歴史文献史実抽出システムの評価

2U－5　丸山　美和*　後藤晋一**　長谷川絵理**　西本秀樹**
（関西大学大学院　総合情報学研究科*　関西大学　総合情報学部**）

1　はじめに

「歴史文献からの動詞抽出システム」は、歴史研究を支援することを目的としたシステムである。今回構築した史実文抽出システムは、膨大な歴史文献から史実動詞を含んだ文（史実文）だけを抽出し、効率良くデータベース化するためのシステムである。現在の日本語全文検索システムの多くは、対象文書中に含まれる単語の情報のみを利用して検索を行っている。しかしその場合、対象の文献が地域、時代により限定されてしまうため、全ての歴史文献を対象とすることは困難である。そこで、動詞によって抽出できるシステムを構築することで、上記の条件に影響されないシステムを目指した。

今回は、構築した動詞キーワード検索による史実抽出システムの精度を評価し、システムを検証する。

2　史実文抽出の流れ

歴史文献から、必要な情報だけを取り出し、それらの情報をデータベース化する、史実文抽出システムの作業手順を次に示す。

手順：

1：文献をスキャナによって取り込み，ＯＣＲを利用し、一文テキスト型データに加工。

2：テキスト型データを動詞抽出システムで読み込み、史実のみを取り出すために、史実動詞が含まれる文（史実文）を抽出。

3：表示された史実文を専門家が判読し、属性値の切り出し。

4：切り出した属性値をレコード化し、データベースへ追加。

5：データベース化したレコードを、さらにＸＭＬタグをもとに保存。

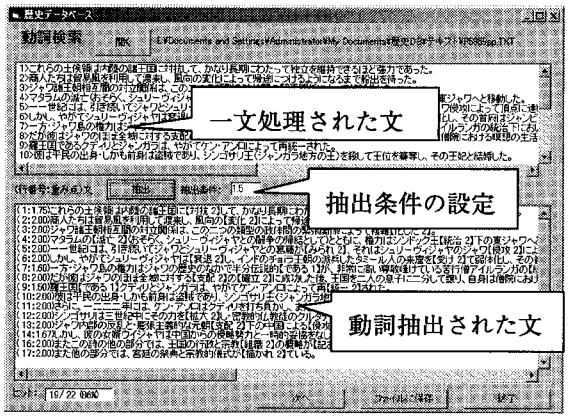

図1　動詞抽出システムの抽出フォーム

3　史実動詞データベース構築

動詞抽出システムは、史実動詞のみを格納した史実動詞データベースを構築し、その中の史実動詞が入った文だけを抽出するシステムである。史実動詞は、例えば「発見する」「支配する」「滅亡する」などの動詞をさし、その史実動詞が入っている文は、逆説的に史実文であると考えられる。


活用あり（過去／終止形）テーブル

ID	漢字(過去)	漢字(現在)	仮名(過去)	仮名(現在)
1	等っ	等う	うばっ	うばう
2	討っ	討つ	討っ	討つ
3	受け	受ける	受け	うける
4	現れ	現れる	あらわれ	あらわれる
5	行っ	行う	おこなっ	おこなう
6	受け入れ	受け入れる	うけいれ	うけいれる
7	襲っ	襲う	おそっ	おそう
8	入り乱れ	入り乱れる	いりみだれ	いりみだれる
9	送っ	送る	おくっ	おくる
10	衰え	衰える	おとろえ	おとろえる

レコード: 1 / 103

図2　史実動詞データベース

この史実動詞に、それぞれ１～３までの値（重み）をつける。この重みにより、自動的に全ての史実文が抽出されるシステムではなく、利用者自身が値と条件を設定することが可能となった。この重み付けにより、利用者の意思が反映されたシステムとなることを目指した。

非抽出規則データベース

歴史文献中に含まれる文として、①史実文　②著者の推論文　③著者による説文　④著者による感想文に分類される。この中で、史実データベースの中に格納されるデータは、①の史実文だけである。

②～④の著者の主観が入った文を可能な限り排除するため、主観的な動詞を非抽出動詞として、非抽出規則データベースを作る。この非抽出の動詞とは、「考えられる」「考察する」といった動詞である。

この非抽出のデータベースにもマイナスの重みをつけ、全ての合計により、史実文のみが抽出される。

4　評価

日本１０進分類法により分類された歴史文献を利用して、再現率と、適合率によって、このシステムの評価を行う。検索方法には、平均法とＭａｘ法があり、条件を１に設定し、比較を行った。また、どの分類の文献が、このシステムに最も合致するかを評価し、このシステムの検証を行う。

図３非抽出規則のデータベース

再現率：抽出された適合文（史実文）件数と対象とするテキスト内の全適合件数の比率

（抽出適合文件数　/　全適合文件数）＊100

適合率：抽出した文における適合件数と抽出結果の全件数の比率（精度）

（抽出適合文件数　/　抽出された全件数）＊100

再現率が高ければ、抽出されるべき史実文の漏れは少なくなり、適合率が高ければ、抽出された全文における史実文の割合が高くなる。

平均法：条件数値を設定し、文章中の史実動詞の重みを加算し、それを史実動詞の個数で割り、文章が条件数値以上の場合、その文章を抽出

Ｍａｘ法：条件数値を設定し、文中に出てくる史実動詞の重みが、その数値以上の場合のみ、文章を抽出

5　結果

条件１以上平均法・Ｍａｘ法で検証を行った。

	ヒット史実文	ヒット全文	史実文のみ
平均法	2316	2867	3029
Ｍａｘ法	2486	3113	3029

表１　方法別による抽出

	再現率	適合率
平均法	2316/3029*100 =76.5	2316/2867*100 =80.8
Ｍａｘ法	2486/3029*100 =82.2	2316/3113*100 =79.9

表２　抽出結果

その結果、適合率をあげるには、平均法が有効であり、再現率をあげるにはＭａｘ法が有効であることがわかった。また、Ｍａｘ法でも、条件を２にあげることで、再現率は落ちるが、適合率はあがることがわかっている。

[参考文献]

[１]César González Sainz,
Roberto Cacho Toca
Resent Research on Paleolithic Arts in Europe and the Multimedia Database
「第５回　公開シンポジウム　人文科学とデータベース　p11－ｐ22(1999)」

[２]丸山美和、歴史文献からの史実抽出と、XML化に関する研究「情報処理学会第 61 回全国大会」

Evaluation of the System for extracting the historical fact from literature

Maruyama Miwa, Goto Shinichi, Hasegawa Eri , Nishimoto Hideki
*: Graduate School of Informatics, Kansai University
**: Faculty of Informatics, Kansai University