保険約款に対する派生文書の矛盾認識

丹治 広樹 , 山本 和英
長岡技術科学大学 電気系
E-mail: {tanji , yamamoto}@jnlp.org

1 はじめに

近年、電子データの増加とともに自然言語処理の活用される場が増えてきた。しかし、保険や金融等の分野では電子テキストデータが増加しているにもかかわらず、未だに校正の大部分を人手により行っている。保険関連の文書には、約款や特約等の基礎書類（以下、基本文書とする。）を流用して消費者向けに改変された文書（以下、派生文書とする。）が多数存在する（図1）。派生文書は基本文書を参照して人手により再度入力される場合もあり、矛盾や表記ミス等が含まれることが多い。そのため、基本文書と矛盾する内容を含んだ文を修正する作業が必要になるが、人手によって認識するには多大な労力と時間がかかる。このように、これらの分野では自動で処理することにより効率化できる部分が多く残されている。

そこで、我々は基本文書と派生文書との間で矛盾がないか半自動で確認し、人手による作業を減らすことを目指す。本稿では、簡単な手法で矛盾をどの程度認識できるか調べるためにパタンマッチングによる基準を作成し、保険関連の文書の矛盾認識を試みた。その結果、基本文書と異なる表現や数値の矛盾、否定表現が一致していない文を抽出できた。

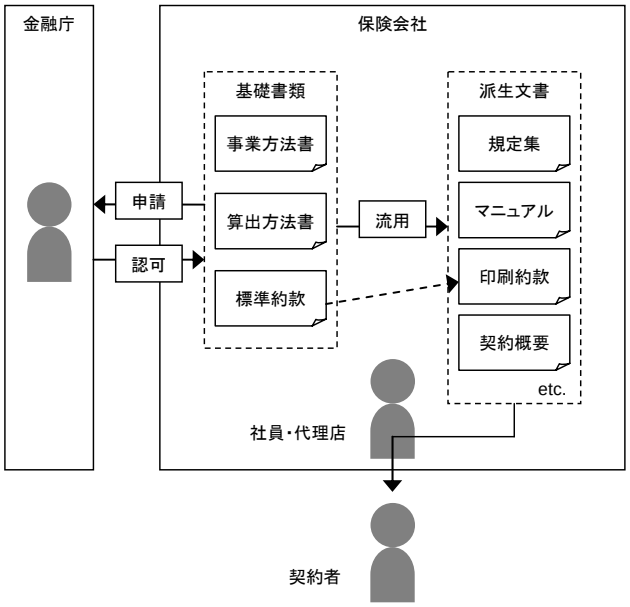

図 1： 基本文書と派生文書の関係

2 関連研究

保険約款に対する言語処理の研究はこれまでに存在しなかった。しかし、保険約款は文体や条項の体系等で法律文と類似した点が多く見られる。法律文を扱っている研究として、岩本ら[1]の研究がある。岩本らは単語頻度の計数等を行い、文章・語彙・構文・文脈・論理構造の観点から法律文の傾向を整理している。高頻度な単語は名詞が多い点、推量を表す文は存在しない点等に保険約款との共通点が見られた。

法律文の特徴を加味した構文解析に関する研究として、山田ら[2]の研究がある。山田らは、法律文コーパスを作成するにあたって、各文に構文情報のタグを付与している。タグのひとつとして、図 2 の例のような括弧表現を含む文について括弧の内外の文字列は各々独立して文を成すとしている。ここで、図 2 中の矢印は係り受け関係を示す。保険関連の文書でも括弧表現を含む文は頻出している。そこで、本研究では山田らの手法に従い、括弧表現の内外で独立に係り受け解析を行った。

矛盾や対立の認識を行った研究として、松吉ら[3]や乾[4]の研究がある。松吉らは事前に収集、整備された事態間関係知識および反義関係知識、動詞項構造シソーラス等を用いて対立関係を認識している。乾は含意関係認識において意味的な情報を重視し、命題に対する書き手の態度（ムード情報）を考慮している。さらに従来の機能表現辞書を使用することを検討している。これらの研究のように知識を収拾し、辞書化するには多大な手間および時間がかかる。本研究では、より簡単なパタンを用いて保険約款とその派生文書に特化した処理を行った。

Li et al.[5]は含意認識(True Entailment Recognition)および非含意認識(False Entailment Recognition)の判定基準を提案している。含意認識は「仮説を言い換えたものがテキストと一致する、またはテキストに含まれる」という考え方に基づき、テキストと一致する言い換えを探す。それに対して、非含意認識は「テキストと仮説は一致しない」という考え方に基づ

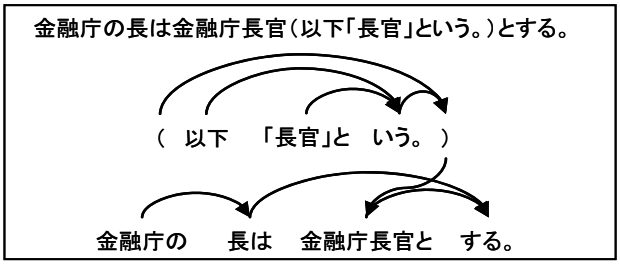

図 2： 括弧表現を含む文の構文解析[2]

き、仮説中のテキストと一致しない箇所を探す。よって、本研究は矛盾を認識するタスクであるため非含意認識に類似しているといえる。Li et al. は、語彙知識や構文情報を用いてテキストと仮説の一致をとる含意認識に対し、非含意認識において数値や場所、否定表現等の不一致を見ることで矛盾を検出している。Li et al. が提案した非含意認識の基準は以下のとおりである。

- 数値の不一致
- 時間や日付の不一致
- 場所の不一致
- 数量詞の不一致
- “Say”関係の不一致
- “Locate”関係の不一致
- 否定および仮定の不一致

3 使用した保険関連文書とその特徴

保険関連の文書には、前述したように基本文書と派生文書の2種類が存在する。基本文書は、省庁に提出する必要があり、法律文に近い傾向をもつ約款や特約等の文書である。章・条・項で区分されており、文末はですます調となっている。ただし、箇条書きの場合は体言止めである。派生文書は、基本文書を基として消費者向けに文章を変えたり、抜粋したりしているパンフレットや契約概要のような文書である。視覚的な読みやすさを考慮しているため、箇条書きや表中に文章を収めたものが多用されている。

本研究では、基本文書として家族傷害保険の普通保険約款および特約条項、派生文書として同保険の契約概要および注意喚起情報の文書を用いた。

4 矛盾認識の基準

保険関連の文書では、基本文書と派生文書との間で情報に差異が生じたり、基本文書に記述されていない項目が派生文書に掲載されたりしてはいけない。また、一部を除いて派生文書で使用されている単語は基本文書でも使用されている。よって、基本文書に出現した単語が派生文書で類義語や異表記に書き換えられていても修正の対象となる。

これらをもとに、本稿における矛盾を含む文とは、基本文書と派生文書との間に情報の差異が生じている文、または基本文書と派生文書で使用されている単語が異なっている文と定義する。本稿では、派生文書が基本文書に出現していない形態素や係り受けを含んでいた場合に矛盾であると判定した。

保険関連の文書において、場所の情報や伝聞を表す表現は滅多に出現しない。しかし、数値や日付、否定の表現は頻出している。よって、本稿では Li et al.の非含意認識の手法を参考にして基準を追加・削除し、以下の基準を用いて矛盾認識を行った。

4.1 数値の不一致

保険関連の文書では、金額や倍率等の数値を扱っている。このとき、基本文書に掲載されていない金額や倍率が派生文書にあってはならない。よって、派生文書に出現した数値が基本文書に存在しなかった場合、基本文書と派生文書は矛盾していると定義する。例えば派生文書に「100 倍」という数値が出現した場合、「100」という数値が基本文書に掲載されていなければ「100 倍」が含まれている文を矛盾と判定した。

4.2 時間や日付の不一致

4.1 節と同様に、入院日数や保険期間等の日付や時間を表す語も頻出する。よって、派生文書に出現した日付や時間が基本文書に存在しなかった場合、基本文書と派生文書は矛盾していると定義する。例えば派生文書に「188 日」や「17:00」といった表現が出現した場合、これらの時間や日付が基本文書に掲載されていなければ「188 日」や「17:00」が含まれている文を矛盾と判定した。

4.3 単語の不一致

保険関連の文書において、一部を除き基本文書で使用されていない単語が派生文書に出現することはない。よって、派生文書に出現した単語が基本文書に存在しなかった場合、派生文書と基本文書は矛盾していると解釈する。例えば、基本文書では「本人」が使用されているが派生文書では「自身」を使用していた場合、「自身」が含まれる文を矛盾と判定した。

ただし、例 1 に挙げるような特定の単語 90 種を例外とした。これらは派生文書で消費者向けの説明を行う際に使用されるものであり、少数かつ汎用的な表現であるため人手で収集して本基準の対象外とした。また、派生文書では基本文書で使用されていない敬語表現や、くだけた表現が使用される場合がある。そこで、例 2 に挙げるような表現 26 種を例外とした。

商品　本書　別紙　下表　連絡窓口

例 1： 派生文書のみに出現する単語の例

お支払いする　まいります　わかりやすく

例 2： 基本文書に出現しない敬語等の表現の例

4.4 「すべて／いずれか」の不一致

保険関連の文書において、「すべて」や「いずれか」が使用されることがある。Li et al. の手法では、“all”や“never”等の数量詞について極性が反転している場合に非含意であると判定していた。この手法では、“all”や“every”と“each”との差は考慮されていない。しかし、保険関連の文書において「すべて」と「いずれか」の差異は大きい。よって、派生文書に「すべて」が出現したときには『すべて→内容語』、「いずれか」が出現したときには『いずれか→内容語』という係り受けが基本文

書に存在しなかった場合、派生文書と基本文書は矛盾していると解釈する。例として、派生文書に「すべての関節の機能を失っていること。」という文が出現した場合を考える。「すべての」は「関節の」に係っているため、基本文書に『すべて→関節』がなければ矛盾と判定した。ただし、本稿において『A→B』はAを含む文節(係り元)とBを含む文節(係り先)との係り受け関係を表す。

4.5 否定表現の不一致

Li et al. は、テキストと仮説のどちらか一方にしか否定表現がない場合に非含意であるとしている。これをもとに、派生文書に否定表現「ない／ぬ／ん」が出現したとき、『内容語→否定表現』が基本文書に存在しなかった場合、派生文書と基本文書は矛盾していると解釈する。例として、派生文書に「傷害保険は、就業中の事故は保険金のお支払いの対象となりません。」という文が出現した場合を考える。この文の係り受け関係を図 4 に示す。「傷害保険は」、「事故は」、「対象と」の3つが否定表現を含む「なりません」に係っている。よって、基本文書に『傷害保険→否定表現』、『事故→否定表現』、『対象→否定表現』の3つがなければ矛盾と判定した。

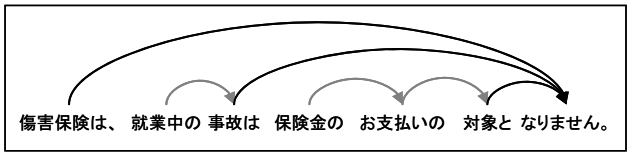

図 4： 例文の係り受け関係

5 評価実験

基本文書として保険約款および特約 794 文、派生文書として契約概要および注意喚起情報 224 文を用いて実験を行った。本研究で使用した文書は実際の保険文書をテキスト形式に直したものであり、図のタイトルや表中の文も存在する。派生文書には矛盾を含んだ 82 文が混ざっている。派生文書の矛盾は、保険関連文書の従事者が実際に起こりうる矛盾をもとに作成した。基本文書と派生文書の間に生じる矛盾の例を表 1 に示す。

4 章の基準を適用して矛盾認識を行った結果、102 文を矛盾と認識し、このうち 70 文が正解であった。このときの再現率、適合率、F値を表 2 に示す。保険関連の文書は金融庁や消費者の手に渡るものであるため、矛盾が残されていてはいけない。よって、本研究においては再現率が重要だと考えている。

本稿において、単語の抽出には形態素解析器「茶筌」(1)、係り受け解析には構文解析器「南瓜」(2)を使用した。これらの解析はIPA品詞体系(3)に基づく。ただし、山田らの手法をもとに括弧表現を含む文は括弧の内外で独立に係り受け解析を行った。

表 1： 矛盾の例

基本文書	派生文書	矛盾の理由
本人	自身	単語の不一致
80%	88%	数値の不一致
該当する場合	該当しない場合	否定表現の不一致

表 2： 矛盾認識の結果

再現率	適合率	F値
0.854 (70/82)	0.686 (70/102)	0.761

6 考察と検討

矛盾を正確に認識できた例として、以下のようなものが挙げられる。ただし、丸括弧内の表現は基本文書で使用されていた単語である。

- 基本文書に出現しない単語を含む文
  「自身(本人)」、「試験運転(試運転)」等の類義語や表記ゆれを獲得することができた。
- 基本文書で明記されていない数値を含む文
  「188日(180 日)」、「88%(80%)」等の数値の間違いによる矛盾を認識することができた。
- 否定表現により意味が反転している文
  「業務遂行上必要な範囲で、委託先に取扱いを委託する場合」のように、否定表現が含まれず、必ず「する場合」になるような文がある。これらが「しない場合」と意味的に反転している文を矛盾として認識することができた。

矛盾していない文を矛盾と認識してしまった例として、以下のようなものが挙げられる。

- 具体例を含んだ文
  派生文書では、消費者向けに具体的な事例を交えて説明する場合がある。本手法では、「国内や海外旅行中に足を骨折した。」、「○山岳登はん(専用の登山用具を使用するもの)、ボブスレー、スカイダイビング、ハンググライダー搭乗等、特に危険なスポーツをしている間」等の具体例を含む文を矛盾と認識してしまった。「海外旅行」や「ボブスレー」等の具体例が基本文書に掲載されていなかったためである。

  この問題は、文の類似度や文中の単語の一致率等を用いて基本文書と関連の薄い文をあらかじめ除外しておく方法により解決できると考えられる。関連の薄い文を除外することで、具体例のように矛盾認識する必要のない文を事前に対象外とすることが可能だと考える。

矛盾している文を矛盾と認識できなかった例として、以下のようなものが挙げられる。

- 否定と肯定の両方をもつ係り受けが含まれる文

  基本文書には「保険金を支払います」と「保険金を支払いません」の両方が含まれている。そのため、派生文書で誤った箇所に『保険金→否定表現』が出現しても矛盾と認識できなかった。

  この問題は、前後の文脈や共起語を考慮することで解決できると考えられる。例えば、「被保険者が急激かつ偶然な外来の事故によってその身体に被った傷害」の場合は保険金を支払うが、「保険料領収前に生じた事故による傷害」の場合は保険金を支払わない。否定表現の周辺を見ることにより、「急激かつ偶然」や「保険料領収前」等の特徴的な表現から否定表現が使用される場面を限定できると考える。

本実験で使用した派生文書には出現していないが、現状の手法では以下のような矛盾についても認識できないと考えられる。よって、今後は多様な矛盾に対応するためにより多くの派生文書を使用して実験する必要があると考えている。

- 箇条書きにより対象が複数文にまたがっている文

  基本文書に「次の各号のいずれかを」という表現を含む文がある。この文は次に書かれている箇条書きを参照している。それに対して、派生文書では「A、B、Cのいずれかを」という表現が使用されている。このように、派生文書1文に対して基本文書は複数文を参照しなければならない場合がある。この問題は今後の課題である。

- 基本文書中の表を参照している文

  図 5 に示すように、基本文書に掲載されている表中の数値が派生文書の文中の数値と関係している場合がある。本手法では表中の不適切な数値を指していても矛盾と認識できない。また、基本文書の表中では“○”および“×”の記号で表現されている部分を、派生文書では文章で表現することもある。しかし、このような例は言語処理の問題ではないと考えるため、本研究の対象外とする。

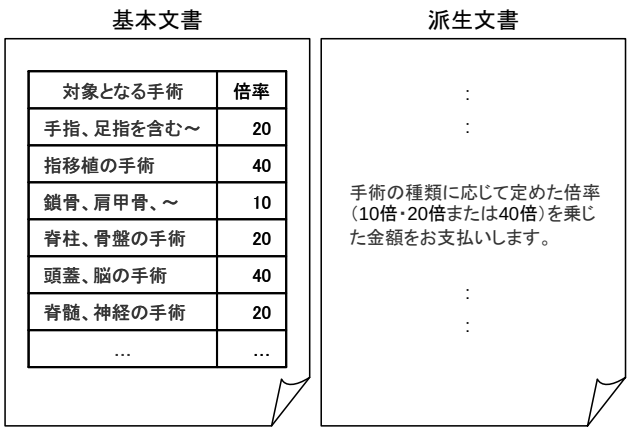

図 5： 表中の数値を参照する例

7 おわりに

保険約款等の基本文書と契約概要等の派生文書との間に生じる矛盾を認識し、校正を支援することを目的に簡単なパタンマッチングを用いた実験を行った。その結果、再現率 0.854、適合率 0.686、F 値 0.761 で矛盾を認識することができた。しかし、具体例を含んだ文を誤って矛盾と認識してしまった点、肯定・否定の両方の文に出現する単語について矛盾を認識できなかった点等に改善の余地がある。今後は、文の関連度や注目する表現の周辺情報を用いて精度を向上させること、より多くの派生文書を使用して実験を行うこと等を予定している。

謝辞

本研究を進めるにあたり、保険関連文書の参考情報を提供していただいた株式会社ミックの細川謙三社長に感謝いたします。

使用したツール

(1) 形態素解析器「茶筌」, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/

(2) 構文解析器「南瓜」, Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/

(3) IPA 品詞体系日本語辞書「IPADIC」, Ver.2.7.0, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/stable/ipadic/

参考文献

[1] 岩本 秀明, 野村 浩郷. 法律文の自然言語処理について. 情報処理学会 研究会報告 NL83-2, pp.7-14, 1991.

[2] 山田 将之, 小川 泰弘, 外山 勝彦. 構文情報付き法律文コーパスの設計と構築. 言語処理学会 第 14 回年次大会, pp.604-607, 2008.

[3] 松吉 俊, 村上 浩司, 増田 祥子, 松本 裕治, 乾 健太郎. 事態間関係知識の整備と類似・対立認識への応用. 情報処理学会 研究会報告 NL-187, pp.15-22, 2008.

[4] 乾 健太郎. 言語情報間の含意・矛盾関係の認識. 月刊言語 2008 年 8 月号, pp.30-37, 2008.

[5] Fangtao Li, Zhicheng Zheng,Yang Tang, Fan Bu, Rong Ge, Xiaoyan Zhu, Xian Zhang, Minlie Huang. THU QUANTA at TAC2008 QA and RTE track. TAC2008, 2008.