# 保険関連文書を対象とした校正支援システム

大平真一, 山本和英
長岡技術科学大学 電気系
E-mail:{ohira,yamamoto}@jnlp.org

## 1　はじめに

近年、保険や金融など紙媒体のテキストデータが中心であった分野において電子データの利用が盛んになっている。しかし、組版ソフトなどを用いて PC 上で作成した書類に関しても、校正は従来通り人手で行われているのが現状である。

保険関連の文書には、約款や特約等の書類（基礎書類）と、基礎書類の内容を消費者向けに編集した書類（派生書類）の 2 種類が存在する。派生書類は基礎書類の内容をわかりやすくまとめるために保険協会が定めたガイドラインに沿って作成されるが、その際に誤字や脱字をはじめとする入力ミスが発生することがある。また、書類作成の過程において参照する書類との内容の矛盾が発生することも考えられるため、派生書類を校正する際には基礎書類内の対応する部分を参照することが必要となる。しかし、1 つの保険に対しての基礎書類・派生書類を合計するとのべ数千ページに及ぶこともあり、すべての文に対して人手で対応付けを行いながら校正するには多大なコストが掛かる。保険関連文書は契約の内容を示すという性質上、誤りが存在したまま流通した場合、大きな損失を生むことも考えられるため、精度を保ったままコストを下げることが求められている。

そこで、我々は校正作業にかかるコストの削減と精度の向上を目的として、派生書類と基礎書類の文単位での対応付けと、対応付けの結果を用いた誤り訂正を行うシステムを提案する。本稿では、保険文書における誤りの分析とそれに基づいた校正支援システムの詳細について説明する。

## 2　関連研究

保険関連文書の校正支援を目的とした研究として、丹治ら [1] の研究が挙げられる。丹治らは保険関連文書以外の言語資源に対して用いられている対応付け手法を派生書類と基礎書類の文の対応付けに適用した。丹治らの行った実験において比較された手法は以下の 3 つである。

1. 内容語の頻度情報による対応付け
   基礎書類と派生書類から内容語を抽出、IDF を用いて単語ベクトルとして内積で類似度を計算し対応付けを行う。
2. 派生書類の手がかり語による対応付け
   派生書類から 1 単語で文の特定ができるような手がかり語を獲得し、基礎書類との対応付けを行う。
3. 基礎書類の手がかり語による対応付け
   基礎書類から 1 単語で文の特定ができるような手がかり語を獲得し、派生書類との対応付けを行う。

比較の結果、頻度情報による対応付けで最もよい結果が得られ、正解率は 7 割程度であった。しかし、文によっては手がかり語を用いた手法でのみ正解が得られた例もあり、使い分けが必要であると示唆された。

本システムでは丹治らが比較した手法のそれぞれの利点を組み合わせることを狙い、内容語の一致率と頻度による重み付けを用いて対応付けを行った。

## 3　保険関連文書

### 3.1　基礎書類と派生書類

基礎書類と派生書類の関係を図 1 に示す。

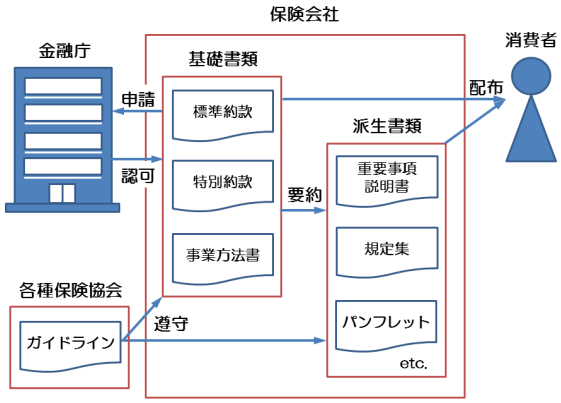


図 1: 基礎書類と派生書類の関係

基礎書類には契約の内容が詳細・正確に記載されることから法律文と近い性質を持っており、章・条・項で区分されているなどの特徴がみられる。また、基礎書類において図や表などは用語集や支払額に関するデータの列挙など補助的に使用されている。

派生書類は契約の内容をパンフレットや数ページ程度の資料にまとめているものである。そのため、章などの区分が基礎書類と全く異なる構造になっていることや 1 つの文や表に基礎書類の複数の内容が対応しているこ

とも多い。また、基礎書類に比べて表現が簡便なものとなっていることや、図や表が多用されていることなどが特徴として挙げられる。基礎書類は省庁から認可された文中に誤りのないものを、派生書類は作成中に発生した誤りを含んでいるものを想定してシステムを設計した。

本稿では基礎書類として自動車保険の普通保険約款および特別約款、派生書類として同保険の重要事項説明書を用いた。重要事項説明書には契約概要や注意点、オプションに関しての説明などが記載されている。

### 3.2 保険協会によるガイドライン

保険関連文書、特に基礎書類は一般的でない語や表現が使用されることで難解なものとなっている。そのため、保険関連文書のわかりやすさを向上させるために各種保険協会がガイドラインを策定している。例えば自動車保険に関するガイドライン [2] では以下の 3 点が保険関連文書における読みにくさの要因として挙げられている。

- 複雑な文章構造
  2 重否定や入り組んだ括弧書きなど
- 難解な用語
  保険特有の専門用語や難読漢字など
- 有効でない図や表などの利用について
  図や表を使わない文のみの説明など

これらの問題の中で『難解な用語』が派生書類を作成する際に直接的に誤りを発生させる要因となる。ガイドラインでは難読漢字のひらがな表記や 1 つの文章中における同音異字の複数使用を控えること、送り仮名の使用に関しての方針が定められている。

本システムでは語の使用法がガイドラインに沿っていない場合についても誤りとして検出できることを目標としている。

### 3.3 保険関連文書における誤りの分析

一箇所の誤りが大きな損失につながる可能性をもつという性質上、保険関連文書は多くの人によって校正が行われる。そのため、誤りを含んだまま流通する書類は非常に少ない。本研究では保険会社での校正によって発見されなかった誤りについて分析することで、人手では発見されにくい誤りを検出することのできる効率的な校正支援システムを目指した。以下に保険会社の校正によって発見されなかった誤りの分類を示す。

- 表記ゆれ
  - 送り仮名
  - かな表記と漢字表記の混在
- 誤植
  - 同音異字が原因の変換誤り
  - 専門用語など一般的でない語の誤字
- その他（数字や記号の誤りなど）

表記ゆれに関しては前項のガイドラインで表現が指定されているものとそうでないものが存在する。表現が指定されている場合はガイドラインに準拠しているか否か、指定されていない場合は表現が統一されているか否かのチェックが必要となる。

同音異字を原因とする変換誤りの例として、「障害」・「傷害」・「生涯」についての変換ミスのように文脈次第で異なる内容の文として解釈できる場合が存在する。そのため、対応する基礎書類内の文と比較を行いながらの校正が不可欠となる。

その他の誤りについては誤りに様々な傾向があり、問題が複雑になるため本稿での校正対象からは除外した。

同音異字を原因とする変換誤りは一部に同じ文字が使われている場合や近い意味を持つ場合があるため、人手での校正の際に見落とされやすい誤りといえる。

人手での校正によって発見されにくい誤りである表記ゆれ・同音異字の問題はどちらも変換ミスであるという共通点があり、読みを利用した誤り検出によって校正の効率化を達成できると考える。

### 3.4 専門用語の抽出

本稿で取り扱う自動車保険関連文書には保険や医療に関する専門用語が多数出現する。専門用語の中には一般的でない語も含まれており、形態素解析誤りの原因となることがある。そこで、基礎書類から専門用語の抽出を行い、結果を誤り訂正に用いた。抽出の流れを以下に示す。

1. 複合名詞の抽出
   IPA 品詞体系辞書 (1) において名詞に分類されるものの内、品詞が「名詞-サ変接続」と「名詞-数」に区分されるものを除いた単語を結合した。
2. 人手による確認
   複合名詞を人手で確認し、「保険関連の専門用語」と「医療などの周辺分野の専門用語」を抽出した。

4769 文からなる基礎書類から抽出できた複合語の異なり数は 766 語、その中で保険関連の専門用語が 145 語、周辺分野の専門用語は 116 語であった。以下に専門用語の例を示す。ここで、括弧内の数字は基礎書類における出現頻度を表している。

保険関連の専門用語

| |
|---|
| 後遺障害保険 (44), 子ども育英費用保険 (8),<br>傷害補償特約 (26), 自動車損害賠償保障 (13), ··· |

周辺分野の専門用語

| |
|---|
| 歯科補てつ (11), 後遺障害等級 (5), 瘻孔閉鎖 (1), 診療報酬明細 (1), ··· |

保険関連の専門用語は商品名や保険法に関する用語がみられた。商品名には低頻度の語が多く、68%が頻度10以下であった。また、周辺分野の専門用語には病名や手術名、医療制度に関する用語がみられた。病名や手術名は全体で1度しか使われないものが多く、周辺分野の専門用語の75%が頻度1、97%が頻度10以下であった。保険の商品名や病名・手術名などの用語は使用される範囲が文書内の特定の章や項だけに限定されているものが多いため、低頻度となりやすい。

# 4 提案手法

## 4.1 提案手法の概要

提案する校正支援システムの概略を図2に示す。

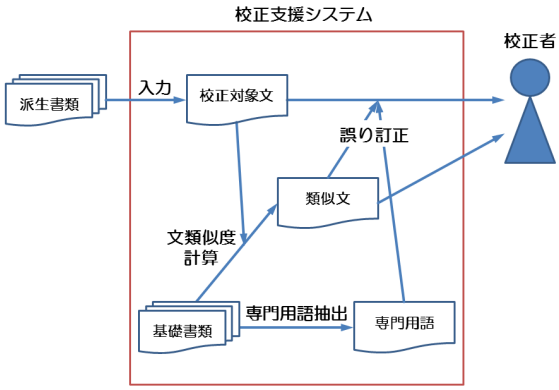


図 2: 校正支援システムの概略

派生書類は基礎書類の内容を簡潔にまとめたものであり、基本的に基礎書類に存在しない内容が書かれることは無い。そこで、本システムでは派生書類の誤り検出のための言語資源として基礎書類を用いる。

以下に提案手法の流れを示す。

1. 校正対象の文と基礎書類内の文の対応付け
   (a) 内容語の抽出
   (b) 文類似度の計算に用いる関数
   (c) 頻度を用いた重み付け
2. 誤り訂正

## 4.2 校正対象の文と基礎書類内の文の対応付け

派生書類は基礎書類の要点を数ページにまとめてあるため、章立てなど全体の構成が全く異なっている。そのため、派生書類内の1文で基礎書類の複数の章の内容が記述されている場合も存在する。

対応付けには文の類似度を示すスコアとして一般的に用いられている内容語の一致率を用いた。

### 4.2.1 内容語の抽出

基礎書類・派生書類の全文に対して形態素解析を行い、IPA品詞体系辞書において「名詞」「動詞」「形容詞」分類される単語の原形を内容語として文類似度の計算に用いた。以下に文と内容語の対の例を示す。

基礎書類内の文と内容語の対

| |
|---|
| 地震もしくは噴火またはこれらによる津波<br>地震, 噴火, これら, 津波 |

形態素解析には MeCab[(2)] を用いた。

### 4.2.2 文類似度の計算に用いる関数

内容語の集合の重なりを見る代表的な係数である *Jaccard* と *Simpson* を以下に示す。ここで、文 $S_1$ の素性の集合を $F_1$、文 $S_2$ の素性の集合を $F_2$ とする。

$$Jaccard(S_1, S_2) = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

$$Simpson(S_1, S_2) = \frac{|F_1 \cap F_2|}{\min(|F_1|, |F_2|)}$$

基礎書類・派生書類には様々な長さの文が混在しているため、文ごとの内容語の数には大きな偏りがある。*Jaccard* を使用した場合、分母が計算を行う2文の内容語の異なり数であることから、内容語の数が大きく異なると類似度が不当に低いものとなることが考えられる。また、*Simpson* の分母は計算を行う2文のうち少ない方の内容語の数であるため、内容語の数が大きく異なると類似度が不当に高いものとなることが考えられる。

そこで、本システムでは分布類似度に関する柴田ら [3] の研究において素性数の大きく異なる類似度計算のために提案された、*Jaccard* と *Simpson* を相加平均した $sim(S_1, S_2)$ を用いた。$sim(S_1, S_2)$ の式を以下に示す。

$$sim(S_1, S_2) = \frac{Jaccard(S_1, S_2) + Simpson(S_1, S_2)}{2}$$

### 4.2.3 頻度を用いた重み付け

保険関連文書では商品の説明の際に事故や事象・病名などの実例を挙げることが多い。専門用語抽出の結果より、商品や病名といった語は頻度が低いものが多く、それぞれの書類で同一の低頻度語が出現する2文は対応関係にあるといえる。そこで、類似度計算の関数の *Jaccard* と *Simpson* の分子を一致した内容語の基礎書類における頻度の逆数を総和したものとした。

頻度情報を考慮した $sim(S_1, S_2)$ の式を以下に示す。ここで、文 $S_1$ の素性の集合を $F_1$、文 $S_2$ の素性の集合を $F_2$ とし、$w$ は $F_1$ と $F_2$ 内の一致した単語、$freq(w)$ を $w$ の基礎書類における頻度とする。

$$sim(S_1, S_2) = \frac{1}{2} \cdot \left( \frac{\sum \frac{1}{freq(w)}}{|F_1 \cup F_2|} + \frac{\sum \frac{1}{freq(w)}}{\min(|F_1|, |F_2|)} \right)$$

### 4.3 誤り検出

保険関連文書の分析によって以下の知見が得られている。

- 変換誤りは人手による校正において発見されにくい
- ガイドラインによって一文中での同音異字の使用は最小限に抑えられている

以上の条件より本システムでは次の手順で誤り検出を行う。

1. 対応関係にある文の抽出
   類似度計算の結果から上位 10 文を入力文と対応関係にある文の集合 $R$ とする。
2. $R$ 内での単語出現頻度のカウント
3. 専門用語のチェック
4. 読みを用いた誤り検出
   入力文内で使用された単語と同音異字の単語が $R$ 内で使用されていた場合、入力文が変換ミスである可能性が高いため、誤りの候補として検出する。
   $R$ 内に同音異字の単語が複数存在する場合は $R$ 内での使用頻度が多いものを選択する。

同音異字の語が 1 文中に存在する例を以下に示す。

> 後遺障害：事故による肉体的な傷害が、 ···

例では「障害」と「傷害」が 1 文中に存在しているが、一方は「後遺障害」という専門用語の一部である。ガイドラインに反して同音異字の単語が 1 文中で使われている場合は一方が専門用語など分割できない場合のみであるため、誤検出を行わないために抽出した専門用語を利用する。誤り検出の際に専門用語をチェックし、一致するものは誤り検出から除外、読みが同じで漢字が異なるものを検出することで上記の例をシステムにかけた場合にも検出ミスが発生しないような設計となっている。

## 5 今後の課題

本稿で提案したシステムは派生書類を作成し、人手で確認しやすい明らかな誤りを除いた後の校正を想定して設計した。今後はシステムを適用する場面を広げ、書類を作成しながら誤りを検出することを想定した改善が求められる。

作成直後の確認作業が行われていない書類には今回想定していない以下のような誤りが考えられる。

- 入力誤りによる本来の文と全く異なる文
- 行や章などの抜け
- その他（無駄な改行や Shift キーが原因の数字・記号の誤りなど）

日本語入力において一般的に利用されているローマ字入力方式では入力誤りの際に無駄なアルファベットが挿入される場合や、母音 (または子音) のみが誤っている場合がある。今回の誤り訂正手法を拡張し、読みからローマ字を推定することでより広範囲の入力誤りに対応できると考える。また、日付などの数値の誤りに関しては丹治ら [4] の手法を用いることで誤りの検出が可能である。行の抜けや章などの広範囲に渡る抜けに関しては、現在のシステムのように文単位での処理では対応不可能なため、異なる手法の導入が必要となる。

## 6 まとめ

本稿では保険関連文書の校正支援を目的として、保険関連文書における誤りの分析を行い、その結果に基づいて基礎書類と派生書類の対応付けと誤り検出を行うシステムを提案した。今後はより効果的な校正支援を目指し、書類作成の際にシステムを利用することを想定した改良を行うことを予定している。


### 謝辞

研究を進めるにあたり、保険約款および特約、重要事項説明書の文書を提供していただいた株式会社ミックの細川謙三代表取締役社長に感謝いたします。


### 使用した言語資源及びツール

(1) IPA 品詞体系辞書 IPADIC,Ver.2.7.0,
奈良先端科学技術大学院大学 松本研究室,
http://sourceforge.jp/projects/ipadic/

(2) 形態素解析器 MeCab, Ver.0.98,
http://mecab.sourceforge.net/

### 参考文献


[1] 丹治広樹, 山本和英. 保険約款と派生書類の自動対応付け. 言語処理学会年次大会, F3-4, pp.868-871, 2011.

[2] (社) 日本損害保険協会
保険約款のわかりやすさ向上ガイドライン.
http://www.sonpo.or.jp/about/guideline/pdf/index/yakkan_guideline.pdf

[3] 柴田知秀, 黒橋禎夫. 超大規模ウェブコーパスを用いた分布類似度計算. 言語処理学会年次大会, D4-7, pp.705-708, 2009.

[4] 丹治広樹, 山本和英. 保険約款に対する派生文書の矛盾認識. 言語処理学会年次大会, D3-9, pp.820-823, 2010.