発話生成を目的としたWikipediaからの文抽出

Sentence Extraction from Wikipedia for Making Utterance for Dialogue System

太田 知宏　鳥海 不二夫　石井 健一郎
Tomohiro OTA　Fujio TORIUMI　Kenichiro ISHII

名古屋大学大学院 情報科学研究科
Graduate School of Information Science, Nagoya University

We propose a method dor extracting sentence from Wikipedia that makes dialogues lively and vivid. Sentences which contain unexpected topics seem to be suitable to this purpose. In order to extract such sentences, we rank sentences by using TF-IDF, collocation, and the number of morphemes. We evaluated the validity of the method to rank sentences. The ratio of articeles that include suitable sentence in the top 10 is 87.2%.

1. はじめに

近年，人間とコミュニケーションを行うコンピュータに対する期待が高まり，対話システムに関する研究が広く行われている．対話システムは，タスク指向型対話システムと非タスク指向型対話システムの二種類に大きく分けられる．

タスク指向型対話システムは，何らかのタスクを達成するために対話を行うシステムである．例えば，チケット予約受付や情報検索を行うための対話システムはこれにあたる．

非タスク指向型対話システムは，対話すること自体が目的であり，人間を楽しませるための対話（雑談）を行うシステムである．例えば，人工無能（chatterbot,chatbot）と呼ばれるような対話システムの多くはこれにあたる．

非タスク指向型対話システムの目的である，「相手を楽しませるような対話」を実現するためには，対話システムが対話を盛り上げるような発話を行う必要がある．しかしながら，現在の非タスク指向型対話システムの多くは対話が成り立つことのみを目的としており，対話を盛り上げるというところまで至っていない．

その理由として，必要な知識量の問題が挙げられる．タスク指向型対話システムは特定のタスクを達成することを目的としており，話題を限定できるので，最小限の知識を備えていればよい．一方，非タスク指向型対話システムは相手を楽しませる対話を実現することが目的となり，話題を限定できないので，必要な知識が膨大となる．

対話の名手と言われるインタビュアには，いくつかの特徴がある．その一つに「話が思わぬ方向に発展させるような話題を持ち込むこと」がわかっている．そのために，インタビュアはまず，対話の前に相手について十分調査を行っている．インタビュアはこの周到な調査により，相手の好む話題や，対話を盛り上げるために必要な情報を得ることができる．その情報を利用して，インタビュアは相手に意外だと思わせるような発話や，相手に話を促すような質問を行っている．

対話を盛り上げる対話システムを設計するため，非タスク指向型対話システム「インタビュア KELDIC（Ken's Laboratory Dialogue Computer）」の研究を進めている [1]．インタビュア KELDIC ではテキスト対話を対象としている．

インタビュア KELDIC の研究では対話システムが熟練したインタビュアのように対話を盛り上げるような発話を行うことを最終的な目標とする．インタビュア KELDIC では，インタビュアと同様に，相手のプロフィールを予め入手し質問を作成しておく．そこで，本論文では「Wikipedia」*1から対話を盛り上げるような文を発話候補文として抽出することを目的とする．入手したプロフィールからキーワードを抽出し，Wikipedia から発話候補文を抽出し，適切な形に文を修正して発話を行う流れになるが，本論分では発話候補文の抽出のみを対象とする．

話を思わぬ方向に発展させるような話題を持ち込むためには，Wikipedia の記事から，一般常識的な内容ではなく意外性のある内容の文を取り出し，その文を利用して対話を発展させていく事が考えられる．本論文で対象とする文は，そのように一般常識的な内容ではなく意外性のある内容の文である．また，Wikipedia を用いる利点は，記事の種類が幅広く，多くの固有名詞についても取り上げられていることや，日々発生する新しい単語についても，随時追加されていることが挙げられる．

2. Wikipedia からの文の抽出手法

2.1 文の抽出処理の流れ

Wikipedia から文を抽出するための処理の流れを示す．まず，前処理として，Wikipedia のデータから記事を一つ取り出す．そして，記事に対しタグを取り除く処理を行った後に形態素解析を行う．

ここで，Wikipedia から文を抽出するために，Wikipedia の各文には順位付けを行う．順位付けを行うために各文に対して文のスコアをつける．文のスコアには三つの指標を組み合わせて使用する．その三つの指標とは，TF-IDF，語の共起，文の長さである．前処理を行った後は，この三つの指標を計算し，各文に対してスコアをつける．

そして，発話生成に不適切な文を取り除いた後に，各文をスコアに従って順位付けを行う．順位付けで上位になる文ほど発話生成に利用できると考え，それらを出力する．

以下，それぞれの処理について順に説明をする．

2.2 前処理

Wikipedia から文を抽出するために，Wikipedia のデータから記事を一つ取り出す．本論文では，処理を行う記事の選択は人間が行う．

連絡先: 太田 知宏，名古屋大学大学院 情報科学研究科，E-mail: ohta@kishii.ss.is.nagoya-u.ac.jp

∗1　日本語版 Wikipedia:http://ja.wikipedia.org/

本研究では，ウィキメディア財団より提供されているデータベース・データ [*2] を利用する．利用するデータは日本語版 2008 年 10 月 19 日の時点のものである．

記事に対して，Wikipedia で文章の構造を記述するために使用されている様々なタグを取り除く．

タグを取り除いた後に，文と文を区切る．文の区切りには改行，もしくは「。」「！」「？」「!」「?」を使用する．区切り文字として「.」を使用しない理由は,「.」を含む固有名詞が存在することを考慮するからである．なお，Wikipedia は日本語版を利用するため，文末が「.」で終わる文は少なく影響は小さい．

文を区切ったのちに，各文に対して形態素解析を行う．形態素解析には日本語係り受け解析器 CaboCha（南瓜）[2] を利用する．形態素解析を行うことで文を単語に分割し，単語の品詞を得ることができる．

2.3 TF-IDF を用いた文のスコア付け

TF-IDF は，文章中の特徴的な単語を抽出する評価式で，重要語に注目した重要文抽出などに用いられている．本論文においても，Wikipedia の記事中の文から重要な文の評価を高くするために，TF-IDF を用いる．

文書 d における単語 t の TF-IDF 値 $w(t,d)$ を以下のように定義する．

$$w(t,d) = tf(t,d) \cdot idf(t)$$

ここで，$tf(t,d)$ は文書 d における単語 t の出現頻度であり，以下のように定義する．

$$tf(t,d) = \frac{n(t,d)}{\sum_{k \in d} n(k,d)}$$

上式の $n(t,d)$ は単語 t の文書 d における出現回数である．また，$idf(t)$ を以下のように定義する．

$$idf(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

上式の $|D|$ は対象とする文書集合の総数である．また，$|\{d : t \in d\}|$ は単語 t の出現する文書の総数である．

ここで，idf は一般的な単語，例えば「こと」「もの」など，の評価値を下げるための値であり，分野に偏りがなく様々な文章から得られたデータであることが望ましい．そこで，$|\{d : t \in d\}|$ の値として，Wikipedia という限定された文章から得るデータではなく，Web 日本語 N グラム [3] の 1-gram に収められている単語の出現回数を用いる．また，それに伴い文書集合の総数として $|D|$ には十分大きな値として 10^{10} を代入する．

文 S_i が与えられたとき，文中の名詞・固有名詞について $w(t,d)$ を求め，その平均を文のスコア $TI(S_i)$ として以下のように定義する．

$$TI(S_i) = \frac{\sum_{k \in S_i} w(k,d)}{\sum_{k \in S_i} n(k,S_i)}$$

*2 日本語版データベース:http://download.wikimedia.org/jawiki/

2.4 語の共起を用いた文のスコア付け

語の共起を表す指標として，本論文では共起回数 (共起頻度) を用いる．共起回数の計算には，Web 日本語 N グラムの 7-gram を使用する．ある中心語に対する共起語を計算する場合は，中心語の前後 6 語以内に出現する語の出現回数を計算する．そのためには，7-gram のデータから 1 番目もしくは 7 番目に中心語が存在するデータを取り出し，共起語が存在するデータの出現回数を合計する．

例えば,「明日」を中心語とし，共起語を「天気」とした場合の共起回数を計算する事を考える．まず，1 番目に「明日」があり，2 番目から 7 番目に「天気」があるデータの出現回数を全て合計する (図 1)．7 番目に中心語が存在する場合にも同様に合計する．

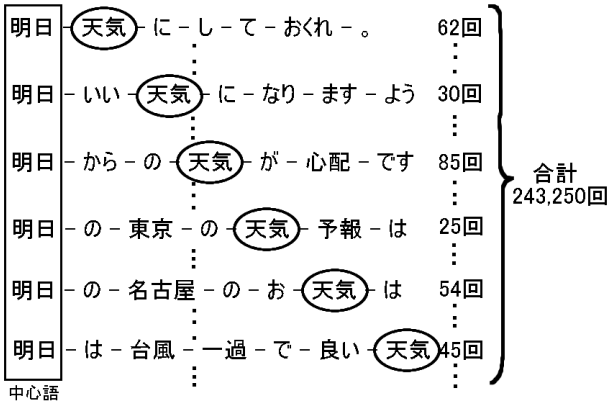

図 1: 1 番目に中心語があるデータの共起語の共起回数

ここで，共起回数が多い単語によって成り立つ文は，人間にとって陳腐で平凡な内容であると考えられる．また，共起回数が少ない単語から成る文は，人間にとって意外性があると考えられる．人間にとって陳腐で平凡な内容の文よりも，意外性のある内容の文の方が，興味を引きやすく，対話が盛り上がりやすいと考えられる．そこで，共起回数が多い単語を含む文のスコアを下げつつ，共起回数が少ない単語を含む文のスコアが高くなるように，スコア付けの式を定義する．ただし，共起回数が 0 回の単語は共起語として数えない．

文 S_i が与えられたとき，一文のスコア $C(S_i)$ を以下のように定義する．

$$C(S_i) = \begin{cases} (c_{max} + 1 - c_i)/c_{max} & (c_i \neq 0) \\ 0 & (c_i = 0) \end{cases}$$

ただし，c_i は文 S_i 中にある全ての共起語の共起回数の平均である．また，c_{max} は文章中に現れる c_i の最大値である．

2.5 文の長さを用いた文のスコア付け

Wikipedia には説明的な文が多く，中には非常に長い文が含まれている．対話において長すぎる文は理解に時間がかかり，対話の流れを悪くすると考えられる．そこで，本論文では文の長さとして形態素の数を用い，人間同士の対話で現れやすい長さの文の評価が高くなるようにする．

ここで，人間同士の対話は挨拶や質問，相槌など様々な種類が考えられる．しかし，1 語や 2 語でも意味が伝わる挨拶や相槌などは文が短くなりやすく，伝える情報が多い質問などは比較的文が長くなると考えられる．そこで，文の種類を制限す

るため，人間同士の対話に対し，発話の種類を記述するタグである SWBD-DAMSL（Discourse Annotation and Markup System of Labeling）タグ [4] が付与されたデータを用いる．対話は人間同士がテキストで行った 59 対話を利用する．

ここで，Wikipedia から本研究で目的とする文を抽出し，発話を生成した際に当てはまるタグは，主に客観的な事実である sd と主観的な意見である sv の 2 つであると考えられる．そこで，人間同士の対話で sd もしくは sv というタグが含まれる発話について，形態素の数毎に発話の出現回数を調べた．対話データのうち sd もしくは sv というタグが含まれる発話の総数は 1864 個であった．形態素の数と文の出現回数の関係は**図 2** の通りである．

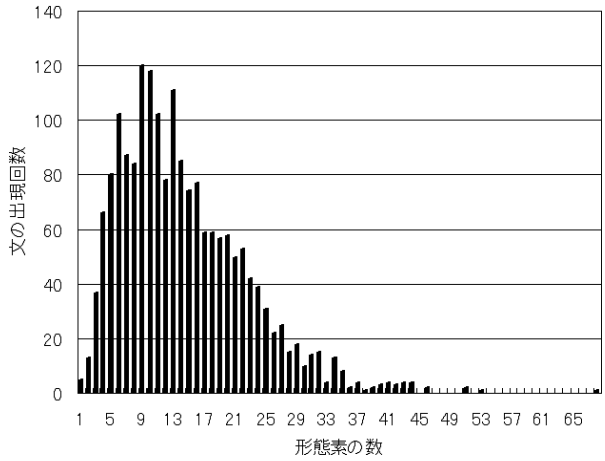

図 2: 人間同士の対話における形態素の数による文の出現回数

文のスコアとして「人間同士の対話における形態素毎の発話の出現回数」を「発話の総数」で割った出現頻度を使用する．すなわち文 S_i が与えられたときの一文のスコア $P(S_i)$ を以下のように定義する．

$$P(S_i) = \frac{n(l(S_i))}{N}$$

ただし，$l(S_i)$ は，文 S_i が含む形態素の数である．$n(l(S_i))$ は形態素の数が $l(S_i)$ 個の発話の，59 対話中の出現回数（**図 2**）である．そして，N は sd もしくは sv というタグが含まれる発話の総数（1864）である．つまり $P(S_i)$ は，人間同士の対話のうち各形態素の数の発話がどの程度の割合で現れたかを示す．

2.6 発話として不適切な文の除去

Wikipedia では箇条書きや表組みなど様々な表記の方法がある．これらは文として成り立っていないものが多く，発話生成には不適切である．また，一文を発話生成に利用するためには，一文のみで意味が理解できなければならない．

まず，以下のような特徴を持つ文は，文として成り立っていない，文中の語が省略されている，などの理由から一文のみで意味が理解できず，発話生成に不適切である．

- 末尾が読点やピリオドではない
- 助詞の「が」もしくは「は」を含まない
- 体言止めで終わっている

また，以下のような特徴を持つ文は，日本語としては正しいが，文脈がなければ一文で意味が理解できないため，発話生成に不適切である．

- 接続詞で始まる
- 指示語を含む
- 特定の単語を含む

2.7 文の順位付け

節 2.3 から節 2.5 で定義した各文に対するスコア付けを組み合わせて，各文に対してスコアをつける．文 S_i に対するスコア $SS(S_i)$ を以下のように定義する．

$$SS(S_i) = TI(S_i) \cdot C(S_i) \cdot P(S_i)$$

この $SS(S_i)$ を Wikipedia の 1 つの記事に含まれる全ての文に対し計算する．そして，節 2.6 で挙げた特徴を持つ文を取り除いた後，スコアが高い順に文を並べる．上位の文ほど発話生成に適した文であると考える．

3. 文の抽出実験

3.1 実験目的と実験方法

本論文で提案した手法を用いて抽出した文が，発話を盛り上げるような文であることを確認する．

まず，Wikipedia の 86 個の記事について，各記事内の文を提案手法で順位付けを行う．順位付けを行ったときにスコアが 0 よりも大きく，かつ上位 30 位以内に入る文について発話生成に適しているかどうかを人間が判断する．ただし，項目によっては文の数が 30 に満たないものも存在する．その場合は存在する順位までの文を判断する．

順位付けを行った各文に対する判断は 20 代の大学生 5 人が行う．1 つの記事に含まれる文について 1 人が担当する．

発話生成に適しているかどうかを判断するにあたって次のことを前提とする．

- 順位付けを行った各記事の題名は既知とする

そして，以下の条件を満たす文を対話を盛り上げるような文であると判断する．

- 内容が興味深い，もしくは意外性がある
- 内容が一般常識的なものではない

例えば，Wikipedia の「ネコ」の記事に「日本の平安時代には位階を授けられたネコもいた。」という一文がある．この文は意外性があり，対話を盛り上げるような文であると言える．逆に，Wikipedia の「トマト」の記事に「トマトは、緑黄色野菜である。」という一文がある．しかし，トマトが緑黄色野菜であることは広く知られている事実であり，対話を盛り上げるような文であるとは言い難い．対話では,「ネコが好きとのことですが，日本の平安時代には位階を授けられたネコもいたそうですよ。」などとして使用する．話題を提起する際などに，ただ「ネコが好きとのことですね。」と発話を行うよりも，対話を盛り上げることができると考える．

3.2 実験結果

各順位以内に，対話が盛り上がるような文があると判断された記事の割合は図 3 の通りである．1 位では 34.9% の記事に対話が盛り上がるような文があると判断された．また，10 位以内では 87.2% の記事に対話が盛り上がるような文があると判断された．

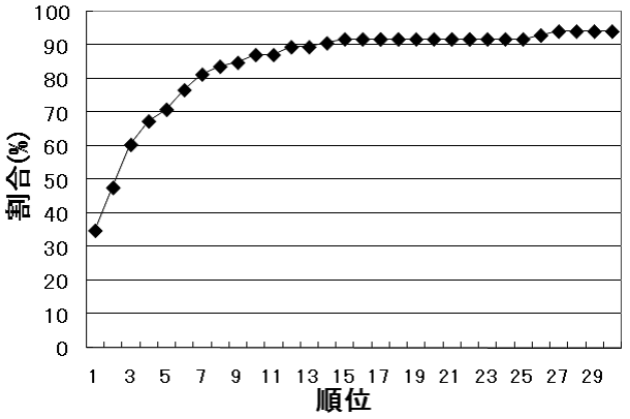

図 3: 各順位以内に対話を盛り上げるような文があると判断された記事の割合

実際に，対話が盛り上がると判断された文には,「イヌは最も古くに家畜化された動物である。」,「ミュージカルでは基本的にポピュラーソングと同じ発声法が好まれ、用いられる。」,「語源はチェコ語で『労働』を意味する robota とされている。」(「ロボット」の記事) などがあった.

3.3 考察

実験の結果から，1 位に対話が盛り上がるような文がある記事の割合は約 35% 程度の精度しか得られないという結果になった．その理由として，TF-IDF，語の共起，文の長さの各手法の重みが考慮されていないということが考えられる．本論文では TF-IDF，語の共起，文の長さによりそれぞれ文にスコアを付け，その積を使用して文の順位付けを行った．そのため，各手法が効果的であるかどうかが不明のまま使用されているという問題がある．

一方，10 位以内には，約 87% の記事で，対話を盛り上げるような文が含まれているという結果となった．対話システムがこの中から効果的な文を自動的に選ぶ処理を行うことで，対話を盛り上げていくことが可能である．

4. まとめと今後の課題

本論文では対話を盛り上げるのにふさわしい発話文候補を Wikipedia から抽出する手法を提案した.

抽出された文が対話を盛り上げるような文であるか否かを人間によって判断した．その結果，10 位以内に対話を盛り上げるような文があると判断された記事は 87.2% であった．よって，Wikipedia から対話を盛り上げるような文を抽出できることが確認された.

実験の結果から，いくつかの課題が明らかになった．Wikipedia の記事から発話生成に適した文を抽出する精度を向上させるための方法として，いくつかの方法が考えられる.

第一に，不適切な文を取り除く精度を向上させることが考えられる．そのためには，以下の 2 つの方法が考えられる.

- 一文のみで意味が理解できない文を取り除く精度を上げる
- 一文のみで意味が理解できない文を意味が理解できるように修正する

一文のみで意味が理解できない文を意味が理解できるように修正する方法として，まず照応解析をすることが考えられる．照応解析とは，省略された名詞句の補完や，代名詞や指示詞などの照応詞の指示する対象を推定することである．日本語では主語の省略が頻繁に起こり，それは Wikipedia の記述でもしばしば確認された．省略された主語を補完し，代名詞や指示語などを対象である名詞と置き換えることで発話生成に適した文の割合が増加すると思われる.

他の方法として，Wikipedia の記述にしばしば見られる章や節の題名を考慮することが考えられる．Wikipedia 固有の方法ではあるが，記述全体の題名だけでなく章や節の題名を考慮することは意味が理解できない文の割合を減らす効果があると考えられる．これは照応解析の一部とも考えられるが，文頭に題名を追加するという程度の実装ならば比較的容易であることが利点である.

第二に，順位付けの式の改善が考えられる．本論文では各手法によるスコア付けに対し重みが付けられていなかったため，精度向上のためには重みを付けることが挙げられる．また，他の素性を追加する事も挙げられる.

その実現の手段として Support Vector Machine（SVM）を用いることが考えられる．平尾らにより SVM を用いた重要文抽出法が提案されており [5]，少ない学習データでも重要文抽出法において有効であるとされている．本研究で抽出する文は重要文とは異なるが，文抽出という点では共通しており，高い効果が得られることが期待される.

これらの手法を実装し，より精度を高める事が今後の課題である.

参考文献

[1] 岡田 譲二 , 鳥海 不二夫 , 石井 健一郎 : インタビュアを模した対話エージェントのための質問文自動生成 , Proceedings of JAWS 2007 (2007)

[2] 工藤 拓、松本 裕治:チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌 (2002)

[3] 工藤拓, 賀沢秀人 : Web 日本語 N グラム第 1 版, 言語資源協会発行

[4] Jurafsky, D. and Shriberg, E. and Biasca, D. : Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, Institute of Cognitive Science Technical Report (1997)

[5] 平尾 努, 磯崎 秀樹, 前田 英作, 松本 裕治 : Support Vector Machine を用いた重要文抽出法 (自然言語), 社団法人情報処理学会 (2003)