二段階洗練化手法による
新聞記事からの人物説明記述の抽出

作成者:

横浜国立大学大学院　環境情報学府

西田　成臣　　森　辰則

発表者:

06T4056N　三沢　博章

はじめに

本実験の目的

・新聞記事から人物の説明記述のまとまりを抽出することを目的とする。

↓そのために、、、

- 解候補となる文章部分に対して、人物説明記述を含むか否かを機械学習により判定を行う。
- 人物説明文の区切りを判定する。

・系列ラベリングに基づく手法も提案し、比較する

本研究におけるタスク定義

・システムのデータ・・・毎日新聞の98年から01年までの4年分を対象

・抽出するデータ・・・人物の略歴、ただし、人物が係った出来事等は抽出しない

人物の説明記述は複数文に渡り連続して書かれる

開始文には説明対象の人名が含まれる

本研究におけるタスク定義(2)

１．説明記述を抽出する対象の人名をユーザから受け付け、その人名が登場する文が人物の説明記述の先頭になっているかを判定

２．１で説明記述の先頭であると判定された文を先頭として後続する文のどこまでが一連の説明であるかを判定

提案手法part1
系列ラベリングに基づく手法

人物説明文が複数文であるという考えのもと

→IOB2法を利用

IOB2法とは？

各文に以下のラベルを割り振る事で各文を表現する。

I:開始文以外の説明文

O：無関係の文

B：人物説明記述文の開始文

提案手法part1
系列ラベリングに基づく手法(2)

分類の対象となる前後、計３文により学習・判定をおこなう

↓つまり、、、

ラベリングを行った文の列において、

『B,I,I,…,I,O』

という系列を見つける。

BからOの直前までを説明記述として抽出する。

提案手法part2
二段階洗練化手法

一段階目の抽出法:

- 解候補の文章が人物に関する説明記述を含むか大まかに判定することを目的とする。
- 解候補は、人名を開始文に含む文章部分である。
- あらかじめ決められた数n文からなる。

提案手法part2
二段階洗練化手法(2)

二段階目の抽出法:

・人物説明記述の末尾となる文を見つけ出す

↓

二文を単位として、

- 二文とも人物説明文であるか
- 一文目が人物説明文であり、二文目が無関係の文であるか

以上の判定を行う。

評価実験

訓練事例・・・新聞記事の98年1月一ヶ月分

評価事例・・・同2月前半半月分

系列ラベリング手法におけるラベルの分類精度:

正解ラベル	判定ラベル	判定数
I	I	702
	O	51
	B	118
O	I	60
	O	16040
	B	65
B	I	7
	O	28
	B	243

	Precision	Recall	F1
I	91.3	80.1	85.3
O	99.5	99.2	99.4
B	66.0	87.4	75.2

評価実験(2)

一段階目の文章部分の分類精度:

Accuracy(%)	Precision(%)	Recall(%)	F1
98.9	92.5	91.6	92.1

以上の結果より、高い精度で人物説明記述文を含むかどうかの分類が可能であると確認できた。

二段階目における説明境界の分類精度:

Accuracy(%)	Precision(%)	Recall(%)	F1
95.9	96.9	97.7	97.3

以上の結果より、人物記述の境界の判定においても高い精度で可能であることが確認できた。

抽出手法の精度の比較

判定の基準:

・開始文から判定を行い、開始文とそれ以外の説明文が過不足なく抽出されたら完全抽出とする。

・説明文の途中で抽出が終わったり、反対に過剰に抽出した場合は部分抽出とする。

・無関係な部分の抽出は誤抽出とする。

		Precision(%)	Recall(%)	F1
系列ラベリング	部分	91.0	88.3	89.6
	完全	88.4	66.8	76.1
二段階洗練手法	部分	94.2	93.3	93.8
	完全	93.2	77.2	84.4

以上の結果より、系列ラベリングに基づく手法よりも、二段階洗練化手法のほうが精度が高いことが証明できた。

おわりに

本研究では、

・系列ラベリングに基づく手法

・二段階洗練化手法

→二段階洗練化手法を用いることで、F値の向上が見られたので、提案の有効性が確認できた。

今後は、

新聞記事からの抽出を拡張し、Web記事からの抽出が出来るよう拡張してゆくことを考えている。