

自然言語における話題の適合性判断システム

北澤 由翔[†] 奥村 紀之[†]

[†]長野工業高等専門学校 電子情報工学科

1 はじめに

本研究では、ユーザーから提供された話題が、背景知識の内容に対する話題として適合しているか否かを判断する、話題の適合性判断システムを構築する。

2 提案手法

本システムは、ユーザーから話題の入力を受け取り、話題から主題語と話題語を取得する。その語を元に、背景知識から話題となりうる語を抽出し、話題候補語群とする。更に、主題語と話題語、話題候補語群を用いて話題の適合性を判断する。図1に本研究のシステムの構成図を示す。

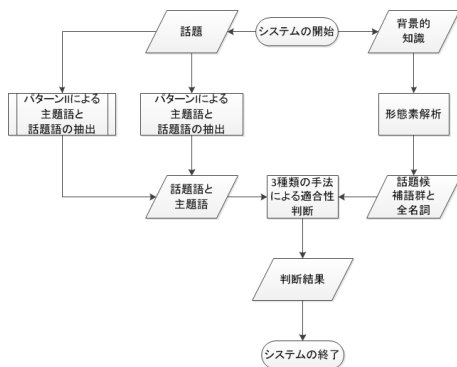


図1: システムの構成図

2.1 背景知識

林は、会話の背景に存在する膨大な知識を「背景知識」と呼称している [1]。本研究では、ある文章を一定量の知識と見立て、人間の知識の代わりに利用した。背景知識に利用した文章の種類は、ニュース、コラム、教科書の3種類である。

2.2 主題語・話題語

野田らはある主題語 p の話題となる語 t は「 p の t 」の形で用いる傾向にあると述べている [2]。よって、本稿での話題は「主題語の話題語」という形式、もしくは日常会話でよく使用される「形容詞+名詞」の形で表されるものとする。

2.3 話題候補語群

背景知識内の主題語が出現している文章内から、名詞・形容詞を抽出し、これを話題候補語群とする。これは、話題は「主題語の話題語」の形で表すことができるため、主題語が出現している文章内の単語は、背景知識内の他単語に比べて話題になり得る可能性が高いと考えるためである。

3 評価実験

本研究では、話題の適合性を判断する手法、適合性判断方法を提案する。また、適合度計算の他に、主題語と話題語の決定方法を提案する。

4 評価方法

システムの評価実験は、3種類の背景知識から作成した、以下の4種類の話題で行う。

1. 文章中の単語を含めた、文章に確実に関係がある話題
2. 文章中の単語から連想できる単語を含めた、文章に関係がある話題
3. 文章中の単語から連想した単語から、更に連想できる単語を含めた文章には関係がない話題
4. 文章に全く関係がない話題

この4種類の評価データにおいて、1, 2の話題は背景知識に対する話題としてふさわしい話題、3, 4の話題は背景知識に対する話題としてふさわしくない話題とする。前者を適合、後者を不適合として評価実験を行う。

4.1 主題語と話題語の決定方法

話題を形成する単語が3つ以上である場合の主題語と話題語の決定方法について提案する。評価データから名詞と形容詞を取得し、取得した順にA, Bとアルファベットを割り当てる。

パターン I

パターン I では、B を主題語、C を話題語とする。

パターン II

パターン II では、それぞれ A-B, B-C, A-C の共起度を計算し、結果値から動的に主題語、話題語を決定する。

A Judgment System of Suitability of Topics in Natural Language Conversation.

[†] Yuka KITAZAWA

[†] Noriyuki OKUMURA

Nagano National College of Technology, Department Electronics and Computer Science, noriyuki.okumura@ei.nagano-nct.ac.jp (†)

適合性判断手法 A

3.1 節で示した評価用データの作成条件に基づいて適合性の判断基準を設定した手法である。

適合性判断手法 B

適合性判断手法 A の、類似語を含む話題を不適合と判断する問題点を考慮した判断手法である。

A では、背景的知識内の主題語、話題語の出現非出現を重要視しているため、主題語と話題語に関連する単語が出現していない話題は不適合と判断した。しかし、不適合と判断された話題には、以下のような傾向が見られた。

- X:主題語が背景的知識内に出現、話題語が非出現かつ、背景的知識内の名詞に話題語の兄弟 (X') が含まれる。
- Y:主題語が背景的知識内に非出現、話題語が出現かつ、背景的知識内の名詞に主題語の兄弟 (Y') が含まれる。

X, Y に該当する話題における X', Y' は、主題語と話題語の類似語である傾向にあった。ここで、兄弟語はシソーラスにおいて、同じノードを親に持つリーフ同士を指す。

以上の問題を考慮し、適合性判断手法 B では兄弟語を使用した。単に兄弟語が存在すれば適合と判断すると、適合判断の許容範囲を広げるだけになる。X では話題語と X' の、Y では主題語と Y' の関連度計算を行い、関連度がしきい値以上だった場合、適合と判断した。

4.2 結果

パターン I で主題語と話題語を選出し、適合性判断手法 A を用いた判断手法を IA, 他も同様に, IB, IIA, IIB とする。表 1 に各手法の正答率を示す。

適合性判断手法	正答率 (%)
IA	68.9
IIA	66.6
IB	71.8
IIB	68.2

表 1: 各適合性判断手法における正答率表

表 2: 適合性判断手法 B における類似語の判定例

判定	主題語	話題語	A	B
成功	母親	団体	不適合	適合
失敗	母親	集まり	不適合	不適合

ただし、ここに示すデータはコラム [3] から抜粋した文章を使用している。

図 2 に、評価データの種類の別正答率を示す。

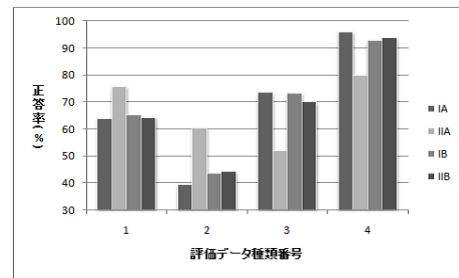


図 2: 各適合性判断手法における話題種類別正答率

5 考察

パターン II では主題語と話題語の選択を動的に行うことで、I に比べ正確な話題を選択できた。正しい話題を使用し、関連語を多く抽出できたことで、適合の許容範囲が広がったため、不適合のデータに対する正答率が低くなった結果、パターン I に比べ全体の正答率が低い。しかし、ユーザーの話題が 2 語で構成されるとは限らないため、適合性判断において主題語と話題語の動的決定は重要である。以上から、パターン II の精度を向上させることで、適切な適合性判断が行えると考える。

適合性判断手法 B では、表 2 のように、A では適正な判断ができなかった話題に対して正しい判断を行えた。また、表 1 から、IA の正答率 68.9% に比べ、IB は 71.8% と約 3% の差異が確認できた。適合性判断手法 B は、A に比べ適切な適合判断が行える手法であるといえる。

6 まとめ

本稿では、背景的知識に対する、話題の適合性判断手法について評価を行った。今後の課題として、人間の考える一段階連想、二段階連想の違いを明確にし、適切な連想範囲で適合判断を行える手法を確立する必要がある。

参考文献

[1] 「相手の嗜好にあった話題を提供する自動発話システムの開発」林 輝大：第 72 回情報処理学会全国大会 6X-7 2010 年 3 月

[2] 「主題語からの話題語自動抽出とこれに基づく Web 情報検索」野田 武史, 大島 裕明, 小山 聡, 田島 敬史, 田中 克己：電子情報学会, DE2006-90

[3] 「ない不便は本当に不便か」大平 一枝：
<http://www.asahi.com/housing/diary/TKY201011080114.html>