HTML構造を用いた文書の分析手法の提案

ニコニコ大百科の概要についての主語補完実験

江畑 拓哉（サイボウズ・ラボユース/ 筑波大学）

初音ミク　単語

(初音ミクは) ---

<タイトル>→概要
=? 主語

© Crypton Future Media, INC.

初音ミクの概要

初音ミクとは、――― である。(初音ミクは) ――― な ――― で、――― だ。(初音ミクの) 誕生日は8月31日 (非公式)。(初音ミクの) CV: ―――

⋮

<タイトル>→概要
=? 主語

楽曲

有名なものとして、(初音ミクの) (楽曲は) ――― 、――― 、――― などがある。

<タイトル>→<項目>
=? 修飾語+主語

楽曲一覧
- ―――
- ―――
- ―――

楽曲 → List
=? 楽曲の一覧

関連項目

- ―――
- ―――
- ―――

関連項目→ List
=? リレーションマップ

コラボレーションした作品や企業

<ニコニコ大百科/初音ミク>

Motivation

対話システム

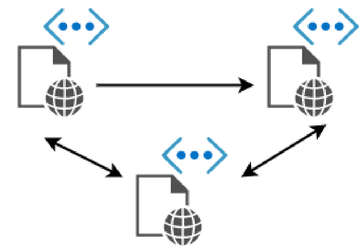
Knowledge Base

Web データ (大百科 etc.)

Fundamental

HTML 内の特定部分の値抽出 (株価など)
文書内の外部リンク同士のネットワーク

```
<!DOCTYPE html>
...
<h2>hoge hoge </h2>
<div id="main">
<strong id="value">VALUE</p>
</div>
<div>...</div>...
```



Proposed

HTMLの構造を読み
構造に基づいた適切な前処理を施す
⇒大百科などある程度規格のあるサイトに有効と予測

```
<h2>概要</h2>
<p> foo-hoge </p>
<p> bar </p>
```

⇒ 主語の欠けている文
⇒ (<単語名> +は) + 文

```
<h2>ステータス</h2>
<div><table><tr>
<td>key_1</td><td>value_1</td>
<td>key_2</td><td>value_2</td>
</tr></table></div>
```

⇒ <単語名>
↳ key_1 = value_1
↳ key_2 = value_2

事前準備・実験

事前準備

1. HTMLパーサの作成
⇒目的：HTML構造を残したまま解析難易度を低下させる
手法：Clojure(Lisp)を用いたDSLへの変換→JSON化
補充：観察により不要と思われるタグを削除・クリーニング

2. 句構造解析
⇒ 目的： 主語が欠けている文の特定
手法：StanfordNLPを用いた句構造解析

実験内容
データ　：ニコニコ大百科の単語記事全体から
ランダムサンプルした 50 記事
主目的　：主語補完ができているか
副目的　：文を適切に抽出できているか
実験概要：サンプルした記事の中の「概要」から
主語の欠けている文を取り出し補完
意味が補完できているかを評価
評価　　：3段階 (可・不可・不明)

結果

1.文を正しく抽出できているか

文を抽出	文数
できている	488 (83%)
できていない	55 (9%)
判断できない	46 (8%)

2.補完に関する実験(266文)
(上:2-1/下 2-2)

補完	/ 補完文	/ 抽出文
できている	67 %	30%
できていない	33%	15%

自然な文を	/ 補完文	/ 抽出文
保っている	57%	26%
保っていない	43%	19%

カイ二乗値:大きいほど関係している
p 値　　:小さいほど信頼できる
(特に 5e-2 より低いと良)

3.独立性検定(266文)

比較対象	カイ二乗値	p 値
1：2-1	96.2	1.25e-21
1 ： 2-2	100.4	1.57e-22
2-1 ： 2-2	168.9	1.27e-38

- 主語が欠けている文は、抽出された文に対して半分近くあった。
- 文の抽出精度は9割を超えることができなかった。
- ヒューリスティックかつ機械学習を用いない手法でも
7割近くの補完を確認できた。
- 意味を補完できても、必ずしも文の自然さを満たせるわけではない。
- 独立性検定から1, 2-1, 2-2 に、かなりの確かさで関連がある。

基本情報	値
記事の種類	単語
サンプル記事数	50
抽出された文	589
文/記事	平均11.78 / 分散11.99
主語が欠けている文	266 (45%)

考察・今後の課題

⇒考察
⇒今後の課題

⇒主語の欠けた文について
何らかの対策をしなければならない、
という問題設定は間違っていなかった

⇒「概要」のみに対象を絞ったことについて
他の部分についてはまだ実装・実験ができていない

⇒独立性検定の結果について
文抽出の精度を上げることで
他精度の向上が推測される

⇒7割近くの精度について
まだ実用には心許ない

⇒使った文の利用先について
(主語, 述語, 目的語) というようなタプルを作れるため
知識ベースの作成の作成補助になると考えられる

⇒サンプル記事数について
より多くの記事を使った実験を行いたい