E-030

コンピュータ会話のためのニュース記事見出し抽出手法

Extraction Method of News Headlines for Computer Conversation

栢下　洋一† 吉村　枝里子‡ 土屋　誠司‡ 渡部　広一‡

Yoichi Kayashita Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，人間と円滑な会話を行うコンピュータの開発が注目されている．人間同士の日常会話では時事情報を基にした会話が頻繁に行われている．そのためコンピュータとの会話においても時事情報を用いることが必要だと考えられる．そこで本研究では，ニュース記事の見出しを時事情報として扱い，話者の発話に対して関連した内容のニュース記事の見出しを抽出する手法を提案する．例えば発話者が「京都で地震があったね．揺れが激しくて恐かったです」と言った場合，‘京都’や‘地震’という語に関係するニュース記事の見出しを取得する．そして，語と語の関連性を定量化することで，記事の見出しと発話内容の関連性を比較し，最も発話内容と関連の高い記事の見出しを抽出する．

2. 関連事項

2.1 概念ベース

概念ベース[1]とは電子化された国語辞典や新聞等から機械的に構築した，語（概念）とその意味特徴を表す単語（属性）の集合のセットを約 12 万語蓄積した知識ベースである．ある概念 A は m 個の属性 a_i と重み w_i（>0）の対によって(1)式のように定義される．

概念 $A=\{(a_1, w_1), (a_2, w_2), \cdots, (a_m, w_m)\}$ (1)

（例.「雪」＝{(雪, 0.61), (白い, 0.30), ･･･, (下る, 0.01)}）

2.2 関連度計算方式

関連度計算方式[2]は概念ベースに定義された語と語の関連の強さを，定量的に評価するものである．関連度の値は 0 から 1 までの値で表現される．

2.3 意味理解システム

意味理解システム[3]とは単文入力に対して 6W1H（who, what, when, where, how, why, whom）と用言の 8 個の成分に分けることのできるシステムである．

2.4 記事関連度計算方式

記事関連度計算方式[4]とは，概念ベースと関連度計算方式を利用したものである．関連度計算方式は語と語の関連の強さを定量的に評価するのに対し，記事関連度計算方式は記事(文)と記事(文)の関連の強さを定量的に評価するものである．

3. ニュース記事見出し抽出手法

3.1 ニュース記事見出し抽出手法の流れ

ニュース記事抽出手法の流れを以下の図 1 に示す．

† 同志社大学大学院工学研究科
Graduate School of Engineering , Doshisha University
‡ 同志社大学理工学部
Faculty of Science and Technology , Doshisha University

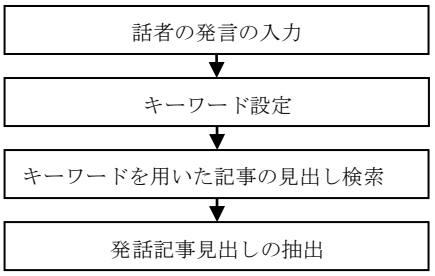

図 1　ニュース記事抽出手法の流れ

まず, キーボートより発話者が発言を入力する. 続いて, その発言に対しシステムがキーワードを設定する．キーワードの設定を行えばそのキーワードを用いて記事の見出し検索を行う．記事の見出しが取得できれば，記事関連度計算方式を用いて取得記事の見出しと発話内容の関連性を定量化し，最も発話内容と関連の高い記事の見出しを抽出する．

3.2 キーワード設定

本研究では会話の内容を表す語をキーワードと定義する．記事検索の際はキーワードを利用する．キーワードは，名詞，形容詞，形容動詞，動詞である．なお，品詞の抽出は形態素解析器「茶筌[5]」を用いる．

3.3 記事見出し検索

記事見出し検索は，時事情報知識ベースを用いて行う．時事情報知識ベースとは，ニュース記事の見出しを格納した知識ベースである．本研究では Web サイトより収集した 2010 年 11 月 8 日から 2011 年 1 月 13 日までの 600 件のニュース記事が格納されている．以下の表 1 に時事情報知識ベースの例を示す．

表 1　時事情報知識ベース

日付	ジャンル	ニュース記事見出し
2010/12/08	海外	韓国，砲撃訓練開始
2010/12/08	スポーツ	イチロー, 10 年連続ゴールドグラブ賞
2010/12/08	国内	あかつき，6 年後に再挑戦

ニュース記事見出しを形態素解析し，それにより得られた各自立語とキーワードとの表記一致を調べる．表記一致した場合，その自立語を含む記事の見出しを関連記事見出しとする．

3.4 記事見出し抽出

3.4.1 記事関連度計算方式による記事見出し抽出

キーワードにより取得した関連記事の見出しの中から，発話に用いるための見出しを一つ抽出する．見出しの抽出には入力文との関連性を考慮することが必要であり，最も関連の高い記事の見出しをコンピュータ側の発話内容として用いることが最適であると考える．そのために，記事関

連度計算方式を用いて入力文と関連記事の関連性を定量化し，関連度の高い記事の見出しを抽出する．

3.4.2 会話の話題を考慮した記事見出し抽出

人間同士の会話では話題が存在し，その話題について会話が展開すると考える．記事関連度計算方式を用いて入力文と関連記事の関連性の定量化を行うが，会話の話題は考慮されていない．したがって，より適切にその関連性を定量化するために，会話の話題である語を含む記事の見出しに対して重みを与える．本研究では，入力文に複数現れる同じ語（例．神戸へ行きました．神戸の街中はきれいですね．→複数現れる同じ語‘神戸’）や，意味理解システムにより抽出された What 格（例．友達と野球をしました→What 格‘野球’）の語は会話の話題となりうると考え，これらの語を含む記事の見出しに重みを与えた．重みづけの手法として，入力文に複数現れる語に関する記事の見出しは実際の記事関連度の値に 1.5 倍，What 格となる語に関する記事の見出しは記事関連度の値に 1.2 倍することで重みを与える．その結果，値が最も高い記事の見出しを抽出する．

4. 評価

アンケートにより入力文となる話者の発言のテストセットを 100 文集めた．ニュース記事の見出しの取得には入力文にできる限り多くの情報が必要であるため，入力文は 2 文以上とした．抽出した記事の見出しが，入力文と関連があり，コンピュータ側の発話内容として適切であるかを調べた．その結果を以下の①～③のグループにそれぞれ分ける．

① 適切な記事の見出しを抽出できている
（被験者 5 名により，適切な記事であると 3 名以上が判断した場合）
② どちらとも言えない
（被験者 5 名の判断で，適切な記事であると 2 名以下が判断した場合）
③ 適切な記事の見出しではない

これらの結果を以下の図 2 に示す．

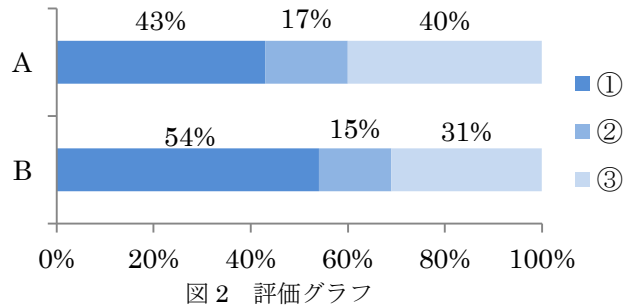

図 2 評価グラフ

図 2 の A に示すグラフは，記事関連度計算方式のみで記事見出し抽出を行った評価結果であり，B に示すグラフは記事関連度計算方式と会話の話題となりうる語を含む記事の見出しに重みを与え記事見出し抽出した評価結果である．B のグラフにおいて記事見出し抽出の成功例と失敗例を以下の表 2，3 に示す．

表 2 記事見出し抽出の成功例

入力文	抽出された記事の見出し
スマートフォンを持つ人が増えましたね．早く手に入れたいです．	スマートフォン，携帯販売台数の 48%
来週東京に行きます．スカイツリーが見てみたいです．	東京スカイツリーが高さ世界 2 位に

表 3 記事見出し抽出の失敗例

入力文	抽出された記事の見出し
今日は成人式ですね．着物姿がきれいです．	ソフトバンクがプリペイド式携帯，コンビニ販売終了へ
KARA のダンスを練習しています．ダンスは難しいです	白鵬 V 翌朝に第 3 子誕生“ゆりかごダンス”で祝杯

5. 考察

会話の話題を考慮し，重みづけを行い記事抽出した結果が，記事関連度計算方式のみで記事抽出した結果と比較すると適切な記事の見出しを抽出できているという割合が全体で 11%向上した．会話の話題となりうる語を含む記事の見出しに重みづけを行い，抽出精度が上がったことから，重みづけの対象とした語が会話の話題となることが多いと考える．ただし，会話の話題となりうる語は他にも考えられ，例えば人名などの固有名詞が挙げられる．有名人は会話に用いられることが多いため，このような語も考慮し重みを与え，入力文と取得した記事の見出しについて適切に関連性を定量化する必要がある．そのため会話の話題として用いられる語はどういう語かさらに調べる必要がある．

6. おわりに

本研究は，発話者の発話内容からニュース記事の見出しを取得し．発話内容と関連の高い記事の見出しを抽出する手法を提案した．その方法として記事関連度計算方式のみで記事抽出する手法と，それに加え会話の話題となりうる語に関する記事の見出しに重みを与え記事抽出する手法を提案した．会話の話題を考慮し重みを与えると，記事関連度計算方式のみによる記事抽出結果と比較して抽出精度が 11%向上した．今後，さらに会話の話題として用いられる語を調査する必要がある．

謝辞

本研究の一部は，科学研究費補助金（若手研究(B)21700241）の補助を受けて行った．

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] 荒木孝允，奥村紀之，渡部広一，河岡司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告，Vol.48，No.3，pp14-24，2007.
[3] 篠原宜道，渡部広一，河岡司，“常識判断に基づく会話意味理解方式”，言語処理学会第 8 回年次大会発表論文集，B6-2，pp.651-654，2002.
[4] 倉田篤史，渡部広一，河岡司，“概念ベースと関連度計算を用いた記事関連度計算方式”，情報処理学会研究報告，2006-NL-171，pp.19-24，2006.
[5] ChaSen - 形態素解析器，奈良先端科学技術大学院大学 http://chasen-legacy.sourceforge.jp/，2011/ 1/ 31.