数量表現を伴う文における含意関係認識の課題分析

成澤克麻 †　渡邉陽太郎 †　水野淳太 †　岡崎直観 †‡　乾健太郎 †

† 東北大学　‡ 科学技術振興機構　さきがけ
{katsuma, yotaro-w, junta-m, okazaki, inui} @ecei.tohoku.ac.jp

1 はじめに

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る。

含意関係認識の課題の1つに、一方もしくは両方の文が数量表現を伴う場合への対応が挙げられる。例えば、現状では以下のような例で含意関係を認識するのは難しい。

(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。

このような文の認識は、先に述べた含意関係認識技術の応用先である質問応答や情報抽出などで重要である。しかし、Sammonsら [8] やLoBueら [6] も指摘するように、現状の含意関係認識の研究は数量表現を伴う文に対して十分な対応ができていない。実際にRTE-6 [5] において (1) のような問題を解決しているシステムは存在せず、一部のシステムが比較的簡単な数量表現の問題に対応している程度である。

この現状は、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による部分が大きい。含意関係認識における数量表現の扱いは既存研究でほとんど無視されてきた問題であり、近年Sammonsらや LoBueらの研究によりこの問題が指摘されたものの、これらの研究も問題の指摘に留まり、具体的な問題事例の提示や解決に必要な処理の分析はされていない。

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。本稿はその第一歩として、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。具体的にはRecognizing Inference in TExt (RITE) [4] のデータと小谷ら [14] の評価セットの２つの日本語含意関係コーパスから数量表現の処理が問題となる事例を集め、その分析を行った。分析の結果、どのような問題があったかを報告し、問題の解決のために必要な処理について述べる。また一部の問題を実際に解決した。

2 関連研究

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。吉田ら [11] とFontouraら [3] はテキストから数量表現を認識し、情報検索に役立てる手法を報告している。RTE-6[5] 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事が分かるが、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。

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 [8] はRTE-5[1] 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 [9][10] 見られたのみであった。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 [15][17][7] 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 [2]。

以上より、本稿では日本語を対象とした数量表現を伴う含意関係認識にどんな課題があるのか明らかにし、解決のために必要な処理を述べる。

3 問題の分類と分析

本稿で対象とする数量表現とは、「5 人」「七個以上」のような数詞+助数詞 (+接頭辞や接尾辞）という形の表現を指す。

我々は RITE で使われた開発データとテストデータ (BC,MC の計 940 文対× 2) と小谷らの評価セット (2471 文対）から数量表現の処理が問題となる事例を集め、その分析を行った。分析の結果、コーパス中で数量表現が問題となる事例は計 118 ペアであった。分析を行うにあたって、述語項構造解析や照応など数量表現の問題とは関係のない部分の問題は全て理想的に解決されていると仮定し、数量表現に固有な問題のみを分析した。

本稿では含意関係認識を行う上での問題を、大まかな難しさの順番に

1. 2 文の構造はほぼ等しく、含まれる表現間の対応付けが問題となる場合（文節レベルの問題）
2. 2 文の構造の異なりが問題となり、それを解決すると 1 のレベルの問題に落とし込める場合（文構造レベルの問題）
3. 上のどちらでもない場合（意味レベルの問題）

の 3 つのカテゴリに分ける。

3.1 文節レベルでの含意関係

数量表現とその数量表現に対応する表現の対応付けが問題となるカテゴリである。さらに細かく問題を分けて、それぞれで必要な処理について述べる。

3.1.1 数量表現間の含意関係

以下の例のように、2 つの数量表現が同義語または上位語下位語の関係にある場合である。

(2) t：この商品は 20 %引きだ。
h：この商品は二割引きだ。

(3) t：宇宙の年齢は１３０億歳だ。
h：宇宙の年齢は１００億歳以上だ。

この認識のためには以下の処理が必要となる。

1. 数の表記の統一：「210000」「二十一万」「21 万」のような表現の違いを吸収する
2. 単位の統一：「割」「%」、「トン」「kg」など違う表現で同じ概念を指す単位を同じ単位に統一する
3. 数の包含関係の認識：数の表記と単位を統一した後、「100 以上」と「130」のような数の包含関係を認識する

3.1.2 数量表現と量を表す表現の間の含意関係

以下の例のように、「たくさん」「全部」などの（数量表現以外の）量を表す表現と数量表現が含意関係にある場合である。

(4) t ：100 人の人々が突然踊り出した。
h ：大勢の人々が突然踊り出した。

量を表す表現には主観に依存する表現と主観に依存しない表現の 2 つがある。「全部」「半分」のような主観に依存しない表現との含意関係を認識するには、これに対応する数量表現の知識（例えば「全部」は「100 %」）が必要となる。

「大勢」のような主観に依存する表現と数量表現の含意関係を認識するには、数量表現が表す量の大小を認識しなければならない。「大勢」が表す量は、その対象によって異なるため、数量に関する事前知識が必要である。

3.1.3 2 つの項をもつ数量表現による含意関係

数や量を表す数量表現が 1 つの項との関係を表すのに対して、割合や順序を表す数量表現は 2 つの項との関係を表す。例えば、「太郎がリンゴを 100 個持っている」の場合は「100 個」は「リンゴ」のみに関係する情報だが、「リンゴは全体の 3 割だ」の場合は「3 割」は「全体」と「リンゴ」の 2 つに関係する情報である。特に割合を表す数量表現について、この項を認識する必要がある。以下の例は、割合を表す数量表現と量を表す数量表現が含意関係にある場合である。

(5) t :人間の遺伝子は予測を含めて 3 万 2615 個で、ショウジョウバエの遺伝子は 約 1 万 5 千個 である。
h :人間の遺伝子は予測を含めて 3 万 2615 個で、ショウジョウバエの遺伝子は その半分程度 である。

この認識のためには以下の処理が必要となる。

1. 割合を表す表現の項（「何の」割合なのか）を認識する
2. その項の数を表す数量表現を認識する
3. 表現に沿って計算を行う（半分なら 1/2 にする計算）

3.1.4 並列関係にある表現と数量表現の間の含意関係

以下の例のように、h の数量表現が t の並列関係で表される表現をまとめあげた表現の場合である。

(6) t_1 :北京の展覧会には、日本、中国、韓国の 漆芸作家の作品が並ぶ。
h_1 :北京の展覧会には、三国の 漆芸作家の作品が並ぶ。

この認識のためには以下の処理が必要となる。

1. 文中の並列関係にある語句の認識
2. 並列関係にある語句の数と、数量表現が表す数が等しいことの認識
3. 数量表現の助数詞が表すものが、並列関係にある語句の上位語であることの認識

また、以下のように個数の情報も含まれる場合は、個数の情報の認識・足し合わせも必要である。

(7) t_1 :ボランティアの責任者から黒メダカ 5 匹とヒメダカ 5 匹をもらった。
h_1 :ボランティアの責任者からメダカ計 10 匹をもらった。

3.2 文構造レベルの問題：被限定名詞の同定

以下の文は含まれる表現がほぼ同じだが、文の構造が大きく異なる。

(8) t :韓国では女性２２・３％が整形経験者である。
h :韓国では整形経験者の女性が２０％以上いる。

このような文構造の違いを吸収すること自体は数量表現に固有な問題ではない。しかし数量表現を含む場合は数量表現が量化する名詞の同定を正しく行うことが必要とされる[1]。

量化される名詞の同定に必要な処理は数量表現のタイプにより異なる。文中で占める位置に注目すると、数量表現は以下の 3 タイプに分けられる[2]（分類名は現代日本語文法 [16] による）。

(9) a. 昨日会った 3 人の学生が来た。（名詞修飾型）
b. 昨日会った学生が 3 人来た。（動詞修飾型）
c. 昨日会った学生 3 人が来た。（添加型）

数量表現が量化する名詞は、名詞修飾型は単純に係り先、添加型は直前の名詞、動詞修飾型の被限定名詞は、自動詞の場合はガ格、他動詞の場合はヲ格となる。このように動詞修飾型が量化する名詞の同定には若干特殊な処理を要する。この処理に失敗すると、次のような表現の含意関係を認識できない。

(10) a. 学生が先生を 3 人招待した。
b. 3 人の学生が先生を招待した。

3.3 意味レベルの問題

語彙的含意関係と構文的言い換えの組み合わせに帰着できない事例が存在する。これらの事例を「意味レベルの問題」として扱う。

3.3.1 テンプレートに基づく立式と評価

意味レベルの問題のうち、一部はテンプレートベースの情報抽出の延長としてアプローチできる可能性がある。以下の例では、t と h の文中において、ある対象の「変化前の数量」「変化後の数量」「変化量」について述べられる。

(11) t :五羽の仔ウサギが産まれて、三羽が死んでしまった。
h : 二羽の仔ウサギが生きている。

また以下の例では、t と h の文中において「ある存在 A の数量」「ある存在 B の数量」「その差の数量」について述べられる。

(12) t :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。
h :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。

これらの例では、それぞれの数量表現の関係性がわかれば、式が立てられる（例えば (9) では「5 – 3 = 2」という式が立てられる）。すなわちこの問題は次の処理を必要とする。

1. 式のテンプレート（例えば「変化前＋変化量＝変化後」）を用意
2. 文中のテンプレートに当てはまる数量表現を抽出
3. テンプレートに沿って式を評価

同様の手法が阿部ら [12] 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。

3.3.2 その他

今までの分類には上手く分類できなかったものをここで挙げる。これらの問題を整理・分析することは今後の課題である。

(13) 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。

[1]本稿では、修飾節となる数量表現は全て、名詞を量化するものとして扱う。「300 グラムの袋」などといった表現は一般的に量化表現とはみなされないが、宇都宮 [13] の定義に従い、これらを「間接量化表現」として扱う。本稿では詳しく述べないが、間接量化表現が動詞修飾型として用いられる文と名詞修飾型として用いられた文との間に含意関係はないので注意が必要である。

[2]厳密には、「修飾節となる数量表現」の分類である。現代日本語文法 [16] の分類には「3 人が来た」「来た学生は 3 人だ」のような非修飾節の数量表現を考慮していない。一方が修飾節の数量表現を含み、もう一方が対応する非修飾節の数量表現を含むような 2 文間の含意関係認識においても少々問題が生じるが、紙面の都合でここでは紹介しない。

(14) t：福島県いわき市は日本一広い市だ。
h：福島県いわき市は福島一広い市だ。

(15) t：山梨県はミネラル水の生産量が日本全体の 50 %を占める。
h：山梨県はミネラル水の生産シェアが日本で 1 位だ。

4 数量表現の規格化

3.1.1 節で述べた数量表現間の含意関係認識を行うために、文中の数量表現の認識と規格化を行うシステムを作成した。規格化とは、図 1 のように数量表現を「[単位],[表す数の範囲]」という規格に変換することを指す。規格化された数量表現間では数の範囲を比較することで含意関係を認識できる。このシステムでは数量表現と似た特徴を持つ時間表現も規格化の対象とした[3]。

接頭辞	*特殊*	*数詞*	*単位*	*接尾辞*	*規格化表現* (*単位, 数の範囲*)
およそ	秒速	5	cm		[cm/s , 3〜8]
		一万	円	以上	[円 , 10000〜∞]
		2〜3	人		[人 , 2〜3]

図 1: 数量表現の構成と対応する規格化表現

認識・規格化は数量表現の構成性に着目して行った（図 1）。我々は数詞以外の 4 つの構成要素で、形態素と機能（例えば「約」の機能は「数の範囲を漠然的にする」）を記述した辞書を作成した。数量表現の認識の際にはこれらの辞書の要素の組み合わせからなる表現を数量表現として認識し、規格化の際は使用した要素が持つ機能を組み合わせることで規格化表現を出力する。時間表現の認識・規格化は正規表現を用いて行った。

提案システムを評価するため、NAIST テキストコーパス中の 2098 文に対してシステムを適用し、文中に含まれる数量表現と時間表現を正しく認識・規格化できたかを人手で評価した。実験の結果、1657 個中 1461 個を正しく規格化し、正解率は 88.2 %となった。結果を分析すると、入力文中には「六冠王」「震度 3」のような数量表現なのか曖昧なもの、「1 個辺り 30 円」のようなどのような規格化表現にするか曖昧なものがあり、規格化する対象の定義がまだ不十分であることが明らかになった。（定義が曖昧だったものは全て誤りとした）他の誤り例としては「五十嵐」「四日市」など人名や地名が数量を含む場合の誤認識があった。

実装したシステムは使用した辞書とともウェブ上で公開している[4]。

[3]今回は数を含む時間表現のみを対象とし、「節分」のような表現は対象としていない

[4]http://www.cl.ecei.tohoku.ac.jp/~katsuma/

5 おわりに

本稿では、含意関係認識課題において数量表現が問題となる事例に焦点を当て、この問題を分析・整理する事を目的として日本語含意関係コーパスから該当する事例を集め、その分析を行った。その結果、問題を大きく 7 つのカテゴリに分け、それぞれの問題を解決するために必要な処理を明らかにした。また、もっとも基本的な要素技術として数量表現の規格化をとりあげ、モジュールを実装し、評価・公開した。

今後の課題としてリソースの整備があげられる。今回対象としたコーパスは小規模であり問題の分析が十分に行えたとは言い難く、また今後問題の解決を図る上でも一定の規模の評価データが必要である。今後は数量表現に関する研究に有為なコーパスを作成し、それを用いて問題の解決と今回分類しきれなかった問題に対しての分析を与える予定である。

謝辞 本研究は、文部科学省科研費（23240018）、(23700157)、および JST 戦略的創造研究推進事業さきがけの一環として行われた。

参考文献

[1] L. Bentivogli, I. Dagan, H.T. Dang, D. Giampiccolo, and B. Magnini. The fifth pascal recognizing textual entailment challenge. In *Proceedings of TAC 2009 Workshop*, 2009.
[2] F. Bond. Determiners and number in english, contrasted with japanese, as exemplified in machine translation. *Unpublished doctoral dissertation, University of Brisbane, Queensland, Australia*, 2001.
[3] M. Fontoura, R. Lempel, R. Qi, and J. Zien. Inverted index support for numeric search. *Internet Mathematics*, Vol. 3, No. 2, pp. 153–185, 2006.
[4] S. Hideki, K. Hiroshi, L. Cheng-Wei, L . Chuan-Jie, M. Teruko, M. Yusuke, S. Shuming, and T. Koichi. Overview of ntcir-9 rite: Recognizing inference in text. In *Proceeding of NTCIR-9 Workshop Meeting* , pp. 291–301, 2011.
[5] 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.
[6] P. LoBue and A. 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. Association for Computational Linguistics, 2011.
[7] S. Nishiguchi. Quantifiers in japanese. *Logic, Language, and Computation*, pp. 153–164, 2009.
[8] M. Sammons, VG Vydiswaran, and D. 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. Association for Computational Linguistics, 2010.
[9] Y. Tsuboi, H. Kanayama, M. Ohno, and Y. Unno. Syntactic difference based approach for ntcir-9 rite task. In *Proceeding of NTCIR-9 Workshop Meeting* , pp. 404–411, 2011.
[10] Y. Watanabe, J. Mizuno, E. Nichols, K. Narisawa, K. Nabeshima, and K. 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.
[11] M. Yoshida, I. Sato, H. Nakagawa, and A. 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.
[12] 阿部一貴, 吉村枝里子, 土屋誠司, 渡部広一. 意味処理を用いた算数文章題演算処理手法の提案. 情報処理学会研究報告. ICS,[知能と複雑系], Vol. 158, p. 1, 2010.
[13] 宇都宮裕章. 量化及び遊離文の認知論的分析. 静岡大学教育学部研究報告. 人文・社会科学篇 50, pp. 1–16, 1999.
[14] 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.
[15] 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.
[16] 日本語記述文法研究会. 現代日本語文法 2 第 3 部格と構文 第 4 部ヴォイス. くろしお出版, 11 2009.
[17] 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.