口語会話文からの話題推定手法の提案

芋野 美紗子　　土屋 誠司　　渡部 広一
大同大学　情報学部
同志社大学　理工学部
{mimono}@daido-it.ac.jp
{stsuchiy, hwatabe}@mail.doshisha.ac.jp

1　はじめに

本稿では，ある一つの話題についての口語会話文を入力し，そこから話題を表す語（話題語）を出力する手法を提案する．話題とは会話の中に継続して存在している主軸の事を指し，例えば発話内で繰り返し出現する語句などはその会話群の話題である可能性が高い．しかし一方で話題とは，会話を形成している発話群全体から新たに想起される語句で表現される場合も多い．「食事」という語句が発話中に出現していなかったとしても,「好きな料理」や「昨夜のメニュー」などについての会話は食事の話題であると言える．会話文から全体に共通する要素を自動的に抽出し，それを表現する語句を新たに想起する機構があればより人間らしい話題推定を行うことができると考える．提案手法は入力を自然な口語会話とし，文法の崩れた発話から得られる名詞群のみからの話題語想起を行う．話題となる語を会話文中以外の語句からも出力するために，語概念連想システム [1, 2] という語の意味定義と関連性の定量化を行う機構を活用する．

2　語概念連想システム

語概念連想システムとは，人間のように柔軟な語句の意味理解を行うことを目指す機構である．語の意味を定義した「概念ベース」および関連性の定量化を行う「関連度計算方式」により語概念連想システムは構築される．

2.1　概念ベース

概念ベース [1] は複数の電子化国語辞書などの見出し語を概念として定義し，人間が持つ概念への常識的な知識をモデル化した知識ベースである．ある概念の意味定義は，属性と呼ばれる他の概念群と属性それぞれの重要さを表す重みによってなされる．概念ベースの具体例を表 1 に示す．

表 1: 概念ベースの具体例

概念	属性
野球	(本塁打,1.06) (硬球,1.00)…
本塁打	(長打,1.95) (安打,0.87)…
投手	(完投,1.78) (失投,1.78)…

概念ベースにおいて属性に現れる語句は，全て概念として定義されている．そのため，概念「野球」の意味定義を行う属性「本塁打」も概念ベースにおいて意味定義がなされている．このような意味定義の連鎖的な構造により，より人間らしい語句の意味定義が可能となる．図 1 に概念ベースの連鎖構造の例を示す．

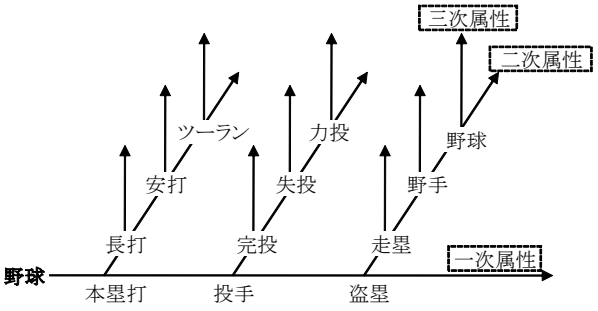


図 1: 概念ベースの連鎖構造

2.2　関連度計算方式

関連度計算方式 [2] は概念と概念の関連性を関連度とよばれる数値で定量的に表現する手法であり，その有効性が示されている．関連度は 0.0 から 1.0 の値を

取り，概念間の関連が強いほど大きな値を示す．関連度の具体例を表 2 に示す．

表 2: 関連度計算の例

概念 A	概念 B	関連度の値
野球	野球	1.00
野球	バット	0.11
野球	叔母	0.00

関連度は概念同士の属性の対応により算出される．互いが持つ属性の内，最も意味が近いもの同士の組を作った上でそれぞれの重みを用いて関連度を算出する．

3　口語会話文からの話題推定手法

提案手法はある一つの話題についての口語会話文を入力し，そこから話題を表す語（話題語）を出力する．提案手法の流れを図 2 に示す．

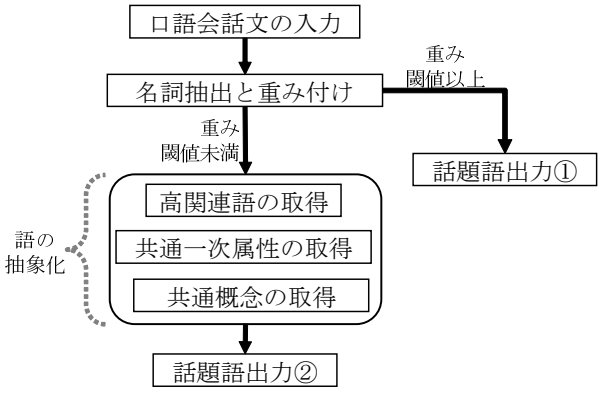


図 2: 提案手法の流れ

まず口語会話文から形態素解析により名詞を抽出し，それらに対し TF・IDF による重み付けを行う．重みに対して閾値を定め，この値を超えた名詞は話題語として出力する．閾値を超える語が存在しない場合は三つの処理を行い「語の抽象化」を行う．

3.1　口語会話文の入力

入力はある一つの話題についての，複数人による口語会話文とする．話題が継続している限りは一つの入力として扱うため，文量に制限はない．本稿においては入力する口語会話文を名大会話コーパス [3] より抜粋した．口語会話文の例を図 3 に示す．

F002:まあ、新人王、は、取るだろうと思ってたけど。(全部はねえ)あとゴールデングラブ、あ、ゴールドグラブ賞は取るだろうと思ってたけど、首位打者、盗塁王、MVPはどう、どうか。ほかは見てないから、わからなかったんですけどね。でも、ご覧になってたんでしょう。あの、何かヤンキースに簡単に負けたとかいうか(そうそうそう)言ってらっしゃった。
F066:全部見てま、全試合は見てなかったけど、かなり見てましたけどね。午前中、最近のほら、よくあの、放送やってるじゃないですか。
F002:ああ、ああ、ああ。私は、あの、夜の短縮版。(ああ)ハイライトの方をよく見てましたけど。やっぱりすごいあれだけアメリカ行っていきなり、ね、何でも取っちゃってすごいなあと思って。　…

図 3: 口語会話文の例

名大会話コーパスの口語会話文は一つの話題が継続したものではないため，文中から目視で同じ話題の続く会話部分を抜粋している．抜粋した会話文に対して話題を表す語を付与し，それを 3 名の被験者で評価した．抜粋した会話文と付与された話題を表す語が正しいと 3 名全員が判断した場合のみを採用する．図 3 に示した例であれば話題は「野球」となった．

3.2　名詞抽出と重み付け

入力された口語会話文から名詞を抽出し，$TF \cdot IDF$ による重み付けを行う．

重み付けの文書集合は名大会話コーパスに収録されている全会話文とした．この時，名詞に付与した重みが閾値 1.0 を超えていた場合はその名詞を話題語として出力し，処理を終了する．例えば，図 4 に示す会話文を入力した場合には文中の名詞「バーゲン」に付与される重みが 4.08 となった．この場合，話題語として「バーゲン」を出力する．

F101:うん、(うん)何かさ、買おう、買おうと思ってなんかいろいろあるけどー。
F093:今、バーゲン？
F101:バーゲンやってたよ、すごい。
F093:ねえ。
F101:やってたけどね、なんかバーゲンでも買えねーやっていうのがいっぱいあったからー、(うん)なかなか買えんかったー。
F093:私もバーゲン行こうかなっと思ったけど、いや、金がない、そう言えばと思って。
＜笑い＞
F101:なんかバーゲンじゃなくってー、(うん)なんかバーゲンじゃないものーの方が欲しくなったりとか、(あー、あー)今、欲しいんだよね。　…

図 4: 話題語「バーゲン」の会話例

$TF \cdot IDF$ による重みが著しく大きいということは口語会話中にある頻出する語が存在し，他の会話中ではその語が出現しづらいことを示す．特に出現頻度は話題の決定において重要であると考える．会話中で何度も繰り返される語が存在する場合，その語を話題の中心として会話が行われている可能性が高いと考えられる．

3.3 語の抽象化-高関連語の取得

重み付けにより話題語が取得できなかった場合は，語の抽象化を行う．抽象化とは会話文中に出現する語の集合から共通要素を抽出し，その共通要素を持ち合わせた別の語を概念ベースを用いて想起することを指す．会話全体が共通して背景に持っている要素が話題であると考え，まずその共通要素が何であるかを自動的に判別する．その上で共通要素をすべて説明できる語を想起すれば，それが話題となるのではないかと考えられる．

まず，3.2 節で付与した重みが最も大きい名詞と会話中のすべての名詞との関連度を算出する．この関連度が高い上位 5 語を高関連語として取得する．重みが大きい名詞は会話中に存在する共通要素を持っていると考えられるため，その名詞と関連の強い他の名詞を取得することで同じく共通要素を持つと思われる名詞の数を増やす．

例として図 3 に示した会話文では，最も重みの大きい名詞として「グラブ」が得られる．この「グラブ」と会話文中のすべての名詞との関連度を算出した結果，上位 5 語は「グラブ，打者，盗塁，選手，間」となった．これらを高関連語とする．

3.4 語の抽象化-共通一次属性の取得

3.3 節で取得した高関連語の一次属性を概念ベースを用いて展開し，共通する属性（共通一次属性）の取得を行う．処理の例を図 5 に示す．

3.3 節で取得した高関連語「グラブ，打者，盗塁，選手，間」の一次属性をすべて展開する．その上で多くの高関連語が共通して持つ一次属性を共通一次属性として取得する．図 5 では「球技」という属性が高関連語のうち 3 語に出現している．同様にすべての一次属性について共通して出現する高関連語の数を調べ，その数が多い上位 5 語を取得する．共通して出現する高関連語の数が同じ一次属性が取得された場合は 3.3 節で算出した関連度が高い高関連語の属性を優先する．

概念 (高関連語)	一次属性
グラブ	野球，ミット，グローブ，フェンシング，**球技**…
打者	打球，直球，バッター，**球技**，代打，走者…
盗塁	ホームスチール，走者，盗塁，走塁，**球技**…
選手	スプリンター，入団，選び抜き，走者，先発…
間	食間，間延び，眉間，波間，晴れ間，京間，欄間…

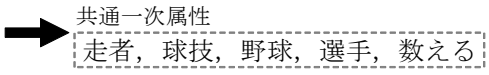


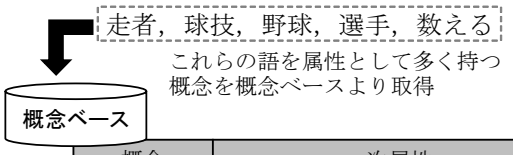
図 5: 共通一次属性の取得例

この例では最終的に共通一次属性として「走者，球技，野球，選手，数える」が得られる．

高関連語は共通要素を持つと思われる名詞群である．その名詞の意味は概念ベースにおいて各々が持つ属性によって定義されている．高関連語が持つ意味のうち，出来るだけ多くに共通して出現するものは共通要素を表すと考えられる．

3.5 語の抽象化-共通概念の取得

共通一次属性は会話文中に存在する共通要素を表す語群であるため，その共通要素で意味定義されている概念はつまり話題を示す語ではないかと考えられる．そこで 3.4 節で得られた「共通一次属性」を多く属性として持つ概念を共通概念として取得し，得られた語を最終的な話題語とする．処理の例を図 6 に示す．



概念	一次属性
生還	争う，還る，**選手**，要塞，**走者**，勝負，**野球**，帰陣，**球技**…
打者	打球，バッター，**走者**，打順，**野球**，**球技**，本塁打，**選手**…
野球	**野球**，**球技**，猛打，盗塁，**選手**，外野，**走者**，外角…
…	…

図 6: 共通概念の取得例

3.4 節で取得した共通一次属性は「走者，球技，野球，選手，数える」である．図 6 に示す通り，例えば生還，打者，野球といった概念はこれらの共通一次属性のうち 4 語を属性として持っている．この数を概念ベース内すべてで調べ，最も多く共通一次属性をもつ概念を最終的に話題語として出力する．共通一次属性の所持数が同値の場合はその概念をすべて話題として出力する．

4　評価

名大会話コーパスの口語会話文から目視で同じ話題の続く会話部分を抜粋したものを 50 セット用意し，それらを用いて評価を行う．会話文には 3 名の被験者による目視によって評価された「正解の話題語」が付与されている．提案手法では最終的に複数の話題語が出力される可能性もあるため，評価としてすべての出力に対する評価（精度）と正解の話題語を出力できているかの評価（再現率）の二つを算出する．

3 名の被験者に対して入力会話文と出力された全ての話題語のセットを見せ，話題語に対して適する順に 2 点，1 点，0 点の三段階評価を付けてもらう．3 名分の評価を合算し，合計点が 4 点以上の話題語を○，2 点および 3 点を△，0 点および 1 点を×としてこの割合を精度とした．

表 3 に評価結果を示す．

表 3: 評価結果

精度○	精度△	精度×	再現率
40.9 %	15.7 %	42.6 %	56.0 %

結果として精度は○が 40.9 %，再現率は 56.0 %となった．再現率に比べ精度が低くなったが，これは 3.3 節以降の語の抽象化処理で多くの話題語が出力された場合に雑音も多く含まれているためである．例えば図 3 に示した会話の場合，正解の話題語「野球」は最終出力に含まれていたため再現率は増加するが，それ以外にも 16 語が話題語として出力されていた．出力された話題語を図 7 に示す．

全てが野球に関連する語ではあるが，話題としては適切でないものが多い．精度評価において○となった話題語は野球のみであり，5 語が△，残りは×となった．全く関連のない語が出力されているわけではないが，話題を表す語として適切なものをさらに選別する必要がある．

出力話題語数：17語

生還，選手，走塁，打者，投手，盗塁，走者，捕手，**野球**，刺殺，球審，憤死，力投，トス，球児，スターティングメンバー，スパイク

図 7: 出力された話題語

5　まとめ

本稿ではある一つの話題についての口語会話文を入力し，そこから話題を表す語（話題語）を出力する手法を提案した．会話文中の語から共通要素を自動的に抽出し，概念ベースの連鎖構造を利用することでそれらの共通要素をできるだけ持ち合わせた語を新たに想起する．この処理により会話文中には出現していない語でも話題として出力することが可能である．ただし，提案手法では会話に関連のある語は出力できても，それが話題を表す語として適切であるかの判断がまだ不足していると分かった．今後は人間がどのような語を話題として適切と感じるのか，そのメカニズムを調査した上で話題語の精錬処理を追加する必要がある．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けて行ったものです．

参考文献

[1] 奥村紀之, 土屋誠司, 渡部広一, 河岡司. 概念間の関連度計算のための大規模概念ベースの構築, 自然言語処理, Vol.14, No.5, pp.41-64, 2007.

[2] 渡部広一 奥村紀之 河岡司. 概念の意味属性と共起情報を用いた関連度計算方式, 自然言語処理, Vol.13, No.1, pp.53-74, 2006.

[3] 藤村逸子 大曽美恵子 大島ディヴィッド義和. 会話コーパスの構築によるコミュニケーション研究, 藤村逸子 滝沢直宏編,『言語研究の技法：データの収集と分析』, ひつじ書房, pp.43-72，2011.