2R-07

利用規約等における重要文の抽出手法の検討

野村 佳太† 前川 司† 水谷 晃三† 荒井 正之†

帝京大学 理工学部†

A method to extract important sentences from user policy and agreement conditions

Keita NOMURA† Tsukasa MAEKAWA† Kouzou MIZUTANI† Masayuki ARAI†

1 はじめに

Web サービスなどの会員登録の際には，利用規約や約款が表示されるが，それらは難解でかつ長文であるため，利用者は読まないことが多い．近年，Web サービスの利用者が増加したことで，規約をよく理解せずに契約してしまい，後でトラブルとなるケースがある．

本研究は，規約の中でも特に重要度の高い内容を，機械的に抽出することが目的である．

関連研究として，新聞記事を対象として単語を TF･IDF 法で重み付けし，文の重みの総和を重要度とする Zechner の重要文抽出手法の評価が挙げられる[1]．また，本研究と同様に EDR 概念辞書を用いた研究として文献[2]などがあげられる．

2 規約や約款等に対する重要度の調査

2.1 調査の概要

6 名の被験者を対象として，10 件の異なるサービスの利用・契約に関わる規約文を読んでもらい，どのような文を重要文と判定するかの調査を行った．最重要文を 4 として，0~4 の 5 段階で重要度を判定してもらった．被験者 3 名の結果を図 1 に示す．

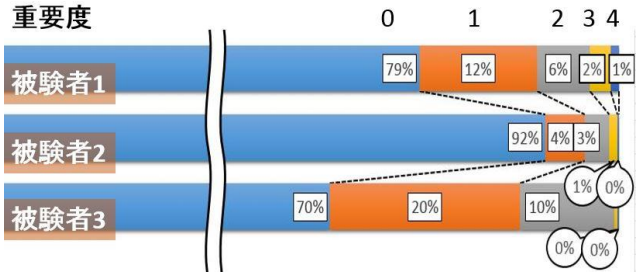

図 1：10 件の規約文に対して重要と判定した文の割合の平均

2.2 考察

図 1 より，被験者によって，その文書中で重要とする文の比率が異なることがわかる．被験者 1 は既婚で子供がおり，年齢は 50 歳代である．被験者 2,3 は 20 歳代の学生であるが，同年代でも重要文の判定に差があることがわかる．

一方で，「ユーザが責任を負う」「提供者のサービスの終了」「個人情報の第三者提供」などの文に対しては，複数の被験者が共通で重要と判定していることがわかった．

以上から，重要文の判定には，個人の属性等に左右される個々人ごとの尺度と個人の属性に依存しない共通の尺度が必要であると考える．

3 抽出手法の概要

2 章の調査結果を踏まえて，図 2 に示す抽出手法を提案する．

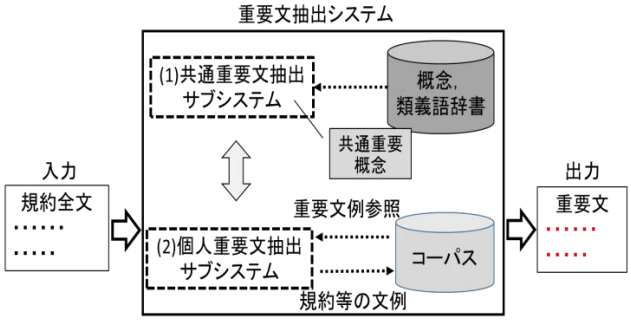

図 2：抽出手法の概要

(1)共通重要文抽出では，共通重要文の抽出を行う．共通重要文を抽出するために，概念辞書の概念の形式で，共通重要概念を予め登録しておく．(2)個人重要文抽出サブシステムでは，ユーザから返された重要度と重要文をコーパスに追加し，それらを個人重要文の抽出時に用いる．

4 共通重要文の抽出方法と評価

共通重要文を抽出するために，文に重みづけをする必要がある．概念体系を類義語の検出に用いることが可能と考えて，ここでは，EDR 概念辞書を用いて，図 2 に示した共通重要概念との整合度(以下，共通重要概念整合度とよぶ)を求める．最終的には，共通重要概念整合度から，文の重要度を判定して抽出する．

まず，図 2 に示した共通重要概念として，2.2 節で挙げたような文に含まれる単語が含まれる概念を，下位概念も含めて登録する．具体的に

A method to extract important sentences from user policy and agreement conditions
†Teikyo University Faculty of Science and Engineering

は，EDR 概念辞書の概念はそれぞれ，一意識別できる識別子(ID)を持っているので，それらの ID をシステムに登録する．

共通重要概念整合度の設定について説明する．まず対象となる文を形態素解析して，名詞だけを抽出する．次に EDR 辞書の構成要素の一部である見出しと説明文の情報の中から，これらの名詞を検索する．これらの名詞が含まれる ID の中で登録した概念中の ID と合致する割合を名詞の共通重要概念整合度とする．最後に文中の名詞の共通重要概念整合度の総和を文の名詞数で割ったものが，その文の共通重要概念整合度となる．

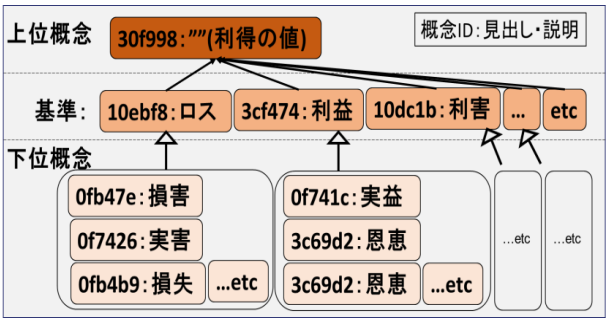

図 3：概念体系の一例

4.1 実験による評価

2 章で用いた規約文一つ一つに対して，共通重要概念整合度を求めた．被験者が設定した重要度の平均値と，プログラムで求めた共通重要概念整合度の関係を図 4 に示す．

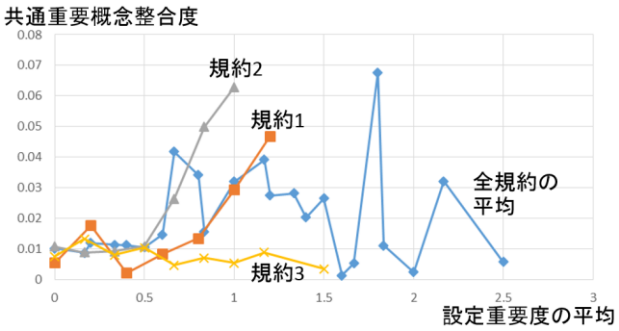

図 4：共通重要文抽出手法での共通重要概念整合度の分布

4.2 考察

図 4 の規約 1，規約 2，平均のグラフを見ると，設定された重要度の平均の値が 1.0 近傍で高い共通重要概念整合度であることがわかる．これは，多くの被験者が共通重要文を重要度 1~2 と設定しているからであると考えられる．たとえば，被験者は 2 章で述べた共通重要文の多くに対して，重要度 1~2 と評価した．

5 個人によって異なる重要文の抽出方法と評価

コーパスに登録された個人ごとの重要文と，その文の重要度を用いて，規約を構成する文から，ユーザが重要と考える文を抽出する．今回は，文と文の比較に多用されている Bag-of-words を使い，抽出の可能性を検討した．Bag-of-words に用いた品詞は動詞と名詞である．2 章の調査で用いた 10 文書のうち，9 文書をコーパスとして用い，1 文書をテスト用文書として，cos 類似度を計算した．

5.1 実験結果

コーパスの文例の重要度と，テスト用入力規約に設定された文の重要度との類似度の関係を表 1 に示す．

表 1：コーパスと入力規約の重要度毎の類似度

コーパス / 入力規約	重要度 0	1	2	3	4
重要度 0	0.05	0.07	0.06	0.06	0.04
1	0.08	0.11	0.10	0.10	0.05
2	0.09	0.12	0.11	0.12	0.07
3	0.08	0.12	0.13	0.11	0.05
4	0.10	0.13	0.13	0.13	0.07

5.2 考察

表 1 より，コーパスに用いた文例と，テスト用の文に設定した被験者の重要度の類似度が低いことがわかる．よって，本稿で提案した手法では，有効な結果が得られないことがわかった．

6 おわりに

規約文書に対して，一般に共通する重要文と，個人毎の重要文を分けて考えて抽出する方法を提案した．各手法について実験を行った結果，共通重要文に対しては概念体系などを利用した手法がある程度有効であると確認できた．一方で，個人毎の重要文設定の指標については，Bag-of-words を用いてコーパスに登録された文と比較するだけでは，有効な結果が得られないことがわかった．まずは Bag-of-words 以外の方法で文を比較することを検討したい．

参考文献

[1]Zechner, K. “Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences,” Proc. of the 16th ICCL, pp.986-989, 1996.

[2] 近藤 恵子,奥村学:言い替えを使用した要約の手法, 情報処理学会研究会報告,NL-116-20, pp.137-142, (1996).