供述調書に現れる数量表現の推論テストセットの構築

小谷野華那[1]　谷中瞳[2]　峯島宏次[3]　福田浩司[4]　橋爪宏典[5]　戸次大介[1]
[1] お茶の水女子大学　[2] 東京大学　[3] 慶應義塾大学
[4] 日本電気株式会社 [5]NEC ソリューションイノベータ株式会社
{koyano.kana,bekki}@is.ocha.ac.jp　hyanaka@is.s.u-tokyo.ac.jp
minesima@abelard.flet.keio.ac.jp　{h.fukuda,h_hashidume}@nec.com

概要

昨今の自然言語処理の発展により、刑事手続きにも意味解析の応用が試みられている。供述調書では頻出する数量表現の意味を正しく処理することが求められる。また数量表現が現れる文の推論では、含意と推意の間で判定が異なり、平叙文と否定文、条件文では含意関係が反転する場合がある。そこで本研究では、供述調書に現れる数量表現に対して、高度な意味アノテーションを付与した数量表現コーパス、および数量表現の推論テストセットの構築を行い、ベースライン実験による評価を行なった。

1　はじめに

昨今の自然言語処理の発展により、刑事手続きの文書にも意味解析技術の応用が試みられている。供述調書には数量表現が頻出するため、数量表現の意味を正しく処理することは特に重要である。そのような用途に供する言語処理技術を開発し、正しく評価するために、実テキストの数量表現の理解を問うデータセットが求められている。

含意関係認識（自然言語推論）は、前提文が真であるとき、仮説文が必ず真（＝含意）か、必ず偽（＝矛盾）か、どちらともいえない（＝中立）かを判定するタスクであり、テキスト間の言語理解の基礎をなすタスクの一つである。推論を構成する関係には、含意 (entailment) と推意 (implicature) がある。例えば、次の前提文 (1) と仮説文 (2) を考えよう。

(1)　男性が道端に 4 人座っていた。
(2)　男性が道端に 5 人座っていた。

(1) を「少なくとも 4 人座っていた」と解釈すればこの推論の含意ラベルは中立になる。しかし、Grice の量の格率 [1] より、もし話者が 5 人座っていたことを知っていたとすればそう言うはずなので、わざわざ（1）のように述べたということは「5 人目は存在しない」と考えられ、推意ラベルは矛盾になる。

このように数量表現が現れる文間の推論では数詞の違いや数量表現の用法によって含意と推意の間で判断が異なるため、両者を区別して考える必要がある。また、平叙文と否定文、条件文では判定ラベルが反転するという特徴もある。

そこで本研究では、刑事手続きに関連する実テキストのうち数量表現を含む文と、NPCMJ[2] から抽出した否定文、条件文を用いて、数量表現の分類や用法などについて高度な意味アノテーションを付与した数量表現コーパスを構築する。さらに、数量表現コーパスをもとに、数量表現の推論テストセットを構築する。本稿では、数量表現コーパスと推論テストセットの設計と、ベースライン実験の評価結果について報告する。構築したデータセットは、研究利用可能な形式で公開する予定である。

2　関連研究

英語では、数量表現を含む推論テストセット [3] があり、含意、矛盾、中立の 3 つのカテゴリに 2,532 件ずつ計 7,596 件の文ペアが含まれている。しかしこの推論テストセットに対しては、数量表現の推論が単純なテンプレートに基づいて構築されているため、いくつかのヒューリスティックスで大部分の問題（全体の約 82%）を解くことができてしまうという批判がある [4]。また、英語の scalar implicature のデータセットとしては [5] があるが、このデータセットはテンプレートから自動で構築されており、文は比較的単純なものが多い。

日本語の推論データセットとしては、形式意味論テストセットの JSeM [6]、英語 SNLI [7] の日本語版である JSNLI [8]、英語 SICK [9] の日本語版である JSICK [10]、旅行情報サイトの評判という実テキストからクラウドソーシングで構築された JRTEC [11]

などがあるが、日本語の数量表現の統語的・意味的な多様性は十分に考慮されていない。

成澤ら [12] は、日本語の含意関係認識において数量表現が問題になる事例に焦点を当て分析を行い、数量表現の規格化のためのモジュール実装と評価を行なった。成澤らは、数量表現が出現する文ペアを 7 つのカテゴリに分類し、正しく含意関係を判定するために必要な処理について述べているが、数量表現自体の分類や、数詞の違いによる含意ラベルと推意ラベルの判定の違いについては言及していない。

本研究では、実テキストにおける日本語の数量表現に高度な意味アノテーションを付与した数量表現コーパスを構築し、そのコーパスをもとに推論ペアを作成し、含意と推意の両方のラベルを付与した推論テストセットを構築する。

3 数量表現の意味と推論

3.1 数量表現の分類

助数辞の分類 飯田 [13] によると、助数辞は分類辞、単位形成辞、計量辞の 3 つに分類される。これに加えて、数量表現には時間や系列の中の順序を表す「3 月」「1 番」のような序数辞（順序数詞）[14] を伴うものがある。そこで、本研究では、飯田の 3 分類と序数辞を合わせた 4 分類を用いる。各助数辞の例を表 1 に示す。

表 1 各分類ごとの助数辞の例とアノテーション件数

タイプ	例	件数
分類辞	人、頭、冊、枚	360
単位形成辞	瓶、箱、袋、パック、切れ	18
計量辞	リットル、円、バイト	293
序数辞	月、日、番、位	23

助数辞の分類には、表層形から一意に定まらず、文脈や用法によって分類が変わる助数辞も存在する。たとえば、「会議室は建物の 3 階にある」に現れる「階」は序数辞であるが、「ここから 3 階のぼったところに会議室がある」に現れる「階」は計量辞である。前者は特定の位置を指しているのに対し、後者は 3 フロア分上の階に上がるという意味であり、会議室が 3 階に位置しているという意味ではない。

数量表現の出現位置 岩田 [15] は、名詞 N についてその数とカテゴリー情報を数量表現 Q が表すものについて、日本語百科事典で定義されている Q ノ NC 型、N ノ QC 型、NCQ 型、NQC 型の 4 分類に加えて、述部型、デ格型の 2 分類を新たに追加している。本研究ではこれに加えて、新たに数量表現の出現位置として、「3 回行った」のように動詞 V を意味的に修飾する QV 型、イベント名詞句 Nv を修飾する NvCQ 型、いわゆる代名詞的用法の「3 人は」など N が脱落しているもの、「1 時間（で）500 円」「1 ヶ月に 1 回」のように時間表現と数量詞が連続する QtQ 型、イディオム・慣習的用法を追加し、数量表現の出現位置について分析を行なった。出現位置の例を表 2 に示す。

表 2 出現位置の例とアノテーション件数

タイプ	例	件数
Q ノ NC 型	3 人の学生が来た	150
N ノ QC 型	学生の 3 人が来た	14
NCQ 型	学生が 3 人来た	161
NQC 型	学生 3 人が来た	64
デ格型	学生が 3 人で来た	43
QV 型	東京に 3 回行った	163
NvCQ 型	渡米したことは 2 回ある	7
N の脱落	3 人はお金を払った	65
QtQ 型	1 時間 500 円かかる	6
イディオム的	1 人暮らし, 8 人兄弟	21

数量表現の用法 岩田が研究対象とした数量表現 Q の用法に加えて、本研究では、名詞 N を修飾する数量表現 Q の用法について新たに 3 つの分類を追加し、動詞 V を修飾する数量表現 Q の用法として 3 つの分類を追加した。さらに、Nv を修飾する数量表現 Q の用法とイディオム用法を追加した。用法の分類とその例を表 3 に示す。

表 3 用法の分類の例とアノテーション件数

タイプ	例	件数
Q が N のカテゴリー情報を表すもの	3 人の学生	377
Q が N を構成する要素の全体数を表すもの	家族 3 人	8
Q が N を構成する要素の一部を表すもの	集団の 1 人	2
Q が N の属性や特徴を表すもの	50 歳の男性	110
V が行われた回数を表す Q	2 回来る	74
V が行われた期間を表す Q	3 日滞在する	61
V が行われた時間を表す Q	9 時に来る	34
Nv を修飾する Q	渡航歴は 2 回	7
イディオム的用法	1 人暮らし	21

3.2 数量表現の推論

意味論と語用論の区別 推論を構成する関係には含意と推意が存在し、含意ラベルと推意ラベルで判断が異なる場合がある。含意では、前提文と仮説文に意味論的に含まれる情報のみから推論を行なっている。これに対して推意では、通常の会話が協調の原則に従って進行していることから、Grice の会話

の格率 [1] にみられるように、文脈や発話者の意図を考慮して、発話に含まれていない情報でも成立すると推測することがある。そのため、推意は含意と異なる判定になる場合がある。

下方含意文脈　M が N の下位概念とすると、通常は下位概念を含む文 $\varphi(M)$ は上位概念を含む文 $\varphi(N)$ を含意する。数量表現の場合、例えば、「200 人」は「100 人」の下位概念であり、「会場に 200 人いる」が真であれば「会場に 100 人いる」も真である。しかし、否定文や条件文の前件などのいわゆる下方含意文脈では、含意関係が反転し、上位概念を含む文 $\varphi(N)$ が下位概念を含む文 $\varphi(M)$ を含意する場合がある。例えば、「会場に 100 人はいなかった」は「会場に 200 人はいなかった」を含意する。

4　数量表現のアノテーション

4.1　供述調書テキストデータ

刑事手続きの実テキストとして、実際の事件記録に沿った教材として作成された法務総合研究所作成の事件記録教材を用いる [16, 17]。本教材には、一冊ごとに、一つの事件を対象として刑事手続上で作成される事件記録が掲載されている。事件の発生時期・場所、登場する人物、団体、地名等は，いずれも実際の事件と関係のない架空のものであるが、取扱状況報告書、被害者、加害者に対する供述調書、被害届、被害状況確認報告書などが、実際の事件記録に沿った形式で編集されているので、検察官が供述間の整合性の確認や矛盾有無を議論する過程を模擬的に検討することが可能である。

本研究でサンプリングしたデータは被害者、加害者に対する供述調書が中心であるが、被害者と加害者間の供述内容が不整合であるケースを取り上げている。特に、数量表現の不一致にもとづく不整合が多くみられるため、数量表現間の正確な推論が重要となる応用領域の一つである。

4.2　数量表現コーパスの構築

本研究では、供述調書テキストデータに含まれる数量表現 613 件と NPCMJ[2] から抽出した否定文、条件文に含まれる数量表現 81 件の計 694 件について、言語学の素養のある大学院生 1 名がアノテーションを行なった。

タグ付け　文中に現れる数量表現に<num>タグを付与し、3 章で述べた助数辞の分類、出現位置、用法についてアノテーションを行った。文中に複数の数量表現が現れる場合は、数量表現 1 つに<num>タグを付与した。また、数量表現のうち、代名詞的用法、数詞 1 を含むもの、遊離数量詞、一定期間に関する数量表現、数量表現が修飾する名詞がなく裸で出てくるもの、「のうち」や「含め」などの要素数に関する記述があるもの、数量詞修飾がついているもの、イディオム、漢数字、否定表現について、それぞれ区別できるようアノテーションを行なった。

コーパスに含まれる助数辞の分類は表 1、出現位置は表 2、用法は表 3 に示したものを使用した。

5　数量表現の推論テストセット

5.1　推論テストセットの作り方

前提文は、4 節でアノテーションを行なった文を用い、仮説文は（3）のように<num>タグが付与されている数量表現について（4）のように意味を変えない最小の節を取り出し、数詞の変更を行い数量詞修飾を付与することで作成した。

(3)　私は，20 年位前にスリで<num>3 回</num>位警察に逮捕され，1 度は刑務所に入ったこともありましたが，その後は真面目に生活していました。

(4)　私は、スリで 10 回以上警察に逮捕された。

前提文 T 中の数詞に対して、数詞が 20 未満の場合は、マイナス 1 の数詞に置き換えた仮説文 H_- と、プラス 1 の数詞に置き換えた仮説文 H_+ の 2 つを作成し、推論テストセットを構築した。数詞が 20 以上の前提文に対しては、数詞をプラスマイナス 5 ずつ行い、H_-H_+ を作成した。数詞が大きいケース（例：2 万 5000 円）については、一番大きい桁（例：2 万）の数字を基準にプラスマイナス 1 を行い仮説文を作成した。表 4 の 1 つ目の例のように、数量表現に「くらい」などの数量詞修飾がついている場合は、上記のルールに従って仮説文を作成すると判定ラベルが全て中立になってしまうため、そのような場合は非文にならない範囲で推意ラベルがなるべく中立にならないように数字を変更した。また、表 4 の 2 つ目のように、数量詞修飾を追加すると非文になる場合は、意味が変わらないように語順を入れ替えて仮説文を作成した。本研究では、イディオム的用法は数詞の変更や数量詞修飾を追加すると非文になるため、推論テストセットからは除外している。

表 4　推論テストセットの例

前提文（T）と仮説文（H_-、 H_+）	正解ラベル $T \Rightarrow H_-$ 含意ラベル	推意ラベル	$T \Rightarrow H_+$ 含意ラベル	推意ラベル
T: 河川敷では 10 人くらいの男性がバーベキューをしていた。 H_-: 河川敷では 2 人以上の男性がバーベキューをしていた。 H_+: 河川敷では 20 人くらいの男性がバーベキューをしていた。	含意	含意	中立	矛盾
T: 私の家族は佐賀県に父宗男 70 歳，母とく子 70 歳が住んでおり，兄が福岡市，弟が広島の南区に住んでいます。 H_-: 私の家族は佐賀県に 65 歳以上の父と母が住んでいます。 H_+: 私の家族は佐賀県に 75 歳以上の父と母が住んでいます。	含意	含意	中立	矛盾
T: 勿論、私ひとりで四升呑みほしたわけでは無い。 H_-: 勿論、私ひとりで三升以上呑みほしたわけでは無い。 H_+: 勿論、私ひとりで五升以上呑みほしたわけでは無い。	中立	中立	含意	含意
T: あと 1000 万円あれば何とか欲しい家が買える。 H_-: あと 950 万円以上あれば何とか欲しい家が買える。 H_+: あと 1050 万円以上あれば何とか欲しい家が買える。	中立	中立	含意	含意

表 5　日本語 BERT を用いたベースライン実験の結果（正答率）

学習データ	含意ラベル 全体	含意	矛盾	中立	推意ラベル 全体	含意	矛盾	中立
JSICK	36.29%	61.28%	4.51%	28.80%	28.38%	61.22%	6.4%	2.12%
JSNLI	42.88%	71.75%	37.70%	17.00%	43.77%	71.43%	26.06%	23.91%

上記のルールで作成した推論テストセットに対して、アノテータ 1 名が判定ラベルを付与した。

5.2　テストセットの概要

本研究で作成した推論データセットには 1124 件の文ペアが含まれており、各文ペアに対して含意ラベルと推意ラベルをそれぞれ付与している。推論テストセットの統計情報を表 6 に、前提文と仮説文の例を表 4 に示す。含意ラベルで中立になるものについて、推意ラベルでは矛盾になる場合があるため、各ラベルの矛盾と中立の数が異なっている。

表 6　推論テストセットの統計情報

	含意	矛盾	中立
含意ラベル	439	244	441
推意ラベル	441	637	46

5.3　ベースライン実験

現状の標準的な言語処理技術がどの程度、数量表現の理解を必要とする推論を扱えるのかを評価するため、日本語 BERT [18] をベースラインモデルとして評価実験を行った。実験では、学習データとして JSICK（5000 件）、JSNLI（53 万件）を用いた。表 5 に含意関係認識モデルの評価結果を示す。全体としては JSICK を学習した場合よりも JSNLI を学習した場合の方が日本語 BERT の正答率が高い傾向があるが、両方とも 5 割を下回っていた。とくに、含意の正答率が 6 割を超えたのに対して、矛盾と中立の正答率は 4 割を下回ったことから、既存のデータセットでモデルを学習すると含意と予測してしまう傾向が示唆される。学習データの違いによる正答率の違いとしては、含意ラベル、推意ラベルの両方とも JSICK より JSNLI を用いた場合の方が矛盾の正答率が上がっており、これは JSNLI の方が学習データの件数が多いことが理由として考えられる。

6　おわりに

本研究では、供述調書内の数量表現を含む文と NPCMJ から抽出した否定文、条件文を用いて、数量表現の分類や用法などの高度な意味アノテーションを付与した数量表現コーパスの構築と推論テストセットの構築を行った。ベースライン実験として日本語 BERT の評価を行い、本研究で構築した数量表現の推論テストセットが、現状の言語処理技術において挑戦的な課題であることを確認した。今後は、数量表現コーパスと推論データセットの拡張を進めるとともに、構築したデータセットを用いて現在の含意関係認識システムの分析と評価を進めていく。

謝辞　本研究の一部は、日本電気株式会社・お茶の水女子大学の共同研究「捜査資料内のテキスト意味解析に関する研究」、JST CREST JPMJCR20D2、

JST さきがけ JPMJPR21C8 の支援を受けたものである。

参考文献

[1] Stephen Levinson. **Pragmatics**. Cambridge University Press, 1983. （S. レヴィンソン『英語語用論』, 安井稔・ 奥田夏子訳, 研究社出版, 1990）.

[2] NINJAL. NINJAL Parsed Corpus of Modern Japanese. (Version 1.0). Technical report, National Institute for Japanese Language and Linguistics, 2016. https://npcmj.ninjal.ac.jp/.

[3] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 2340–2353, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[4] Nelson F. Liu, Roy Schwartz, and Noah A. Smith. Inoculation by fine-tuning: A method for analyzing challenge datasets. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 2171–2179, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8690–8705, 2020.

[6] Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In **New Frontiers in Artificial Intelligence**, pp. 58–65, 2017.

[7] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, 2015.

[8] 吉越卓見, 河原大輔, 黒橋禎夫. 機械翻訳を用いた自然言語推論データセットの多言語化. 第 244 回自然言語処理研究会, 2020.

[9] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 216–223, 2014.

[10] 谷中瞳, 峯島宏次. JSICK: 日本語構成的推論・類似度データセットの構築. 第 35 回人工知能学会全国大会, 2021.

[11] Yuta Hayashibe. Japanese realistic textual entailment corpus. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 6827–6834, 2020.

[12] 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 発表論文集, 2012.

[13] 飯田隆. 日本語と論理. NHK 出版, 2019.

[14] 奥津敬一郎. 拾遺日本文法論. ひつじ書房, 1996.

[15] 岩田一成. 日本語数量詞の諸相. くろしお出版, 2013.

[16] 法務総合研究所. 事件記録教材: 法科大学院教材. 第 15 号 (窃盗被疑事件). 法曹会, 2014.

[17] 法務総合研究所. 事件記録教材: 法科大学院教材. 第 15 号 (暴行機凝事件). 法曹会, 2014.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, 2019.