自然言語処理が産業日本語の作業低減に貢献するために

—保険関連文書に対する自動校正の事例の紹介—

What should NLP researchers work for reducing cost of technical Japanese production ?
—A case study on Japanese proofreading—

長岡技術科学大学准教授 **山本　和英**

PROFILE 豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士（工学）。1996 年～2005 年（株）国際電気通信基礎技術研究所（ATR）、2002 年～現在まで長岡技術科学大学、現在准教授。自然言語処理、及びテキストマイニングの研究に従事。

yamamoto@jnlp.org

1　はじめに

企業において業務目的で作成された言語テキスト（以下、産業日本語）は蓄積され続け、すでに多くの企業において膨大な規模になっている。これに対処するため、自然言語処理の果たすべき役割は日を追うごとに増加している。

その一方で、私の見る限り、自然言語処理分野での様々な研究はこれら産業向け日本語処理の要求に十分に応えているようには見えない。その最大の理由は、自然言語処理研究者の関心と企業でテキスト処理を希望する技術者の需要に大きなずれがあるからだと考える。すなわち、研究者は常に新しい技術・手法の提案が求められている。一方、産業日本語を扱う現場においては、必要なのは必ずしも最先端技術ではなく、むしろ安定して学術的な評価が確立している技術のほうが多いのではないだろうか。この相違は分野に関わらず普遍的で、例えば医療の分野において最新の治療技術や薬剤を必ずしも多くの人が求めている訳ではない、というのと状況は同じである。

もう一つ、性能評価に対する考え方においても学術と現場では大きな隔たりが見える。例えばある課題 10 問を解くシステムがあった時に、学術においては 10 問のすべてを解くことを要求され、全解答にわたって平均的に良好であることが良いシステムと考える。一方、産業日本語を扱う現場においては完璧な解答の多さがより重要である。すなわち、自然言語処理システムはあくまでも人間の作業を支援するシステムであって、支援システムの最も重要な評価基準はどの程度人間の作業が削減できたかである。よって、前述の例で言えば、10 問のうち 1 問は完璧に解答でき、残り 9 問は全くできない（場合によっては解答放棄する）ほうが良いシステムである場合が多い。なぜなら後者は作業量が 10%削減できるのに対し、前者は結局出力結果をすべて見直す必要があって作業量削減に結び付かないからである。

以上をまとめると、産業日本語向けに自然言語処理の研究開発を行う場合は、従来のような平均的に良いシステムではなく、真に作業低減に結びつくための研究開発が必要であると考える。このような発想に基づいて行われている研究は極めて少数であり、今後学術界は発想を転換し、産業界の需要に応えて研究開発を進めていかなければいつまでも実用化できない。

本稿では、以上のような問題意識のもとで進めている我々の研究[1] を紹介する。我々は、共同研究として保険関連文書の自動校正の研究開発を行っている[1-3]。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いるが、その校正は依然人手である。そこで、この作業量低減を目指して我々は研究開発を行っている。

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。ここで冒頭の議論に照らして重要なのは、以下の２点である。

●訂正ではなく検出のみに注力する

訂正候補が提示されても人手の作業時間はそれほど低下しない。それよりも、検出の精度向上に集中すべきである。

●絶対に検出漏れしない

過検出（誤りのない文を誤りと判断する）の低減よりも検出漏れの撲滅を優先する。

2　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。

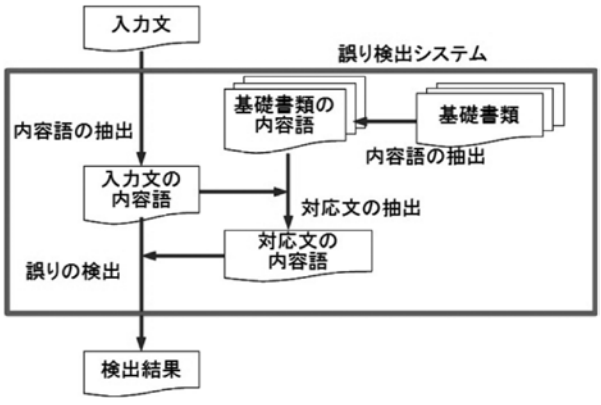

図１　誤り検出システムの概略

2.1　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。

2.2　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

3　誤り検出性能の評価

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２　誤り検出正解の例

テストセットの誤りを検出した結果を表１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	63,615
		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	1,825
計			65,718

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。

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３　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。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。

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

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。

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４　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られる。

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。

4　おわりに

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紹介した。この手法では、誤りが名詞の変換誤り１か所のみの場合において、再現率 99.6%・精度 100% で誤りを検出することができた。

冒頭で議論したように、このようなシステムを実際の現場で利用するには検出漏れなし（再現率 100%）であることが必要である。現段階では 100% は達成できていないので若干の過検出を許容してもこれを達成したい。なお、現在は名詞の変換誤り１か所のみという非常に限定された誤りしか対象にしていないが、仮にこれを検出漏れなしにできればこれだけでも十分に有用であると考えている。

本研究室では、真に産業利用できるために何が必要かという観点を決して忘れずに、今後も技術開発を推進し

ていく。

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] 林　秀治、山本　和英。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。言語処理学会第 20 回年次大会、pp. 618-621, 2014.

[2] 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、pp. 868-871, 2011.

[3] 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、pp. 243-246, 2012.