2011年度　卒 業 論 文

数量表現を伴う文における含意関係認識の課題分析

2012年3月

情報知能システム総合学科
(学籍番号: A8TB2165)

成澤 克麻

東北大学工学部

概 要

テキスト t から仮説 h が推論可能であるとき、t と h を含意関係にあると呼び、二文の含意関係を認識することを含意関係認識と呼ぶ。含意関係認識は、質問応答や文書要約等の自然言語処理の応用分野で必要とされる技術である。含意関係認識を行うための課題は様々だが、その 1 つに文が数量表現を伴う場合への対応が挙げられる。しかし既存研究において数量表現の問題に着目した研究は行われておらず、分析が足りていない。そこで本論文では、現存する日本語含意関係認識コーパスを対象に、数量表現が問題となる事例を抽出し、課題の分析を行った。

今回の分析から、問題の半数は必要とされる処理がある程度明確だが、残りの半数については様々な推論の複合であり数量表現の問題として切り分けにくく、必要とされる処理が明確にしにくい難しい問題であることが判明した。複雑な推論が必要な難しい問題に対して、どのような枠組みで含意関係認識を行っていくべきなのかを考えることは、今後の含意関係認識の研究の上で必須であるが、数量表現という側面から事例を分析した今回の分析は、これを考える上での 1 つの参考となるであろう。また、処理が明確になった問題において、意味上の並列関係の認識や数量の主観的な大小の認識などこれまで自然言語処理においてあまり注目されていなかった技術が多く必要とされており、数量表現という切り口から問題を捉え解決を図ることはやはり必要であると考えられる。

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、システムを実装し、評価・公開した。数量表現が示す数量の情報を表層の違いを吸収して得るこの技術は、数量表現を扱う上で必須の技術である。評価実験においては、新聞記事コーパスを用いたテストデータにおいて適合率が 0.95 となった。

目 次

第1章　序論

The PASCAL Recognising Textual Entailment(RTE) Challenge[4] に代表されるように、含意関係の認識は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。例えば質問応答において、h の文章をクエリとして同じことを述べている文章を探す際に、t のように数量の情報しか書かれていない文章も抽出するべきである。この例に限らず、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るが、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。RTE-6 [8] においては (1) のような問題を解決しているシステムが存在せず、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。Sammons ら [14] や LoBue ら [10] においても、最新の含意関係認識システムが冒頭の例のような数に関わる推論が必要な事例が解決できていないと指摘されている。

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 つに、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が挙げられる。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 らの研究により数に関わる推論の必要性が指摘されたものの、これらの研究は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) [6] のデータと小谷ら [19] 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。

第2章　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 [17] と Fontoura ら [5] 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[8] 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 [2],[7],[9],[13],[11]。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。[11] は RTE6 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している。

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 [14] 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 [10] 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 [15][16] 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 [20][22][12] 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 [3]。

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。

第3章　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。

我々は現存する全ての日本語含意関係認識コーパスといえる、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。本稿の最後に付録として、抽出した全ての事例をまとめたので参照されたい。

本稿では含意関係認識を行う上での問題を、問題の整理のため

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

このカテゴリの問題はデータ中に最もよく見られる問題であり、関連研究で述べた通り、英語ではこれに対応している既存研究もいくつかみられた。ただし、単位の問題をどれほど解決しているかなど不明な点は多く、また数の表記や単位の問題（助数詞の豊富さ）のように日本語に固有な問題も少なからずある。今回我々は、この問題を解決すべく、日本語の数量表現の規格化を行うシステムを作成した。これについては 4.1 節で詳しく述べる。

本稿における数量表現の定義とは異なるが、「10 メートルを超える」「10 人から 20 人まで」のような数量表現と他の語が組合わさって、ある数量を表す表現がコーパス中には見られた。このような表現の扱いは検討中だが、ひとまず「ある数量を表している表現」として、もとの数量表現の定義を拡張し、このカテゴリの問題として扱う。

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」「半分」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 ：日本人のおよそ 5 割がそれを信じている。

h ：日本人のおよそ半分がそれを信じている。

(5) t ：AKB48 のライブに 10000 人が押し寄せた。

h ：AKB48 のライブに大勢が押し寄せた。

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。該当する表現の数は限られるため、対応は十分に可能だと考えられる。

「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、量を表す表現が何の量の大小について述べているのか（例 4 では「突然踊り出した人々の人数」）を認識した上で、対応する数量表現がその量の大小の条件を満たすのかを認識する必要がある。（例 4 では「100 人は突然踊り出した人々の人数としては多いのか？」）これを行う上での問題としてまず考えられるのは、主観は個人によって異なるということである。これに関しては、ある程度一般的な人の感覚を参照する、すなわち誰にとってもほぼ等しく言えそうな量の大小を認識させることでの解決が図れる。数量表現の大小を見極めるためには、ある対象の数量に関する大小の知識（例えば、AKB48 のライブに◯人くれば多い、もしくは◯人が普通の人数、といった知識）を対象ごとに知

識獲得すれば良さそうに思えるが、この「対象」の候補がいくらでも存在するため非常に難しい。例を挙げれば、AKB48 ではなく他のグループではどうなのか？「武道館で行われる AKB48 のライブに〜」や「田舎で突発的に行われた AKB48 のライブに〜」だったらどうなのか？などと考えていけば、単純に知識獲得するだけでは解決は難しそうである。

1 つ注意したいのが、割合に関する大小の話である。「死者の 8 割」という表現は、それ自体が「大勢の死者」なのかどうかは死者に関する大小の知識（例えば、死者が◯人であれば多い、など）が必要だが、知識を必要とせず「死者の大部分」なのである。当然「何割以上であれば大部分なのか？」といったルールを決める必要はあるが、これは上で述べた話とは異なる話であることに注意されたい。

最後に、直接量を表す表現ではないものの、量に関する情報をもつ表現についても、このカテゴリの問題として言及する。

(6) t：父は還暦を迎えた。

h：父は六十歳になった。

(7) t：水深 1 メートルくらいのところで取れるモズクは、海藻らしい独特のとろみがあり、歯ごたえがよく、磯の香りも味わえる。

h：浅瀬で取れるモズクは、とろみはあまりなく、食感も風味もイマイチだ。

この例では、「還暦」＝「六十歳」、「水深 1 メートルくらいのところ」＝「浅瀬」と結びつける必要がある。「還暦」や「浅瀬」といった表現は、言葉がもつ情報として数量情報をもってはいるが、「半数」や「たくさん」のように言葉そのものが量を表しているわけではない。このような事例には、「還暦」＝「六十歳」のように数量表現に一対一対応に結びつけられるものと、「銀婚式」＝「25 回目の結婚記念日」のように数量表現 + 修飾される名詞のような表現に結びつけられるもの、「浅瀬」のように対応する数量表現の範囲が明確でなく、一対一対応には結びつけられないもの、などの種類がある。「還暦」のようなパターンはあまり存在せず（今のところ「歳」に関するものしか確認していない）、主観に依存しない表現と同じような解決が考えられる。「銀婚式」「浅瀬」といった事例への対応は検討中である。

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。

(8) 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。

(9) 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。

(10) t_2 :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が生息している。

h_2 :日本には 3 種類のカブトエビが 生息している。

(11) t_1 :言葉には「つくる楽しみ」と「使う楽しみ」があります。

h_1 :言葉には 二つの楽しみ があります。

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。

(12) 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。

(13)　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、数量表現には他の表現にはみられない独特な用法があり、数量表現を含む文の文構造の違いを吸収する際には、この用法について把握しておく必要がある。この用法とは、以下のようなものである。

(14) 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)

数量表現が文中で占める位置に注目すると、数量表現は以上の3タイプに分けられる[1](分類名は現代日本語文法 [21] による)。動詞修飾型(数量表現の遊離現象とも呼ばれる)や添加型の用いられ方は他に見られない数量表現独特な用法である。この3つの例が言い換え関係にあるとすると、これらの文構造を吸収するために最低限必要な処理として、数量表現が限定する名詞の同定が必要であると考える。よって本節では、この問題を数量表現が量化する名詞の同定問題として捉える。

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。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。名詞の同定に失敗すると、次のような表現の含意関係を認識できない。

(15) 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。

(16) 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。

[1]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 [21] 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。

(17) 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 [18] 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。

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。ここでは、大まかな分類とともに事例を紹介する。(厳密な整理ではないので注意されたい)

(18) t : イラク国立博物館で、貴重な展示品十数万点が略奪された。

h : イラク国立博物館は、略奪で壊滅的被害を受けた。

(19) 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。

例 (14) を解くための一つのやり方としては、「展示品が略奪される」＝「略奪で被害を受ける」であること、「展示品十数万点」は「たくさんの展示品」であること、「たくさんの展示品が略奪されること」は「略奪で壊滅的被害を受けた」であること、という 3 つのステップを踏む方法が考えられる。例 (15) を解くためにも、「70 億人」が「大勢」であり、「大勢が水不足」であることは「深刻な水不足」であると結びつけるのが妥当だと考えられる。このように、数量表現を推論を経てある種の程度表現のようなもの（「十数万点」は「壊滅的」、「70 億人」は「深刻」）に結びつけることが必要となる処理がこの分類には見られる。

(20) t : マドンナは今回三人目を妊娠した。

h : マドンナは三人子供がいる。

(21) t：甲子園は準決勝を迎えた。

h：8 チームが残っている。

これらの事例では、言葉がもつ数量情報を把握をかなり高度に行う必要がある。

(22) t：ミカンを 7 つまで数えた。

h：ミカンは 3 つしかない。

(23) t：らっきょう漬けは、原材料の食感、漬け込み前の乳酸発酵、調味液の三位一体の味だ。

h：らっきょう漬けは、調味液だけで味わいが生まれるのではない。

この事例では取り立て助詞が問題となる（下の事例は数量表現が表れないため厳密にはこの分析の対象とならないが、興味深い例として紹介する)。取り立て助詞には「だけ」「しか」「も」などがあり、ある集合の個数の情報を言外に述べる役割をもつものがある。このような情報をいかに扱うかを考える必要がある。

(24) t：木下大サーカスの観客数は、現在年間 120 万人で、米国リングリング・サーカスに次ぐ世界第 2 位だ。

h：米国リングリング・サーカスの観客数は、世界一だ。

(25) 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。

(26) 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。

これらは高度な推論が必要となる例である。例 (26) では、「山梨県のシェアが日本全体の 50%以上ならば、他の県のシェアは必ず 50%以下である」「よって山梨県のシェアは一位である」という推論が必要になる。

第4章　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、数の表記の統一（半角数字に統一）や基本的な単位の統一（「kg」「トン」「グラム」のような表記は、全て「g」に直される）を行いながら、図 4.1 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単位と数の範囲を比較することで含意関係を認識できるため、数量表現間の含意関係の問題は規格化が正しく行われれば解決できる。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[1]。

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 4.1: 数量表現の構成と対応する規格化表現

数量表現の認識・規格化は、数量表現がもつ構成性に着目して行った（図 4.1）。数量表現は、およそ「接頭辞」「特殊助数詞」「数詞」「助数詞」「接尾辞」に分割できる。我々は数詞以外の 4 つの構成要素で、表現とその機能（例えば「約」という表現の機能は「数の範囲を漠然的にする」、「メートル」という表現の昨日は「単位を m 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。用いた辞書を付録としてのせているので参照されたい。時間表現の認識・規格化は、時間表現のパターン数は少ないため正規表現を用いて行った。

システムは大まかに以下のような流れで処理を行う。まず始めに、前処理を行う。この前処理部分では、文中に含まれる数詞を全て半角数字に直す処理を行う。この後に時間表現の認識・規格化を行う。この処理は正規表現によるマッチングで行われ、文中から該当する表現がなくなるまで続けられる。最後に数量表現の認識・規格化を行う。ここでは、始めに文頭から見て始めに位置する半角数字列を探索し、半角数字列があれば、数字列の前の文字列が「特殊」辞書内の単語に後方一致するかどうかを判断し、後方一致しなければ引き続きその文字列が、後方一致すれば一致した分の特殊の単語を除いた更にその前の文字列が、「接頭辞」辞書内の単語に後方一致するかどうかを判断する。数字列の後の文字列についても同様に「助数詞」「接尾辞」と前方一致するかどうかを判断する。最終的に、一致した部分が数量表現となる。最後に、マーキングした文章と、規格化された表現を出力する。

提案システムを評価するため、NAIST テキストコーパス（新聞記事コーパス）中の 2098 文に

[1]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

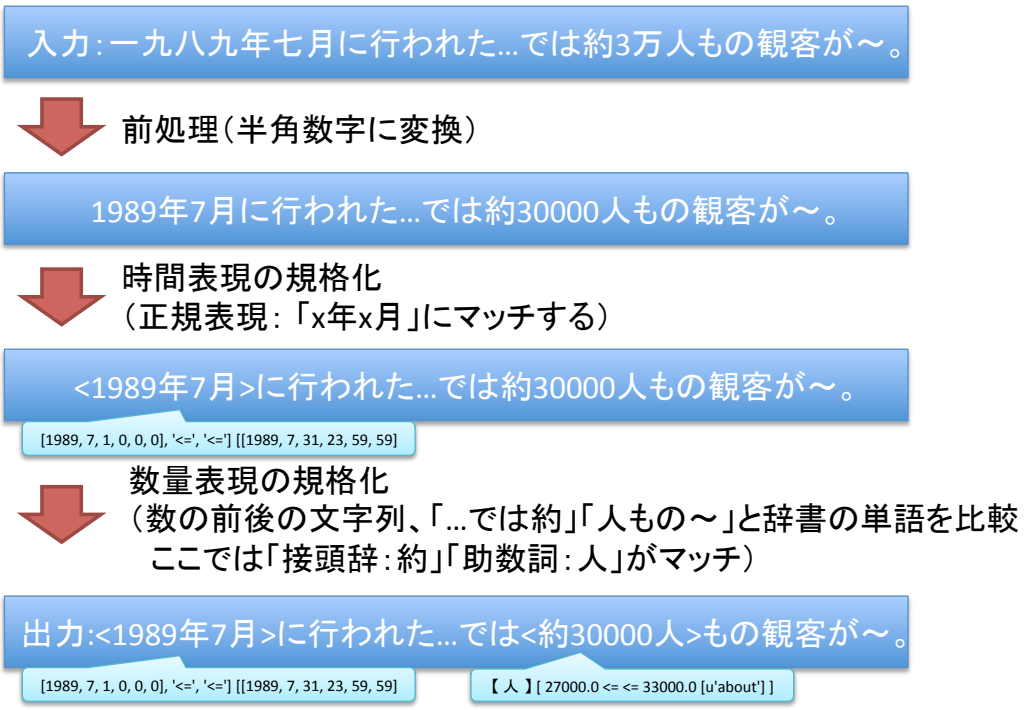

図 4.2: フロー

対してシステムを適用し、文中に含まれる 1492 個の数量表現・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、適合率は 0.95、再現率は 0.91 となった。今回用いた辞書データは、数時間程度で構築した決して完全とは言えない辞書データであったが（特に SI 単位系などは未対応なものが多い)、結果より基本的な表現はほとんど規格化できたといえる。結果を分析すると、入力文中には「六冠王」「震度 3」のような数量表現として扱うべきなのか曖昧なもの、「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。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。他に誤りだった例として「３―５個」など、規格化の対象として予期していなかった表現（この例では「―」を範囲を表す表現に含めていなかった）や「五十嵐」「四日市」など人名や地名が数量を含む場合があった。

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[2]。

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

第5章　結論

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。分析においては、問題をその性質から大きく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。今回の分析から、問題の半数は必要とされる処理がある程度明確だが、残りの半数については様々な推論の複合であり数量表現の問題として切り分けにくく、必要とされる処理が明確にしにくい難しい問題であることが判明した。複雑な推論が必要な難しい問題に対して、どのような枠組みで含意関係認識を行っていくべきなのかを考えることは、今後の含意関係認識の研究の上で必須であるが、数量表現という側面から事例を分析した今回の分析は、これを考える上での1つの参考となるであろう。また、処理が明確になった問題において、意味上の並列関係の認識や数量の主観的な大小の認識などこれまで自然言語処理においてあまり注目されていなかった技術が多く必要とされており、数量表現という切り口から問題を捉え解決を図ることはやはり必要であると考えられる。

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。

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、システムを実装し、評価・公開した。数量表現が示す数量の情報を表層の違いを吸収して得るこの技術は、数量表現を扱う上で必須の技術である。評価実験においては、新聞記事コーパスを用いたテストデータにおいて適合率が0.95となった。今後は、扱うべき数量表現の定義を厳密に定めながら、システムの性能を向上させていく。

謝 辞

本研究を進めるにあたり、終止熱心なご指導を頂いた乾健太郎教授、岡崎直観准教授、渡邉陽太郎助教に感謝致します。乾教授と岡崎准教授には、私の稚拙な質問や相談に対しても問題の本質を的確に汲み取ったご助言を何度も頂きました。本当に感謝しております。渡邉助教には普段から私の議論に付き合って頂き、様々な知見を頂くとともに、精神的にも支えられました。ありがとうございます。

また、日常の議論を通じて多くの知識や示唆を頂いた研究室の皆様に感謝します。特に、水野淳太さんには日常の些細なことから研究上の話まで、非常に多くの事について相談させて頂き、感謝の念にたえません。本当にありがとうございました。

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。

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] 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.

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

[4] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 177–190, 2006.

[5] 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.

[6] 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.

[7] A. Iftene and Moruz M-A. Uaic participation at rte-6. In *Proceedings of the Third Text Analysis Conference (TAC 2010) November*, 2010.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

付 録A　数量表現を伴う事例

事例を分類とともに挙げる。実際は問題となっていないが、似通った性質を持つものについても参考までに挙げてある（参考事例には*を付与している）。小谷らの評価セットについては、アノテートされていた「カテゴリ」「推論判定」の情報も記述する。

■数量表現間の含意関係

- 一般

394 語彙 (体言):同義語 ◎ この商品は 20%引きだ。 この商品は二割引きだ。
t：東京・上野動物園で約５００人の若者が東京芸術大学の入試のために一斉に動物を描いた。 h：５００人の東京芸術大学の受験者が動物の絵を描いた。
t：クラスター爆弾を使用する場合、親爆弾に含まれる子爆弾は６５０個としている。 h：クラスター爆弾とは、親爆弾の中に数百個の子爆弾が入った兵器だ。
t：青森県住宅供給公社の元主幹は約１４億円を横領した。 h：青森県住宅供給公社の元経理担当主幹は、１４億４６００万円を横領した。
t：熱帯低気圧のうち中心付近の最大風速が約１７メートル以上になったものを「台風」と呼びます。 h：台風は中心付近の最大風速が１７メートル以上の熱帯低気圧をいう。
t：アルコール依存症は８０万人以上いると推計されている。 h：全国のアルコール依存症者は、約２２０万人といわれる。
t：宇宙の年齢は１００億歳だ。 h：宇宙の年齢は１３０億歳以上だ。
t：たばこを吸う飼い主に飼われていた猫は、吸わない飼い主の猫に比べて２・４倍も悪性リンパ腫になりやすいことを突き止めた。 h：たばこを吸う飼い主と暮らしている猫は、たばこを吸わない主人を持つ猫に比べて、２倍以上に悪性リンパ腫になりやすいことが分かった。
t：０３年と０４年の全国の盗難数を比較するにあたり、イモビライザー付きの車両に絞ると、ランドクルーザーは２５９台から５５２台と、２・１倍になった。 h：イモビライザーを搭載したランドクルーザーの盗難被害が、前年の約２倍になっていた。
t：無党派層の４割近くを取り込んでいる。 h：無党派層でも３割以上の支持を受けている。
t：約６割の学生は写真は合否に影響するとの意識を持っていた。 h：就職活動で履歴書に張る写真の良し悪しが合否に影響すると考えている学生が６割前後もいることが調査で分かった。
t：チャイルドシート着用率が警察庁と日本自動車連盟の調査で５２．４パーセントだと分かった。 h：チャイルドシートの着用率が約五割にとどまっていることが調査で分かった。
t：現在でも市中肺炎の４分の３は細菌によるものです。 h：市中肺炎では肺炎球菌によるものが約４割を占め最も多い。
t：日本の衆院議員に女性が占める割合は約９％だ。 h：世界の衆院議員に女性が占める割合は９％だ。
t：マウスは４万個近い遺伝子を持ち、その９５％以上が人間の遺伝子と共通すると考えられている。 h：マウスの遺伝子は人間の遺伝子と９割以上共通する。
t：韓国では男性９・３％、女性２２・３％が整形経験者であった。 h：韓国は整形をする女性が２０％以上いる。

- 「超えている」「満たない」「切った」のような数量を変化させる表現

t：世界中に６００頭程度しかおらず、極めて絶滅の危険が高いマウンテンゴリラが、ルワンダで殺された。 h：マウンテンゴリラは世界に７００頭に満たない希少動物だ。
1867 推論:時間軸・数量 ○ 日本の人口は一億を超えている。 日本の人口は九千万を超えている。
865 語彙 (用言):言い換え ○ 日本の人口は一億を超えている。 日本の人口は一億以上だ。
t：アフガニスタン産のアヘンは世界市場の４分の３を占めている。 h：アフガニスタンのアヘン生産量は、世界市場の７割を超えている。
t：朝の読書活動を実施する小学校は全国で９０００校を超えた。 h：全国に朝読書を取り入れる小学校が９０００校ある。
t：ソメイヨシノの寿命は１００年もない。 h：ソメイヨシノの寿命は一説に約６０年といわれる。
t：人間の腸には、大腸を中心に１００種類以上の細菌がすみついている。 h：ヒトの大腸には５００種類以上の細菌がすんでいる。
t：スマトラ沖大地震と大津波による死者数は１７万５４５８人となった。 h：スマトラ沖大地震は大きな津波を起こし、１５万人以上の死者を出した。
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。
t：チャイルドシートの使用率が５割を切った。 h：チャイルドシートの使用率が４７％に減った。
t：６歳未満の子供のチャイルドシート使用率は４７％だ。 h：６歳未満の子供のチャイルドシート使用率は５割を切っている。

- 「○人に○人」「○日に○個」、またその他の特殊な数量表現

t：関節リウマチは世界の１００人に１人がかかる。 h：関節リウマチは「２００人に１人の割合で発症する」という。
t：人の体には１日に２００個から３００個の「がんもどき」ができていると言われます。 h：人の体には１日に２００個から３００個の「がんのようなもの」ができていると言われます。
t：日本人１人は毎日約９メートルのトイレットペーパーを消費しているという。 h：日本人は毎年約９メートルのトイレットペーパーを消費しているという。
t：日本人はＡ型４割、Ｏ型３割、Ｂ型２割、ＡＢ型１割である。 h：日本人の血液型はＡ、Ｏ、Ｂ、ＡＢの各割合がほぼ４、３、２、１で構成されている。

- 参考

* t：イワシの豊凶は４０年周期で繰り返す。 h：イワシの豊凶は１００年周期で繰り返す。
* t：ビジネスマンの３人に１人が、会社に着いてから朝食をとっていることが、アサヒ飲料の調査でわかった。 h：３人に１人は、会社で朝食をとるビジネスマンがいる。
* t：北朝鮮拉致被害者の５人は帰国以来付けていた北朝鮮のバッジを外し、拉致問題解決を願うブルーリボンだけを胸に付けて懇談などに臨んだ。 h：北朝鮮拉致被害者の５人は帰国以来付けていたブルーリボンを外した。
* t：温泉法によると、温泉は水温２５度以上か、遊離炭酸（ＣＯ２）などの物質が基準値以上に溶存する温水・鉱水と定義されている。 h：「温泉」とは温泉法に基づき源泉の温度が２５度以上の湯か、もしくは定められた物質を一定以上含むものである。
* t：徳島県の上勝町が、ごみを３５種類に分別して回収する取り組みを続けている。 h：徳島県上勝町は、３４種類ものごみの分別を断行している。

■量を表す表現との含意関係

t：１８～２９歳の若い世代で「科学技術の話題に関心がない」と答えたのは５２％に上った。 h：半数以上の若者が科学技術には関心を払わない。
t：「自分は冷え性」と思っている女性は全体の約７割、学生の８割に上る。 h：女性の大半が自分は冷え症だと訴えている。
t：阪神大震災の被害に遭った人の多くが、建物の下敷きになり、死者の８割以上が建物の下敷きで亡くなっている。 h：阪神大震災では、死者の多くは建物の下敷きになり火災から逃げられなかったといわれる。
t：関東大震災では東京、神奈川を中心に死者、行方不明者が十数万人、家屋の全壊や焼失も数十万戸あったが、死者のうち８割以上が火災によるといわれる。 h：関東大震災では、死者の多くはがれきではなく火災によるといわれる。
t：メディア教育開発センターの調査で一部の大学生の語彙力が低いことが分かった。 h：メディア教育開発センターの語彙力調査で、私立大１年生の１９％が「中学生レベル」と判定された。

- 「銀婚式＝ 25 回目の結婚記念日」「若者＝ 18 29 歳」「水深 1 メートルくらい＝浅瀬」のような言い換え

351 語彙 (体言):定義的—推論:時間軸・数量 ◎ 佐藤夫妻は 25 回目の結婚記念日を迎えた。 佐藤夫妻は銀婚式を迎えた。
t：１８～２９歳の若い世代で「科学技術の話題に関心がない」と答えたのは５２％に上った。 h：半数以上の若者が科学技術には関心を払わない。
t：歌手、宇多田ヒカルさん（１９）が４月に卵巣腫瘍（しゅよう）の摘出手術を受けていたことが明らかになりファンを驚かせた。 h：卵巣や子宮などの婦人科の病気は決して中高年者だけに起こるものではない。
t：水深１メートルくらいのところで取れるモズクは、海藻らしい独特のとろみがあり、歯ごたえがよく、磯の香りも味わえる。 h：浅瀬で取れるモズクは、とろみはあまりなく、食感も風味もイマイチだ。
428 語彙 (体言):同義語 ◎ 成人を迎えた。 二十歳になった。
464 語彙 (体言):同義語—語彙 (用言):言い換え ◎ 父は還暦を迎えた。 父は 60 歳になった。
? 1855 推論:時間軸・数量 ◎ この商品は 20%引きだ。 8 掛けで買える。
* t：日本の財界や政界では７０、８０歳を過ぎても元気な指導者が多い。 h：日本の財界や政界では引き際を知らぬ年寄りが多い。

■ 2 つの項をもつ数量表現

* t：人間の遺伝子は予測を含めて３万２６１５個で、予想されていた数の３分の１であり、線虫やショウジョウバエのわずか２倍にとどまり、新たな謎も出てきた。 h：線虫やショウジョウバエの遺伝子は約１万５千個で、人間の遺伝子はその３分の１である。

■並列関係にある表現

t：北京の展覧会には、３国の漆芸作家の作品が並ぶ。 h：中国・北京に日本、中国、韓国の代表的な漆芸作家や研究者らが集い、シンポジウムと展覧会が開かれた。
t：伝統的な塩漬け卵黄の入った月餅のほか、朝鮮ニンジン月餅やフルーツ月餅、フカヒレ月餅、アイスクリーム月餅などが話題を呼んでいる。 h：さまざまな月餅が登場している。
t：ボランティアの責任者から黒メダカ５匹とヒメダカ５匹をいただき、我が家の家族になった。 h：ボランティアの責任者からメダカ計１０匹をもらった。
1850 推論:時間軸・数量 × 昨晩のおかずは肉ジャガと青菜のお浸しだった。

おかずは一品だけだった。
1857 推論:時間軸・数量 ◎ 京都で一泊、博多で二泊の出張をした。 三泊の出張をした。
t：日本には３種類のカブトエビが生息している。 h：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。
t：第８３回オール読物新人賞（文芸春秋主催）は志川節子さんの「七転び」と竹村肇さんの「パパの分量」に決まった。また、第７１回小説現代新人賞（講談社主催）は、やはり竹村さんの「ゴーストライフ」と橘かがりさんの「月のない晩に」に決まった。 h：竹村さんは、新人賞をダブル受賞した。
t：地震波は２種類あり、それぞれ速度が異なる。 h：地震波には初期微動のＰ波（縦波）と、大きな揺れをもたらすＳ波（横波）があるが、Ｐ波の速度は毎秒６～８キロで、Ｓ波の速度は毎秒３～４キロと約半分である。
t：古代ギリシャのピタゴラスには、同じ日の同じ時刻に南イタリアのメタポンティオンとクロトンという二つの場所で多くの人に目撃され、またギリシャのオリンピアの競技会にも姿を現したという不思議な言い伝えがある。 h：言い伝えでは、同じ日の同じ時刻にピタゴラスの目撃情報が三件ある。
* t：石破長官は英国、オランダ、フランスの３か国を訪問する。 h：石破長官はイギリス、蘭、仏を訪問する。
* t：日立製作所は、色覚異常の人に配慮し、それぞれのボタンの上部に「青」「赤」「緑」「黄」の文字を表示した。 h：ある社員は「４色ボタンは機能表示を色だけに頼る点で非常に特殊。なぜ色のみで識別する規格が決まったのか」と疑問を投げかけた。
* t：日本の周辺には北米プレート、太平洋プレート、フィリピン海プレート、ユーラシアプレートの四つのプレートがある。 h：日本は、四つのプレートがひしめき合う所である。
* t：言葉には「つくる楽しみ」と「使う楽しみ」の二つがあります。 h：言葉は二つの楽しみがあります。
* t：現在の震度階級は、低い方から０～４、５弱、５強、６弱、６強、７の計１０あり、全国数百カ所に設置した震度計が測っている。 h：現在の震度には１０段階の階級がある。
* t：スマトラ沖地震では、陸、海、空の３自衛隊が一体となって国際緊急援助活動に取り組んだ初のケースとなった。 h：スマトラ沖地震では、陸上自衛隊、海上自衛隊、航空自衛隊が協力して国際緊急援助に取り組んだ。
* t：茶は、茶葉をすべて発酵させる紅茶、発酵させない緑茶、半分発酵するウーロン茶の三つに分けられる。 h：お茶は発酵させるかさせないかで、発酵茶、半発酵茶、不発酵茶の３種類に分かれる。

■被限定名詞の同定（t,h の数量表現のどちらか一方が動詞修飾型 or 添加型の文のみ挙げる）

t：マウスは４万個近い遺伝子を持ち、その９５％以上が人間の遺伝子と共通すると考えられている。 h：マウスの遺伝子は人間の遺伝子と９割以上共通する。
t：韓国では男性９・３％、女性２２・３％が整形経験者であった。 h：韓国は整形をする女性が２０％以上いる。

■テンプレートに基づく立式と評価

t：原油高に伴う燃料価格の高騰を受け、国際貨物運賃の付加運賃が１キロ当たり３６円から４２円になる。 h：原油高に伴う燃料価格の高騰により、国際貨物について、付加運賃が値上げされることになった。
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。
t：現代人の１回の食事の咀嚼回数は弥生時代の６分の１以下、食事時間も５分の１になった。 h：現代人の、食物を噛む回数は昔より少なくなっている。
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。
t：文系と理系の生涯賃金の差は５０００万円である。 h：理系の生涯所得は文系より５０００万円も少ない。
t：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。

h：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。
624 語彙 (用言):含意—推論:時間軸・数量 ◎ 五羽の仔ウサギが産まれて、三羽が死んでしまった。 二羽の仔ウサギが生きている。
t：風呂に入ると、浮力で体重は９分の１になる。 h：お風呂に入ると浮力を受ける。
* t：地球儀の製造販売「オルビイス」（大阪市）によると、サッカー・ワールドカップがあった０２年は前年より約３割多く売れた。 また今、アテネ五輪で地球儀が人気を呼んでいる。 h：世界的なイベントのある年に売れ行きが良くなる実績がある。

■その他

- 数量表現の程度表現などへの言い換え

t：ブッシュ大統領は人並み外れた心臓機能の持ち主である。 h：ブッシュ大統領の心機能・循環機能の高さは、同年代の米国人男性の中で「上位１％以内」との判定を受けているという。
t：近い将来、世界は深刻な水不足になると懸念されている。 h：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。
t：人種の違いを動機とする犯罪は、英国で年２０万件以上発生している。 h：イギリスでは人種差別犯罪が後を絶たない。
t：バグダッドのイラク国立博物館で、メソポタミア文明の遺跡の発掘物など貴重な展示品十数万点が略奪されていたことが分かったが、本物と模造品を見分けて持ち出すなど、考古学に詳しい専門家が略奪に加わっていた可能性が浮上している。 h：イラク国立博物館は、略奪で壊滅的被害を受けた。
t：病院内の麻酔科医だけで麻酔業務ができる病院は４割に満たず、多くは外科などからの応援が不可欠になっている。 h：麻酔医の数が追いついていない。
160 語彙 (体言):定義的 × あの店は創業 300 年の老舗だ。 あの店は歴史が浅い。
t：携帯電話販売と通信サービスに携わるネプロジャパン（東京都中央区）が１０代から５０代までの男女５５００人を対象に、機種変更や解約によって要らなくなった携帯電話をどうするか尋ねたところ、「ショップに返却する」と回答したのは２６％にとどまった。 h：携帯電話のリサイクルが進まない。
1565 推論:結果→原因 ○ 四十日連続の真夏日を記録した。 暑い日が毎日続いている。
t：「ななめドラム」型洗濯機は乾燥時間が従来機種の約半分で済む。 h：ななめドラム型洗濯機は、時間を選ばず洗濯物を乾燥までできる。
t：富山県・黒部峡谷の宇奈月－欅平（けやきだいら）間（２０．１キロ）でトロッコ電車を運行している黒部峡谷鉄道は、落石による線路破損で出平（だしだいら）－欅平間（１１キロ）で運休している。 h：黒部峡谷鉄道のトロッコ列車は、落石事故で一部区間の不通が続いている。
* t：研修医の受け入れ数が制限され、病院では研修医が減った。 h：医師不足は、厳しさを増している。
* t：鳥インフルエンザに感染したカラスが見つかった。 h：鳥インフルエンザが鶏だけでなく、カラスにまで広がった。

- 単語がもつ数量情報（序数詞、その他）の利用、推論

1871 推論:時間軸・数量—推論:省略の類推 ○ 花子は泰子に「もしかして三人目？」と聞いた。 泰子には子供が二人いる。
t：マドンナは今回三人目を妊娠した。 h：マドンナは三人子供がいる。
1866 推論:時間軸・数量 ○ 第三次世界大戦では、多くの貴い人命が犠牲になった。

これまでに三度の世界大戦があった。
1694 推論:原因→結果 ◎ 小泉純一郎首相は 3 回目の内閣改造と党三役人事を行う。 第 3 次小泉内閣が誕生する。
1864 推論:時間軸・数量 ◎ 台風 7 号が発生し、日本に接近の恐れ。 すでに 6 つの台風が発生している。
172 語彙 (体言):定義的 × 甲子園は準決勝を迎えた。 8 チームが残っている。
610 語彙 (用言):含意—語彙 (体言):定義的 ◎ 甲子園は準決勝を迎えた。 あと 3 試合で優勝が決まる。
340 語彙 (体言):定義的—語彙 (体言):対義語 ○ 冬季オリンピックがトリノで開催される。 夏季オリンピックが開催されるのは 2 年後である。
1851 推論:時間軸・数量 × 来年は 3 回目の年女だ。 今年は 34 歳だ。
t : 応募者の中から抽選で３組６人を１０月８日、大阪空港で行う命名式と体験搭乗に招待する。 h : 応募者から抽選で３人をペアで１０月８日、大阪空港で行う命名式・体験搭乗に招待します。

- 取り立て助詞

1844 推論:時間軸・数量 × ミカンを 7 つまで数えた。 ミカンは 3 つしかない。
*t : らっきょう漬けは、原材料の食感、漬け込み前の乳酸発酵、調味液の三位一体の味だ。 h : らっきょう漬けは、調味液だけで味わいが生まれるのではない。

- その他

t : 木下大サーカスの観客数は、現在年間１２０万人で、米国リングリング・サーカスに次ぐ世界第２位だ。 h : 米国リングリング・サーカスの観客数は、世界一だ。
t : 山梨県産のミネラルウォーター生産量は日本全体の５０％と圧倒的シェアを占める。 h : 山梨県はミネラル水の生産シェア 1 位である。
324 語彙 (体言):定義的 ○ 太郎は、握力がクラスで一番だった。 太郎は、クラスの誰よりも握力が強い。
1856 推論:時間軸・数量 ◎ ミカンを 7 つまで数えた。 ミカンは 7 つ以上ある。
1930 推論:順接・逆接+具体化 ○ 日本チームは逆転勝ちしたが、4 失点した。 日本は 5 点取った。
1893 推論:順接・逆接+一般化 ◎ 卵を割ったら黄身が二つ入っていた。 一般に卵には黄身が一つである。
1860 推論:時間軸・数量 ◎ 佐々木容疑者は他に２件の犯行を自供し始めた。 佐々木は３件の犯罪を犯している。
249 語彙 (体言):定義的 ◎ 十の位が八、一の位が九だ。 八十九だ。
t : 円周率は３・１４１５９……とかぎりなくつづく数だ。 h : 円周率は３だ。

t : 日本人に不足しがちな栄養素の第一がカルシウムだという。 h : 日本人に不足がちな栄養素の代表がカルシウムだという。
t : 昨年の新入生に「学生生活に必要なもの」を尋ねた調査で携帯電話が 1 位だった。 h : 携帯電話は学生にとって必携である。
1843 推論:時間軸・数量 × このパーキングは 20 分 100 円だ。 1 時間で 400 円になる。
* t : タンク内に塩化水素を２０％含む洗浄剤７リットルをじょうろでまいた。 h : タンク内に塩化水素７リットルをまいた。
* t : 福島県いわき市。 面積が１２３１平方キロと「日本一広い市」だ。 h : 世界一面積が広いのはいわき市だ。
* t : 国内で１年間に使われる割りばしは２８０億ぜんだ。 h : 割り箸を数えるときは「ぜん」です。

付 録B　規格化のための辞書

「助数詞」部分の辞書（右側は変換される単位。基本的にそのまま変換されるだが、「ニュートン」のような単位は、より一般的な「N」に変換される）

ニュートン	N
カナダドル	カナダドル
シンガポールドル	シンガポールドル
オクターブ	オクターブ
パーセント	%
オクターヴ	オクターブ
メートル	m
リットル	l
デシベル	デシベル
ピクセル	ピクセル
アンペア	A
ケルビン	K
カンデラ	cd
フィート	フィート
カラット	カラット
カロリー	cal
US ドル	ドル
タイトル	タイトル
シリーズ	シリーズ
ポイント	ポイント
ジャンル	ジャンル
ステージ	ステージ
パターン	パターン
ラウンド	ラウンド
グラム	g
ビット	ビット
バイト	バイト
フラン	フラン
クローネ	クローネ
ポンド	ポンド
ユーロ	ユーロ
レース	レース
ウォン	ウォン
ルピー	ルピー
インチ	インチ
ノット	ノット
モーラ	モーラ
コース	コース
ページ	ページ
テイク	テイク
タイプ	タイプ

ゲーム	ゲーム
チーム	チーム
ペソ	ペソ
選手	選手
拍子	拍子
音節	音節
区画	区画
切れ	切れ
人前	人前
ドル	ドル
モル	mol
海里	海里
ペア	ペア
トン	トン
種類	種類
集落	集落
手法	手法
言語	言語
地域	地域
議席	議席
カ国	カ国
ヶ国	カ国
か国	カ国
ケ国	カ国
ヵ国	カ国
年分	年分
民族	民族
種目	種目
分割	分割
母音	母音
箇所	箇所
ケ所	箇所
ヶ所	箇所
ヵ所	箇所
カ所	箇所
か所	箇所
個所	箇所
文節	文節
年生	年生
回生	回生
単位	単位

次元	次元
番地	番地
連勝	連勝
連敗	連敗
重奏	重奏
年制	年制
試合	試合
文字	文字
作品	作品
世代	世代
大会	大会
得点	得点
方向	方向
店舗	店舗
世帯	世帯
師団	師団
艦隊	艦隊
要素	要素
領域	領域
音素	音素
段階	段階
連隊	連隊
階級	階級
連覇	連覇
路線	路線
bite	バイト
便	便
勝	勝
敗	敗
等	等
人	人
個	個
つ	つ
枚	枚
面	面
段	段
本	本
匹	匹
羽	羽
灯	灯
頭	頭

本	本
張	張
戸	戸
軒	軒
棟	棟
杯	杯
匹	匹
枚	枚
架	架
体	体
柱	柱
府	府
党	党
氏	氏
団体	団体
局	局
番	番
脚	脚
本	本
基	基
着	着
具	具
羽	羽
頭	頭
席	席
献	献
柄	柄
玉	玉
杯	杯
巻	巻
枝	枝
尾	尾
港	港
掛	掛
番	番
封	封
筋	筋
挺	挺
条	条
錠	錠
丈	丈

幅	幅
株	株
刎	刎
座	座
騎	騎
行	行
服	服
包	包
果	果
菓	菓
足	足
領	領
丁	丁
俵	俵
膳	膳
喉	喉
斤	斤
叺	叺
貫	貫
篇	篇
尊	尊
棹	棹
台	台
両	両
連	連
部	部
頁	ページ
球	球
部	部
句	句
門	門
問	問
戦	戦
畳	畳
棹	棹
反	反
卓	卓
口	口
壷	壷
通	通
振	振

腰	腰
剣	剣
刀	刀
票	票
帖	帖
句	句
輪	輪
片	片
機	機
名	名
拍	拍
躯	躯
隻	隻
粒	粒
顆	顆
札	札
冊	冊
品	品
mol	mol
℃	℃
rad	rad
円	円
割	割
種	種類
級	級
度	度
こ	個
倍	倍
%	%
回	回
日	日
弦	弦
校	校
次	次
項	項
歳	歳
才	歳
国	国
州	州
件	件
区	区

話	話
選	選
位	位
合	合
階	階
銭	銭
波	波
節	節
bit	ビット
bps	bps
期	期
切	切
音	音
手	手
尺	尺
寸	寸
県	県
章	章
泊	泊
曲	曲
列	駅
線	線
社	社
弾	弾
組	組
役	役
桁	桁
字	字
点	点
店	店
石	石
厘	厘
版	版
藩	藩
号	号
課	課
作	作
軍	軍
集	集
州	州
周	周

m/h	m/h
m/s	m/s
m/m	m/m
rpm	rpm
項	項
cd	cd
Pa	Pa
Ω	Ω
Wb	Wb
Hz	Hz
sr	sr
ha	ha
°	度
cc	cc
m	m
g	g
N	N
l	l
s	s
A	A
K	K
J	J
W	W
C	C
V	V
F	F
S	S
T	T
H	H
%	%
t	トン

「特殊」部分の辞書。右側が対応する処理。

時速	/h
毎時	/h
分速	/m
毎分	/m
秒速	/s
毎秒	/s
風速	fusoku

「接尾辞」部分の辞書。右側が対応する処理。

くらい	about
ばかり	about
以上	＜＝
以下	＞＝
未満	＞
前後	about
程度	about
ほど	about
前半	zenhan
後半	kouhan
近く	about
余り	kyou
強	kyou
弱	jaku
以降	＜＝
頃	about
超	＜＝
台	dai
代目	daime
代	dai
半ば	nakaba
数万	suuman
数億	suuoku

「接頭辞」部分の辞書。 右側が対応する処理。

約	about
ちょうど	none
だいたい	about
ほぼ	about
およそ	about
ほとんど	about
数	some
何	some
全	none
地下	tika
海抜	kaibatsu