機械学習を用いたステークホルダ分析方法の提案と評価

M2015SE001　市川　裕也

指導教員　青山　幹雄

1. 研究背景

要求獲得においてステークホルダを明らかにすることが重視されている. しかし, 現在の要求獲得方法は分析者に依存することから, 大量データを分析することが困難であり非効率的である. 一方, 大量データ分析に適した機械学習が注目されている. しかし, システム開発におけるデータを機械学習で分析し, 構造化する方法は確立されていない.

本稿はクラスタリングと自然言語処理の技術を用いて議事録データから発話の構造化を行い, ステークホルダ分析する方法を提案する.

2. 研究の課題

本稿では以下の 2 点を研究課題として設定した.

(1) 文書データを構造化し, ステークホルダを特定する方法の確立

(2) 提案の有効性と妥当性をプロトタイプで評価

3. 関連研究

3.1. ステークホルダ構造化分析の提案[2]

データ分析に基づく要求獲得プロセスに基づいて発話内容を分析し, 発話意図を可視化する研究がある. この研究で提案されている発話意図の分類方法では, 形態素解析器で文書の分析を行い, 予め定義した品詞と語彙を分析結果から得られた語彙と品詞を比較することで, 発話者の意図を特定している.

3.2. グラフデータベース[8]

SNS のユーザ間の関係などの半構造化データを表現するために用いられる.

グラフ DB で扱われるプロパティグラフモデルはノード, 関係, プロパティで構成される(図 1). ノードと関係にプロパティを設定することができ, 設定することにより, 一意に特定することができる.

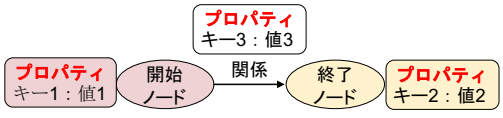

図 1 プロパティグラフモデル

3.3. 機械学習[7]

機械学習を応用した教師データに基づく予測分析は, スパムメール判定などの文書分類に用いられている. しかし, 要求獲得において機械学習を用いた研究は行われていない.

機械学習の評価を行うために, 混合行列(表 1)を用いて正解率(式(1))を算出する方法がある.

表 1 混合行列の定義

	予測+	予測-
正解+	true positive(TP)	false negative (FN)
正解-	false positive(FP)	true negative(TN)

$$\text{正解率} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

4. アプローチ

アプローチの全体像を図 2 に示す. システム開発にはステークホルダの意見が重要となってくる. そのため, 議事録などの大量の文書データからステークホルダの意図を抽出し, 話題の内容を構造化する方法を提案する. そのため, 人手による学習モデルの定義を行い, この学習モデルを用いて文書分析の構造化結果をグラフで表現する. さらに, このグラフを用いてステークホルダ分析を提案する.

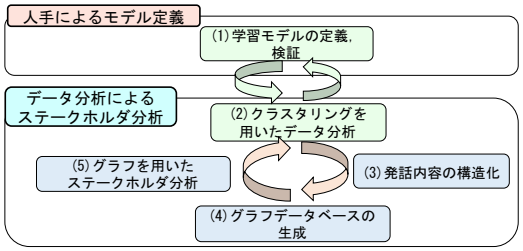

図 2 アプローチの全体像

5. ステークホルダ分析プロセス

ステークホルダ分析支援プロセスを図 3 に示す. 文書から発話者の意図を抽出する. そして, 文書から発話構造の抽出を行う. さらに抽出した文書構造をグラフで表現し, このグラフを用いてステークホルダ分析を行う.

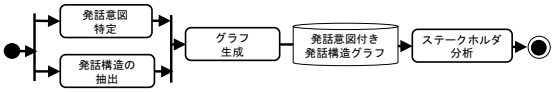

図 3 ステークホルダ分析プロセス

5.1. 発話意図特定プロセス

発話意図とは, 発話内容から抽出できる発話者の意図である. 研究[2,3]より, 発話意図は語尾の情報

から「報告」,「返答」,「受入」,「問い」,「要望」,「示唆」の 6 種類とした. この定義に基づき形態素解析器から得られた語尾の 4 単語と 4 つの品詞を教師データ(表 2)として定義する. この教師データを用いることで, 発話内容から発話意図を取得する.

表 2.発話意図分類モデルの教師データ(部分)

発話意図	語尾単語	品詞 1	品詞 2	品詞 3	品詞 4
報告	レクを実施する	名詞	助詞	名詞	動詞
返答	その通り			連体詞	名詞
受入	検討する			名詞	動詞
問い	実施するのか	名詞	動詞	名詞	副助詞
要望	宣言して欲しい	名詞	動詞	助詞	形容詞
示唆	中心だと思う	名詞	助動詞	助詞	動詞

5.2. 発話構造の抽出プロセス

発話内容を構造化するために係り受け解析を用いたトリプルの抽出方法を提案する. トリプルとは, 文書の「主語」,「述語」,「目的語」の対とする. また, 抽出したノードは名詞とし, 関係は動詞として扱う. 以下に 2 つのトリプル抽出方法を示す.

5.2.1. 共通文節の抽出方法

係り受け解析器を用いて文節に分割し, 係り受け先の文節が共通している文節を抽出する.

5.2.2. 詳細の抽出方法

5.2.1 節では文書の主語, 述語, 目的語は抽出できるが, この構造以外にも重要な情報がある. そこで, 文節の語尾の品詞が格助詞「の」が抽出したら, 主語を格助詞の前の名詞とする. さらに, 述語を「詳細」とし, 目的語を係り受け先の文節とする.

5.3. 発話意図グラフの生成

5.1 節で抽出した発話意図と 5.2 節で抽出したトリプルをグラフで表現する. 手順を以下に示す.

(1) 話題と発話者の関係の定義(図 4)

大規模システム開発では発話者の役割が変わる. 話題に対して発話者を特定するために話題ノードと発話者ノードを定義する. 話題ノードのプロパティは id と topic とする. 発話者ノードのプロパティは id, name とする. さらにノード間の関係を Participant とする.

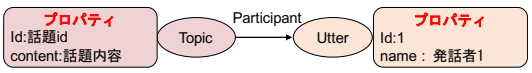

図 4 話題と発話者の関係

(2) 発話者と発話内容の関係の定義(図 5)

発話者の発話内容を特定するために発話内容ノードを定義する. 発話内容ノードのプロパティは発話内容 id, 話題 id を値とする topic_id, 5.2 節で定義した発話意図の値とする label とした. このノード間の関係を Remark とする.

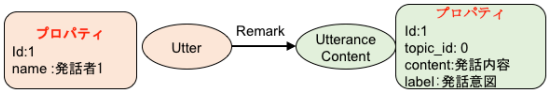

図 5 発話者と発話内容の関係

(3) 発話内容とトリプルの関係の定義(図 6)

発話内容と 5.2.1 節で抽出した最初のノード間の関係を定義する. 発話内容とノード間の関係は Intention とした.

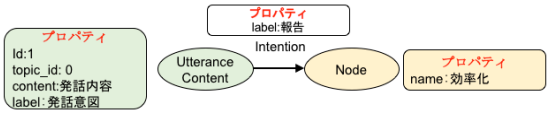

図 6 発話内容とノードの関係

(4) トリプル間の関係の定義(図 7)

5.2.1 節のノード間の関係を Relation とする. また関係のプロパティは name と設定し, その値として抽出した共通の文節を格納する. 5.2.2 節のノード間の関係を Detail とする. 関係のプロパティに話題ノードの id を付与することで話題毎の分析ができる.

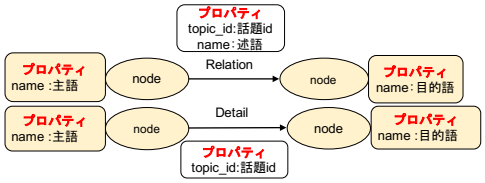

図 7 ノード間の関係

6. プロトタイプ

提案を評価するためにプロトタイプを作成した(図 8). 発話意図の特定を行うために機械学習 Jubatus[4] と形態素解析器 MeCab[5]を用いて発話意図分類モデルを作成した. 次に, 発話構造の抽出を行うために形態素解析器と係り受け解析器 CaboCha[1]を用いて発話構造を抽出する. 最後にグラフデータベースのクエリを作成し, グラフデータベース Neo4j[6]でクエリを実行し発話意図付き発話構造を可視化する.

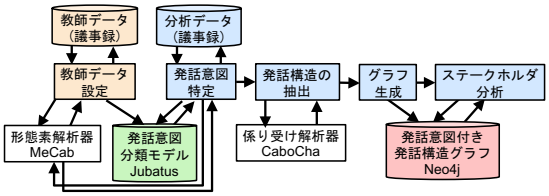

図 8 プロトタイプ

7. 実システムへの適用

公共情報システムに関わる議事録データの 27 回から 28 回目の約 12,000 字を分析した結果を図 9 に示す. また, 抽出したノードの対応関係を表 3 に示

す.

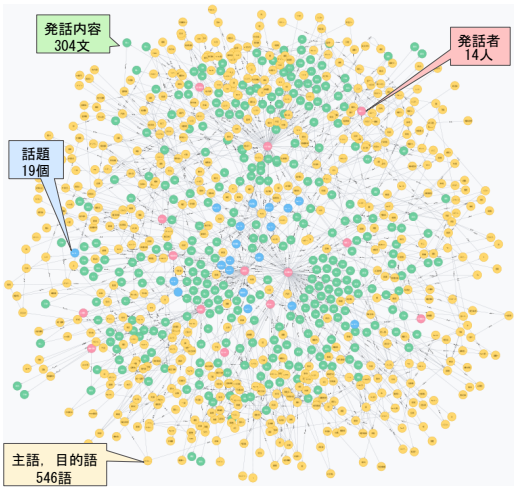

図 9 発話意図付き発話構造グラフ

表 3 議事録から抽出したノード数

回数 [回]	語彙数 [字]	話題 [個]	発話者 [人]	発話内容 [文]	トリプル [語]
27	4,286	3	8	112	221
28	7,483	16	13	192	402

8. 適用結果の評価

エラー! 参照元が見つかりません。章の結果を用いてステークホルダ分析を行う. 分析方法は以下の 3 つとした.

(1) 発話内容から抽出した話題構造を行う.
(2) 発話意図分類モデルの結果と人手が行った分類結果を用いて発話者の役割の比較を行う.
(3) 発話意図分類モデルの正解率を示す.

8.1. ステークホルダの話題構造モデルの分析

1 つの話題に対して発話内容から抽出されたノードについて評価する. 抽出されたノードのリンク接続数を調査すると,「組織内 LAN」,「コスト削減」などのノードが上位に挙がった. この結果から,「組織内 LAN」のノード間の関係を図 10 に示す.

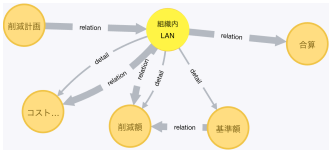

図 10「組織内 LAN」ノードの周辺の関係

「組織内 LAN」から「削減額」,「削減計画」,「基準額」などが関係付けられていることから, 組織内 LAN のコスト削減が重要な課題であることが明らかとなった. また, 議事録からもコスト削減について議論していたことを確認できたため, 発話内容から発話構造を抽出し, 各話題に対するステークホルダの意図を抽出できた.

8.2. 発話者分析

8.2.1. 発話者の役割分析

話題内での発話者の役割を求めるために, 発話意図の重み, 影響度, 関与度を定義した[2].

(1) 発話意図の重み付け

発話者の役割を調べるために, 発話意図の重み付けは研究[2]に基づいて定義した (表 4).

表 4 発話意図の重み

	報告	返答	受入	問い	要望	示唆
重み(%)	5	5	15	15	20	40

(2) 影響度と関与度の算出

発話者から影響度と関与度を定義する. 関与度は発話者の発話数の確率を求めるため式(2)で算出する. 関与度が高ければプロジェクトに関与している.

影響度は, 発話意図の重みと発話数を用いて式(3)で算出する. 影響度が高ければ, 意思決定に大きな役割を果たしていることを示している.

$$\text{関与度} = \frac{\text{対象にしている話題の特定発話者の発話数}}{\text{対象にしている話題の発話数の合計}} \quad (2)$$

$$\text{影響度} = \frac{\sum(\text{重み} \times \text{特定発話者における各発話意図の発話数})}{\text{特定発話者の発話数}} \quad (3)$$

8.2.2. ステークホルダマトリクスの作成

議事録全体の 3.5%にあたる 5,000 字(110 文), 7%にあたる 10,000 字(330 文), 14%にあたる 20,000 字(660 文)を教師データとしてラベル付けを行った. 組織内 LAN に関する 28 文の発話内容の分類結果から発話者 4 人の影響度と関与度を算出した. この結果と人手による発話意図分類から算出した結果を図 11 に示す.

C 氏, D 氏の発話意図分類結果は発話意図分類モデルと人手による分類が一致したため, 影響度と関与度の算出結果が一致した. しかし, A 氏の分類モデルによる抽出では「報告」と判定されなければならないのに対し, 発話意図の重みが高い「示唆」と「要望」が得られたため影響度に差が生じた.

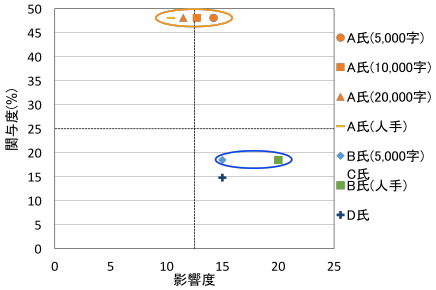

図 11 ステークホルダマトリクス

8.2.3. 教師データの効果

8.2.2 節の A 氏の影響度と誤差率を図 12 に示す. 教師データを増やすことで誤差率(式(4))が減少する

ことから，教師データを増やすことで線形となることを発見した．

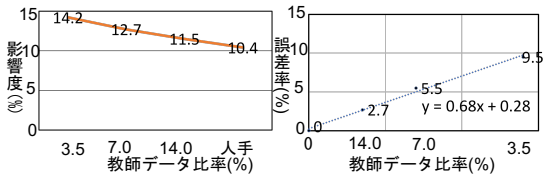

図 12 A 氏の影響度の推移と誤差率

$$\text{誤差率}=\frac{\text{人手による影響度} - \text{学習モデルによる影響度}}{\text{話題に対する発話者数}} \quad (4)$$

8.3. 発話意図分類モデルの評価

28 の発話内容に対して発話意図モデルが分類を行った集計と，5.1 節の定義を用いて人手で行った集計の結果を図 13 に示す．最も誤差が生じたのは報告と要望であることが明らかである．

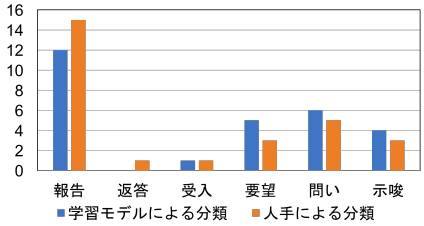

図 13 発話意図分類の比較

次に，式(1)を用いて正解率を算出した結果を表 5 に示す．正解率の算出結果から低いものはなかった．しかし，算出過程から語尾が名詞で終わる発話内容に誤差が生じていたことが明らかとなった．

表 5 発話意図の正解率

	報告	返答	受入	問い	要望	示唆
正解率(%)	89.2	94.6	100.0	92.9	96.4	85.7

9. 考察

9.1. 話題構造モデルの有用性

発話内容から抽出した話題構造モデルを分析し，ステークホルダの意図を抽出した．これにより話題構造モデルの有効性が示せたと考えられる．しかし，「もの」や「こと」など口語特有の表現で意味を一意に特定できない場合があった．そのため，トリプルの抽出に課題があることが明らかとなった．

9.2. 発話意図分類モデルの妥当性

発話意図分類モデルの有無で影響度と関与度の算出と混合行列から正解率を算出した．影響度の比較から教師データを増やすことで人手による分類結果に収束することが明らかとなった．また，正解率が高いことから発話意図分類モデルの定義が妥当であると考えられる．しかし，名詞で終わる語尾については誤った分類が行われたことが明らかとなった．

9.3. 関連研究[2]との比較

(1) 発話意図の抽出

関連研究[2]では文書からトリプルを抽出するには形態素解析器を使用することに留めていた．さらに，人手による抽出が必要であると示している．本稿ではトリプル抽出方法を議事録へ適用し評価したことにより，人手に頼らずトリプルを抽出することを示した．

(2) ステークホルダ分析

関連研究[2]では話題毎の分析は行っていないが，本研究ではプロパティグラフモデルの特性を利用し，話題毎に分析可能になることを示した．

10. 今後の課題

(1) 発話内容からトリプルを抽出するときに，ノードを一意に特定するためにプロパティを用いてデータを付与することが必要である．

(2) 発話意図分類モデルで語尾情報を頼りに設定したが，名詞で終わる語尾は人手による分類結果と一致しないため，語尾で終わる文書データを学習データとして設定する必要である．

11. まとめ

本稿は教師データを基にステークホルダの意図の抽出方法と係り受け解析器を用いて議事録を構造化し，ステークホルダを特定する方法を提案した．

文書から発話者の意図を抽出するために発話内容の語尾に注目して発話意図分類モデルを定義した．また，形態素解析器と係り受け解析器を用いて議事録の構造化を行い，グラフデータベースで表現した．さらに，この構造を発話者の役割と発話内容から抽出したノード間の関係を分析し，発話者の関与度と影響度を示すことでステークホルダ分析支援プロセスの有効性を検証した．

参考文献

[1] CaboCha, Yet Another Japanese Dependency Structure Analyzer, https://taku910.github.io/cabocha/ .

[2] 藤本 玲子 ほか，セマンティックグラフモデルによるデータ駆動要求獲得方法の提案とステークホルダ分析への適用評価，情報処理学会 SES2016 論文集，Sep. 2016, pp. 179-186.

[3] 福本 淳一 ほか，日本語文章の構造化解析，情報処理学会研究報告，NL-85-11，Sep. 1991，pp. 81-88.

[4] Jubatus, http://jubat.us/ja/.

[5] MeCab, Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/.

[6] Neo4j, https://neo4j.com/.

[7] W. Richert, et al., Building Machine Learning System with Python, Packt Pub., 2013.

[8] I. Robinson, et al., Graph Databases, O'Reilly, 2013.