結束性と場面の変化に着目した物語文要約システムの構築

Narrative Text Summarization Focusing on Cohesion and Scene Changest

松倉 慎一郎[1)]
指導教員　岩下 志乃[1)]

1）東京工科大学大学コンピュータサイエンス学部コンピュータサイエンス学科　岩下研究室
キーワード：物語要約，結束性，語の吸引力，自然言語処理

１．はじめに

近年､Web 小説投降サービスの広がりに合わせて、Web 上で公開される物語文書が増加している。しかし、Web 上で公開されている作品には、あらすじなどの要約文がないことが多く、ユーザーが新たに読みたい作品を探すことに時間がかかるという問題がある。そこで、簡単に作品の内容を把握するために要約文を自動的に作成することが重要であると考えられる。

物語文書の要約文を作成する既存研究としては、重要度を計算して文を抽出し、さらに文間を補完する文も抽出することで、合わせて要約文とするものがある[1]。また、文同士の結束性を利用して要約文を作成する手法を提案する研究がある[2]。本研究では、文同士の結束性と、場所の移動や時間の経過などの場面転換を利用して要約文を生成するシステムを作成することを目的とする。

２．研究概要

今回作成する要約文出力システムは、入力に対して自動的に要約文書を出力することを目指す。入力には物語文書全文を用いる。

入力された文書に対して形態素解析を行い、そこで得られた要素から場面が変化している文を検出し、その前後の文を要約文候補として抽出する。同様に、形態素解析で得られた品詞などの要素を素性として､SVM を用いて文同士の結束性の有無を判定し、結束性が無いと判定された文を要約文候補として抽出する。具体的な素性を表 1 に示す。

表 1：素性一覧

1	名詞、形容詞、動詞について一致する単語を利用した文同士の類似度
2	接続表現を累加、逆説、因果、並列、転換、例示、その他に分類したものについての文中での出現箇所を利用した数値
3	指示表現についてコソア系について分類したもの文中での出現箇所を利用した数値

最後に二つの手法を用いて得られた要約文候補を用いて要約文を生成、出力する。

３．場面の変化の検出について

場面の変化の検出について、どの要素を用いるかの手がかりとするため、赤ずきん[3]を用いて場面が変化した文に関して手作業でラベル付けを行い、その後 KH Coder[4]を用いて特徴語の抽出を行った。その結果、場面が変化した文の特徴語として、「オオカミ」、「赤ずきん」、「狩人」などの登場人物や、「家」、「森」などの場所を示す名詞が含まれるという特徴が見られた。

このことから登場人物について、語の吸引力[5]を用いて場面の変化を検知することを試みた。場面の変化の検知について、登場人物の単語の中で、その文における吸引力が最も高い語に着目し、前後の文のその語の吸引力と比較して、吸引力が大きかった場合に場面が変化したと判定する。

この方法で同様に場面の変化の検知を実施した。まず登場人物の単語に着目したところ，会話文について誤検出が多かったことから，入力文の会話部分の削除を行った。さらに表2の結果を踏まえ、

着目する語に場所に関する単語も加えた。

手作業で場面変化のラベル付けを行った結果と比較して精度を求めた。その結果を表 2 に示す。会話文を削除することにより、適合率の上昇が見られた。また、着目する単語に場所に関する単語を加えることによる精度の上昇は見られなかった。

表 2：吸引力による場面変化検出の精度

会話文	単語	正解率	適合率	再現率	F 値
有	人	0.732	0.258	0.444	0.327
無	人	0.661	0.364	0.444	0.400
	人 ＋ 場所	0.606	0.308	0.444	0.365

3．文章結束性について

文章結束性について、結束性を検知するため、文同士の類似度、主語の有無、指示表現をそれぞれ抽出して、素性とした。その素性を用いて「シンデレラ」「赤ずきん」について SVM にて分類を行った。分類の結果、誤った判定をしてしまった文の一例を表 3 に示す。

表 3：不正解となった文

正解ラベル	文
結束性あり	そして、うばは、いっそくの小さなガラスのくつをシンデレラにあたえました。せかいのどんなものよりかわいらしい、すてきなくつでした。
結束性なし	王子さまは、今日のシンデレラが、今までの中でいちばんうつくしい、と思いました。すうじつご、シンデレラと王子さまはけっこん式をあげました。

正解ラベルが結束性ありの文では、同じ「くつ」についての文であるが、共通の単語が「くつ」のみで、接続表現や、指示表現がない文を誤って結束性なしと判定している。正解ラベルが結束性なしの文では、時間が経過した後の文ではあるが、共通の名詞として、「シンデレラ」「王子」という登場人物がある文を結束性ありと誤って判定している。

「赤ずきん」と「シンデレラ」を用いて学習したモデルで新たに「白雪姫」について結束性の判定を行った。結果として「白雪姫」全 270 文中 157 文が、結束性が無いと判定された。ただし、その 157 文中の 69 文が会話文であった。そこで、会話文を削除して判定を行った所、結束性が無いと判定された文は 69 文にまで減少した。

5．おわりに

本研究では、時や場所の変化や文同士の結束性を利用して要約文を生成するシステムを作成することを目指す。

今後は、文同士の結束性、場面の変化の検出に関して精度の向上を目指す。また、抽出した文章を自然な日本語に直す処理についても作成する予定である。

参考文献

[1] 横野 光，“整合性を考慮した物語要約システムの構築”， 自然言語処理， 15 巻， 5 号，pp.47-71，2013.

[2] 山本 悠二，増山 繁，酒井 浩之，“小説自動要約のための隣接文間の結束性判定手法”，言語処理学会年次大会発表論文集， 12 巻，pp.1083-1086，2006.

[3] 赤ずきん | 青空文庫，https://www.aozora.gr.jp/cards/001091/files/59835_72466.html

[4] KH Coder | 使い方を知るためのチュートリアル，https://khcoder.net/ [22/08/03 最終アクセス]

[5] 赤石 美奈，“文書群に対する物語構造の動的分解・再構成フレームワーク”， 人工知能学会論文誌，21 巻，5 号，pp.428-438，2006.