保険関連文書を対象とした文章校正支援のための変換誤り検出

林 秀治　　山本 和英
長岡技術科学大学 電気系
{hayashi, yamamoto}@jnlp.org

1　はじめに

近年、保険や金融などの書類が紙媒体が主流であった分野においても電子データの利用が増えてきている。しかし、今でもこのような電子データの書類であっても、校正は人手で行われるのが主流である。

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パンフレットなどの書類(派生書類)の2種類がある。派生書類は保険協会が定めたガイドラインに沿って基礎書類から作成されるが、その際に誤字や脱字などの入力ミスが発生することがある。また、派生書類作成の過程で基礎書類の内容と矛盾が生じる場合もある。そのため、派生書類を校正する際、基礎書類から対応する部分を参照する必要がある。しかし、基礎書類・派生書類を合計するとのべ数千ページにも及ぶ場合があり、全てを人手で対応をつけながら校正を行うのには多大なコストがかかる。また、保険関連文書は誤りが存在したまま流通した場合大きな損失を生んでしまうため 、誤りの検出を行う場合、検出漏れをいかになくすかが重要になる。

そこで、我々は校正作業の支援のため、基礎書類と派生書類の文単位での対応付けと、その結果を用いて誤りの検出漏れをなくすことに重点をおいた誤り検出を行うシステムを構築する。

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、丹治らの研究[1]や大平らの研究[2]がある。

丹治らは内容語の頻度情報における対応付け、派生書類の手掛かり語による対応付け、基礎書類の手掛かり語による対応付けの3つの手法で派生書類と基礎書類の文の自動対応付けを行った。その結果頻度情報による対応付けで最も良い結果が得られ、その結果は7割ほどであった。しかし、文によっては手掛かり語を用いた手法でのみ正解が得られ、使い分けが必要であるとされた。

大平らは丹治らの研究を基に、IPA品詞体系辞書において「名詞」「動詞」「形容詞」に分類される単語を内容語として、内容語を使った文類似度と頻度を用いた重み付けによる対応付けと、対応付けされた文を使い、読みを用いた誤り検出を行う手法を提案している。

本研究では大平らの研究に基づいて、誤り検出の漏れを減らすことに重点をおき、内容語を使った対応付けと誤り検出を行った。

3　基礎書類と派生書類の性質

保険関連文書には、約款や特約などの基礎書類と基礎書類の内容を消費者向けに編集したパンフレットなどの派生書類の2種類が存在する。基礎書類は、法律文に近い性質をもつ約款や特約などの文章で章・条・項で区分されている、文末には丁寧語を用いる、箇条書きの場合は体言止めであるといった特徴を持つ。派生書類は保険協会が定めたガイドラインに沿って基礎書類をもとに消費者向けに編集したもので、パンフレットや契約概要などがそれにあたる。読みやすさを考慮しているため、派生書類は基礎書類に比べて表現が簡便なものになっている。この派生書類を作成する際に、誤字や脱字などの入力ミスが発生することがある。これらの性質から同音異字を原因とする変換誤りが発生しやすく、また人手での校正でも見落としやすい。

これらの性質から、本研究では同音異字の変換誤りを主な検出対象とした。

4　提案手法

入力された派生書類の文の誤りを検出するためには、その派生書類を作成するのに使用された基礎書類から対応する文章を探し出す必要がある。そこで、入力文と基礎書類それぞれが持つ内容語を用いた文の対応付

け及び誤り検出の手法を提案する。

誤り検出システムの概略を図 1 に示す。

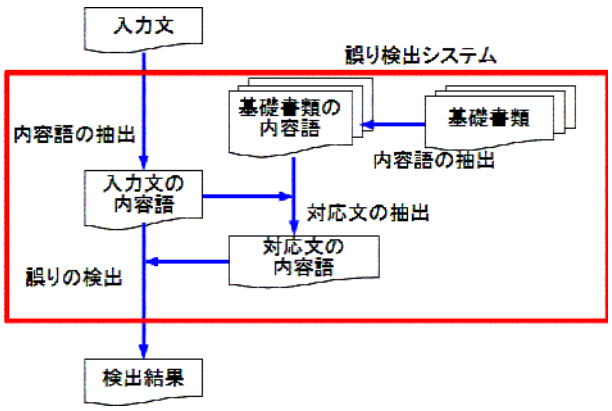

図 1: 誤り検出システムの概略

4.1 入力文と基礎書類の対応付け

入力された文に対応した文を基礎書類から抽出する。

基礎書類を MeCab(1) を使って形態素解析し、1 文毎に出現した内容語を保存する。このとき同時に、内容語とその内容語が出現した行数の対も保存する。このとき、複合名詞に対応するため、名詞が連続して出現する場合はそれらを連結し、一つの内容語して扱う。入力文も同様に形態素解析を行い内容語を抽出する。抽出した内容語が出現した基礎書類の行数を、内容語と行数の対から取得し、入力文が含む内容語を最も多く含む文を入力文に対応する文 (対応文) とする。

4.2 誤りの検出

対応付けで得られた対応文を使って入力文の誤り検出を行う。

入力文が含む内容語のうち、対応文に含まれない内容語を誤りとして検出する。その例を以下に示す。

入力文：保健証券等に記載の自動車をいいます。
入力文が含む内容語：
　保健証券等、記載、自動車、いい
対応文：保険証券等に記載の自動車をいいます。
対応文が含む内容語：
　保険証券等、記載、自動車、いい
入力文の内容語『保健証券等』が対応文が含む内容語にないため誤りとして検出される

例 1：誤り検出の例

5 評価実験

対応付けの精度及び誤り検出の精度、再現率を確認するために、テストセットを使った評価実験を行った。

5.1 テストセットの作成

誤り検出の精度と再現率を確認するため、以下の手順でテストセットを作成した。

1. 基礎書類中の文を形態素解析する
2. 1 文中に現れた名詞 1 つを、IPA 品詞体系辞書 (2) から同じ読みの名詞を取得し、その名詞に置換する
3. 置換前の文（原文）、置換した語 (置換語) 以外はそのままの文、置換語のセットを保存する
4. 以上の処理を基礎書類の全文の全名詞に対して行う

このような手順で作成されるため、置換語が誤り語となる。また、作成されるテストセットは基礎書類のままの誤りがない文も含むため、1 文中の誤りの数は 0 か 1 となる。

原文とそこから作成された文の例を以下に示す。

約款に機才の番号の読み方
約款に既済の番号の読み方
約款に奇才の番号の読み方
約款に既裁の番号の読み方
約款に鬼才の番号の読み方
約款に記載の番号の読方

例 2：原文『約款に記載の番号の読み方』から作成された文

今回 1825 文の『自動車保険の約款』から 65718 文を作成した。

6 実験方法

テストセットを入力文、テストセットを作成するのに使った『自動車保険の約款』を基礎書類として対応付け及び誤り検出を行う。対応付けは、抽出した対応文が原文と一致した場合を正解とする。誤り検出は、置換によって解析結果が変わり区切り位置も変わる場合があるので、置換語と検出した語が完全に一致した場合以外にも、検出した誤り語が置換語の一部と一致するか、置換語が検出した誤り語の一部と一致した場合を正解とする。誤り検出正解時の例を以下に示す。

完全一致：
　置換語が『保健』のとき『保健』を検出
検出した語が置換語の一部に一致：
　置換語が『支払い』のとき『支払』を検出
置換語が検出した語の一部と一致：
　置換語が『不通』のとき『不通保険約款』を検出

例 3：誤り検出正解の例

7　実験結果と考察

7.1　対応付け

対応文として原文と同じものを抽出し、対応付けに成功したのは 65718 文中 51056 文であった (精度 77.7%)。また、この 51056 文のうち一致する内容語が一つもなく対応文を抽出できなかったものが 1263 文あった。その例を以下に示す。

保険商権等
医道承認書等
配偶車

例 4：対応文抽出に失敗した入力文

対応文を抽出できなかった文はすべて、例 4 のような章の見出しなどの内容語が 1 つしかない文であった。このような文の場合、置換された後の内容語が基礎書類に含まれていなければ一致する内容語が存在しないので、対応付けは不可能である。しかし、対応文が抽出できなかった場合、入力文の内容語がすべて誤りとして検出されるため問題はない。

また、対応文は抽出できたが正しいものを抽出できなかった 14563 文の主な失敗原因は以下のようなものがあった。

1. 内容語が 1 つしかない文で置換を行ったとき、置換した後の語を含む文が基礎書類に存在した
2. 正解となる文を含み、それ以外の内容語を含むより長い文が基礎書類に存在した
3. 含まれる内容語は同じで、それ以外の部分が異なる文が基礎書類に存在した

それぞれの例を以下に示す。

原文　：備考
入力文：鼻腔
対応文：１４．鼻・副鼻腔の手術

例 5：内容語が 1 つしかない文で対応文を抽出した例

原文　：用途・車種
対応文：※１別表に掲げる用途・車種をいいます。

例 6：原文を含み、他の内容語を含む長い文がある文

原文　：記名被保険者の配偶者
対応文：ア．記名被保険者の配偶者

例 7：内容語以外が異なる文

1. のパターンでは置換語が対応文中に存在し、誤りとして検出されないため問題である。しかし、このような文は主に先ほど述べたような見出しである場合が多いため完全一致する文が存在しない場合誤りとするなどの対処法が考えられる。

2. のパターンではより長い文が対応文とされてしまうため、原文よりも含まれている内容語が増え置換語が誤りとして検出されない可能性がある。今回はそれぞれが含む内容語の数を考慮していないため、同じ内容語を多く含みそのなかで一番内容語の数が少ない物を選択するなどしてある程度数が減らせると考えられる。

3. のパターンでは内容語は同じであるため、誤りを検出する上では問題ない。

7.2　誤り検出

テストセットの誤りを検出した結果を表 1 に示す。誤りを含む 63893 文中、誤りを検出できたのは 63615 文であった (再現率 99.6%)。誤りを含まない 1825 文はすべて誤りが検出されなかった。また、誤りとして 2 語以上検出された文が 432 文あったが、これらはすべて対応文では 1 つの内容語として扱われていた複合名詞が置換によって 2 語以上になってしまい、それらすべてを誤りとして検出していたため、検出は正しくできていた。その例を以下に示す。

表 1: 誤り検出結果

全体	誤りを含む文	誤り検出有り	63615
		誤り検出無し	278
	誤りを含まない文	誤り検出有り	0
		誤り検出無し	1825
計			65718

対応文での内容語：核燃料物質
置換後の内容語　：かく燃料物質
検出された誤り　：かく
　　燃料物質

例 8：誤りを含む複合名詞が分割されて検出された例

このため、検出の精度は 100%であった。検出に成功した文のうち、対応文の抽出に失敗したものが 498 文あったが、すべて 7.1 節のパターン 2. か 3. であった。

誤りが存在するが、誤りを検出できなかった文は 278 文であった。その内訳を表 2 に示す。

対応文の抽出に失敗した 107 文は全て 7.1 節の 1. か 2. のパターンであった。対応文の抽出に成功したが誤りの検出に失敗した 105 文は、置換語が原文に元から含まれている文であった。その例を以下に示す。

表 2: 検出に失敗した文の内訳

全体	対応文抽出失敗		107
	対応文抽出成功	置換語が原文に有り	105
		置換語が原文に無し	66
計			278

原文　：事業を営む者が預託を受けている物
入力文：事業を営む物が預託を受けている物

例 9：置換語『物』が原文に含まれている文

今回は、内容語の出現回数を考慮していないためこのような検出漏れが発生したが、出現回数を考慮した場合、出現回数に差異が生じたときその文に含まれる同じ内容語がすべて誤りとして検出されてしまい、検出精度が大幅に落ちてしまうことが予想される。このような文では、N-gram の頻度情報を使うなどの例外的な処理を行うなどの対処法が考えれる。

その他の検出に失敗した 66 文だが、抽出に失敗した内容語の種類としては 19 種類だけであった。その 19 種類を以下に示す。

さん すんで トウ ほう もの ようじ 急
旧 元 小 相 多 打 超 当 内 否 比 非

これらの語はすべて形態素解析を行った際に、内容語以外の品詞としてされてしまっていて、内容語として抽出されていなかった。これらの語は文によって品詞が変わってしまうため、別途に処理を行うなどして対処する必要がある。

8　おわりに

本研究では、保険関連文書の校正を支援するために基礎書類と派生書類の対応付け及びその結果を用いた誤り検出手法を提案した。その結果誤りが名詞の変換誤り 1 か所のみの場合において、77.7%の精度で対応付けを行い、再現率 99.6%・精度 100%で誤りを検出することができた。

今回は基礎書類と同じ文章で、変換誤りが 1 つだけ発生した場合の対応付け及び誤りの検出を行った。今後は対応付けの精度、誤り検出の再現率のさらなる向上を目指し、誤りが 2 個以上の場合や文章の長さが変わる場合での検証も行いたい。

また、将来的には保険文章以外にも使用できるようにしたい。

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. 言語処理学会第 17 回年次大会，pp868-871，2011

[2] 大平真一, 山本和英. 保険関連文書を対象とした校正支援システム. 言語処理学会第 18 回年次大会，pp243-246，2012