR9-RITE: Leveraging Diverse Lexical Resources for Recognizing Textual Entailment The 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies (NTCIR-9), December 2011.

国内会議・研究会論文

- 成澤克麻 , 渡邉陽太郎, 水野淳太, 岡崎直観, 乾健太郎. 数量の大小の自動判定：「彼は身長が 2m ある」は高いか低いか. 言語処理学会第 19 回年次大会論文集, March 2013.
- 成澤克麻 , 比戸将平, 海野裕也, 松井くにお, 鈴木隆一, 田代光輝, 丸山宏 NIFTY-Serve におけるフォーラムデータの分析. 第 5 回知識共有コミュニティワークショップ論文集, November 2012.

- 岡崎直観, 成澤克麻 , 乾健太郎. Web 文書からの人の安全・危険に関わる情報の抽出. 言語処理学会第 18 回年次大会論文集, March 2012.
- 成澤克麻 , 渡邉陽太郎, 水野淳太, 岡崎直観, 乾健太郎. 数量表現を伴う文における含意関係認識の課題分析. 言語処理学会第 18 回年次大会論文集, March 2012.