保険約款と派生書類の自動対応付け

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、電子テキストデータの増加とともに様々な分野で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。

保険に関する文書には、約款や特約等の文書(以下、基礎書類とする。)を流用して消費者向けに改変されたパンフレットや重要事項説明書等の文書(以下、派生書類とする。)が多数存在する。これらの関係を図 1 に示す。派生書類は基礎書類や協会指針のガイドラインを参照して人手により再度入力される場合もあり、誤字・脱字や入力ミス等が含まれていることがある。また、文章の校正を繰り返すうちに基礎書類との矛盾や語彙の差が生じる可能性もある。そのため、基礎書類と対応していることを確認しながらの校正が必要になるが、派生書類は延べ数万ページにも及ぶため、人手で確認するには多大な労力と時間がかかる。

派生書類を校正するには、数十章で構成されている基礎書類の中から内容が対応している部分を探し出さなければならない。しかし、大量に存在する派生書類の各文に対して人手で対応を見ることは容易ではない。

そこで、我々は派生書類の各文を基礎書類と自動的に対応付けすることで人手作業の削減を目指す。本稿では、類似文書検索をベースとした手法およびルールベースによる手法を試みた。その結果、基礎書類の章と1対1で対応する文に関しては類似文書検索の手法によって正解率約7割で対応付けすることができた。

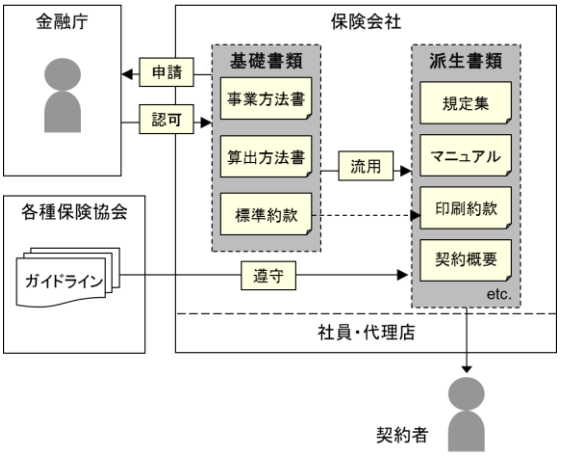

図 1： 基礎書類と派生書類の関係

2 関連研究

特殊な文体や体系をもつ文書を扱ったテキストの対応付けの研究として、丸川ら[1]および新森ら[2]、田村ら[3]の研究等が挙げられる。

丸川らおよび新森らは特許に関する文書を対象とし、特許請求項と「発明の詳細な説明」の項を対応付けしている。丸川らは複数文間での対応付けにローカルアラインメント DP マッチングを用いて順位付けすることで、計算量の削減と柔軟な対応付けを可能とした。新森らはまず特徴的な手がかり表現をパターン化し、対応付けの対象となる文を獲得した。次に特許請求項を構造解析し、文末表現等のパターンを用いて対応付けの要素を抽出してからローカルアラインメントをとることで高精度な対応付けを実現している。特許文は長い 1 文で構成されているために並列構造が多く、1 文内に多くの情報を保持している。それに対して保険関連文書は箇条書きや文をまたいだ列挙が多くみられるため、この手法を保険関連文書に適用するには改善が必要だと考える。

田村らはコールセンターの通話記録とコールメモのトピック対応付けをしている。不要発話を除去したうえでトピックごとに分割することによって、頻度情報を用いた対応付けで従来手法より正解率を 5 ポイント程度上昇させている。しかし、適合率および再現率はそれぞれ 0.5 に満たないため、実用レベルでないと考える。

類似文書検索の手法を用いたテキスト対応付けの研究として、池田ら[4]の研究が挙げられる。池田らは blog とニュース記事を対象に、頻度情報を用いた類似度で対応付けしている。ニュース記事および blog について TF･IDF や IDF を用いてベクトルを作成し、コサイン類似度または内積により類似度を計算している。池田らの研究では、blog ベクトルの重み付けには IDF のみを用い、ニュース記事との類似度を内積により計算した場合、一般的に用いられるコサイン類似度よりも精度が高いという結果を得ている。blog は多数の話題に触れることもあり、文の長さと類似度に関連が薄いためだと考えている。そのうえ、blog という更新頻度の高い文書の性質を利用して、語の出現頻度の推移で重み付けすることによって精度を改善している。本稿で対象とする基礎書類と派生書類の間には、blog とニュース記事の場合と同様に文書の性質に差がある。そのため、この手法を参考に頻度情報を用いた手法を試みた。

3 使用した言語資源

保険関連文書には、大別して基礎書類と派生書類の 2 種類が存在する。基礎書類は省庁に提出する必要があり、法律文に近い性質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、文末には丁寧語を用いている。ただし、箇条書きの場合は体言止めである。派生書類は基礎書類をもとにして消費者向けに文章を改編および抜粋している、パンフレットや契約概要のような文書である。基礎書類を簡潔にまとめているため、1 文中で基礎書類の複数章に触れることも多々ある。逆にサポートや苦情、相談室に関する情報等、基礎書類にない項目も記載されている。視覚的な読みやすさを考慮しているため、箇条書きや表形式で文章を収めたものが多用されている。

本研究では、基礎書類として自動車保険の普通保険約款および特約条項、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、契約概要および注意喚起情報、保険のオプションに関する記述や個人情報保護に関する記述等が記載されている。

4 対応付けの手法

本稿では、派生書類の各文について対応している基礎書類の章を提示することを目的とする。そこで、頻度情報を用いた手法、派生書類の語を用いた手法および基礎書類の語を用いた手法の 3 種類を試みた。

4.1 頻度情報による対応付け

頻度情報を用いて基礎書類と派生書類を対応付けする流れを以下に示す。

1. **基礎書類からの単語ベクトルの作成**

   頻度情報としては TF･IDF を用いるのが一般的である。しかし、池田ら[4]は blog ベクトルの作成時に IDF のみによる重み付けで TF･IDF を用いたときよりも良い結果を得ている。そこで、本稿でも IDF を用いた単語ベクトルも作成した。単語ベクトルには名詞、動詞、形容詞を用いた。ただし、「する」、「場合」、「こと」の 3 単語は頻出かつ手がかりにならないためストップワードとした。TF と IDF の処理単位は条を用いた。

2. **派生書類からの単語ベクトルの作成**

   1.と同様に TF･IDF または IDF を用いて単語ベクトルを作成した。処理単位は文を用いた。

3. **ベクトル間の類似度の計算**

   作成した単語ベクトル間の類似度を計算する。類似度の計算にはコサイン類似度を用いるのが一般的である。しかし、派生書類では 1 文中で複数の項目に触れるため、文の長さで正規化するコサイン類似度よりも内積が適切だと考え、内積による対応付けも試みた。計算された類似度が最大となる章を提示した。

4.2 派生書類の手がかり語による対応付け

派生書類を人手により校正するとき、文全体を見て基礎書類との対応をとらなくとも 1 単語で特定できる場合が多い。まずは派生書類の文から手がかり語を獲得する。手がかり語とは基礎書類の章を特定するのに有効な一語である。次に基礎書類の中でその語を検索し、照合した部分の周辺を見て対応しているか否か判断する。以下に例を示す。

例） お車の**入替**の場合（自動車を新たに取得し（以下省略）
⇒ 約款 第 7 章 第 8 条（被保険自動車の**入替**）に対応

この手順をもとに、派生書類の各文から手がかり語を獲得して基礎書類の手がかり語との一致を見た。派生書類において対象の文から特有の語を抽出するために、派生書類中での IDF が最大となる語を用いた。基礎書類で手がかり語を検索し、照合した数が最も多い章を提示した。手がかり語には名詞、動詞、形容詞を用いた。

4.3 基礎書類の手がかり語による対応付け

人手による校正で用いる手がかり語は、基礎書類の本文よりもタイトル等の特徴的な位置に出現することが多いと考えた。 そこで、基礎書類における以下の 4 項目を手がかり語の基準として用いた。

- 章のタイトルに出現した語
- 各章の「第 1 条（用語の定義）」で定義された用語
- 用語の定義文に出現した語
- 「用語の定義」以外の条のタイトルに出現した語

上記の基準により獲得できる手がかり語の例をそれぞれ例 1～例 4 に挙げる。

これらの基準を用いて基礎書類の章ごとに手がかり語を獲得した。派生書類の各文について章ごとに獲得した手がかり語との一致を見て、照合した数が最も多い章を提示した。手がかり語には名詞、動詞、形容詞を用いた。

搭乗者　人身傷害　盗難

例 1： 章のタイトルに出現する手がかり語の例

記名被保険者　対人事故　免責金額

例 2： 用語の定義に出現する手がかり語の例

後遺障害　衝突　火災　落下

例 3： 用語の定義文に出現する手がかり語の例

費用　入替　支払う　告知義務

例 4： 条のタイトルに出現する手がかり語の例

表 1： 頻度情報による対応付けの結果

ベクトル	類似度	正解率
IDF	内積	**0.692**
	コサイン類似度	0.569
TF･IDF	内積	0.392
	コサイン類似度	0.573

表 2： 手がかり語による対応付けの結果

抽出元	基準	正解率
派生書類	IDF	0.392
基礎書類	章のタイトル 定義された用語 用語の定義文 条のタイトル	0.425

5 評価実験

基礎書類として、自動車保険の約款および特約計 48 章 3,868 文を使用した。派生書類として、重要事項説明書のうち約款または特約の章と 1 対 1 で対応している 487 文を使用した。基礎書類と対応のない文 453 文および 1 対多で対応しているもの 80 文を人手で除外した。

単語ベクトルの作成や手がかり語の抽出時には形態素解析器「茶筌」[1] を用いた。品詞体系は IPA 品詞体系日本語辞書[2] に準ずる。

頻度情報による対応付けの結果を表 1 に示す。ここで、複数の章で最大の類似度と同じ値を得たとき、いずれかの章に正解を含む場合には正解とした。基礎書類および派生書類の手がかり語による対応付けの結果を表 2 に示す。ここで、複数の章で手がかり語のヒット数が最多かつ同じであった場合、いずれかの章に正解を含むものを正解とした。

6 考察および検討

6.1 頻度情報による対応付け

表 1 の結果より、IDF を重みとして単語ベクトルを作成し、内積により類似度を計算した場合に正解率が最も高かった。保険関連文書において 1 文中に繰り返し使用される語は「保険」や「補償」等、対応付けの参考にならない語が多数あった。これが原因で、TF が類似度の計算に悪影響を及ぼしていた。また、派生書類には複数の項をまとめた文があり、1 文のすべてが基礎書類の同じ部分に対応するわけではない。このように基礎書類との対応は文の長さに依存していないため、文の長さで正規化しているコサイン類似度よりも内積の方が適していた。

対応付けできなかった文を観察した。これらの多くは、人手による校正では 1 単語で対応がとれた文であった。「対人賠償」や「変更」等の特徴的な語をもっていても文内の別の単語により類似度が下がってしまった。短めの文や全体的に同一の章に出現する語で構成された文は対応がとれていた。

6.2 派生書類の手がかり語による対応付け

表 2 の結果より、手がかり語を用いた手法はいずれも正解率 4 割程度にとどまった。派生書類の手がかり語は IDF により決定し、基礎書類の章ごとでのヒット数を計数した。しかし、IDF では正確に重要語を選定できていないためにこのような結果になったと考える。派生書類には章や条といった明確かつ詳細な内容の区分はないが、段落のような区切りは存在する。その区切りの中では同一の章に対応する文が多く、手がかりとなり得る語が複数文にわたり出現することもある。本手法での処理単位は文であるため、複数文にわたり出現する手がかり語を獲得できなかったと推測する。このような問題を解決する案として、語の出現の連続性の考慮や事前にトピック分割をすることが挙げられる。前者は出現頻度が同数である場合に、離散的に出現した語と連続的に出現した語で重みを変える等が考えられる。後者は派生書類をトピックごとに分割して処理単位を広げることで、同一の章に対応する文の区切りごとに対応付けすることが可能となる。

6.3 基礎書類の手がかり語による対応付け

基礎書類の手がかり語は 4 つの基準により決定した。例 1～例 4 に挙げたような特有の表現を獲得できている。しかし、定義された用語の中には「記名被保険者」や「治療」等、多くの章で毎回定義されている用語もある。各章で条の数や定義される用語の数、定義文の長さが大幅に変わっており、特に定義文の長い章と照合しやすい傾向にあった。これらの問題を解決するために、獲得した語を候補として取捨選択や順位付けする必要がある。実際に手がかりとなる語はこれらの中でも章特有の語や、章の中で重要な意味をもつ語に限られる。また、獲得した手がかり語の中に照合する語がなかったために対応がとれなかった文もあった。これに対しては、手がかり語候補の抽出範囲を条文まで広げ、出現位置で重みを付ける等して差別化しながらも派生書類の出現する語の被覆率を上げる必要があると考える。

表 3： 派生書類全文での対応付け

手法	再現率	適合率	F 値
頻度情報	0.244	0.368	**0.293**
	0.021	**0.986**	0.041
派生書類の手がかり語	0.074	0.368	0.122
	0.060	0.641	0.110
基礎書類の手がかり語	0.072	0.374	0.120
	0.019	0.747	0.036

7 追加実験

4 章で述べた手法をもとに、基礎書類と対応のない文および 1 対多で対応している文も加えた派生書類 963 文を用いて実験した。頻度情報による対応付けには IDF を重みとして単語ベクトルを作成し、内積により類似度を計算した手法を採用した。類似度や手がかり語のヒット数に閾値を設け、閾値を超える章がない場合に対応なし、閾値を超える章が複数ある場合に 1 対多の対応と判断した。類似度は 0.1 刻み、手がかり語のヒット数は 1 刻みで閾値を設定した。これらの結果を表 3 に示す。それぞれ上段が最大 F 値の点、下段が閾値 0 を除く最大の再現率での結果である。頻度情報による手法では閾値 0.3、手がかり語のヒット数の閾値は 1 とした。

表 3 の結果より、頻度情報による対応付けでも最大 F 値 0.3 程度であった。実用化するには再現率 1 が望ましいが、再現率 0.986 では適合率が 0.02 程度であった。これは基礎書類との対応がない文に対しても多数の対応をとっていること、基礎書類の章と 1 対 1 で対応している文に対して複数の対応を出力していることが大きな問題となっている。

派生書類および基礎書類の手がかり語による対応付けは、手がかり語が 1 語も照合しない文があったために、閾値 1 でも再現率が 0.7 程度であった。これは手がかり語の選定方法に問題があるためであり、必要な手がかり語を確実に獲得できる手法が必要となる。重要語抽出の手法を用いて手がかり語を抽出する等が考えられる。適合率については、頻度情報と同様に不要な対応を多数とっているため 0.1 に満たなかった。使用した派生書類 963 文のうち 453 文と約半数に基礎書類との対応がないため、これが大きな問題となっている。基礎書類との対応がない文については不要文削除等の手法を用いて事前に除外することが考えられる。

8 おわりに

本稿では、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自動対応付けを行った。IDF による重み付けで単語ベクトルを作成し、内積により類似度を計算した方法を用いて、正解率約 7 割で対応付けすることができた。一方で手がかり語を用いた手法では正解率 4 割程度にとどまった。獲得した手がかり語に適切でないものが多数あったためだと考える。

今後は手がかり語を正確に獲得するために重要語抽出の手法を適用することを考えている。基礎書類との対応がない文および基礎書類の複数カ所と対応している文についても対処する必要がある。本稿の手法を流用した方法では、対応がない文にも複数の対応をとったり、1 対 1 で対応している文に不要な対応をとったりしている。そのため、不要文除去の手法により対応のない文をあらかじめ自動で除外する等のアプローチが必要だと考えている。

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) 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] 丸川 雄三, 岩山 真, 奥村 学, 新森 昭宏. ローカルアラインメントを用いたテキスト間の柔軟な対応付け. 情報処理学会 研究会報告 NL151-4, pp.23-28, 2002.

[2] 新森 昭宏, 奥村 学. 特許請求項解読支援のための「発明の詳細な説明」との自動対応付け. 自然言語処理 Volume 12 Number 3, pp.111-128, 2005.

[3] 田村 晃裕, 石川 開, 安藤 真一. 不要発話特定を導入した通話とコールメモ間のトピック対応付け -差分マイニングの性能改善に向けて-. 言語処理学会 第 16 回年次大会, pp.1062-1065, 2010.

[4] 池田 大介, 藤木 稔明, 奥村 学. Blog とニュース記事の自動対応付け. 言語処理学会 第 11 回年次大会, pp.1030-1033, 2005.