단일문서요약을 위한 실마리어추출방법

김예화, 정만흥

실마리어는 문서분류와 질문응답, 문서요약 등의 체계들에서 중요하게 리용되며 실마리어를 추출하기 위한 많은 방법들이 제기되였다.

감독학습방법에 기초한 실마리어추출알고리듬[2, 3]에서는 단어출현빈도와 TF-IDF, *n*-그람, 단어길이, 위치정보, 공기빈도수와 같은 특징들을 효과적으로 리용하였다.

무감독학습방법[1]에서는 웨브페지의 순위결정을 위하여 HITS알고리듬, PageRank알고리듬을 리용하였다.

우리는 문서의 구문적 및 통계적특징들을 반영한 그라프토대의 학습자료를 구축하고 주어진 문서에 대한 실마리어를 추출하는 방법과 PageRank알고리듬을 리용하여 실마리어의 순위화를 진행하여 추출요약을 진행하는 방법을 제기하고 그 성능을 평가하였다.

1. 실마리어추출

우리는 문서의 구문적특징을 추출하기 위하여 문서를 그라프로 표현한다.

그라프에서 매 마디점에는 단어가 대응되며 마디점표식은 유일하다. 즉 단어가 본문에서 한번이상 나타난다고 해도 매 개별적단어들에 대하여 오직 하나의 정점만이 창조된다.

문서의 어느 한 문장에서 단어 a가 단어 b를 선행한다면 a에 대응하는 정점에서 b에 대응하는 정점으로 향하는 직접릉이 있게 된다.

문장의 종결구두점들(종지점, 물음표, 느낌표 등)이 2개 단어들사이에 존재하면 릉은 창조하지 않는다.

문서의 그라프로부터 매 마디점들의 특징 즉 그 마디점으로 들어오는 입력차수와 출력차수, 빈도수와 빈도단어분포 등을 추출할수 있다.

1) 감독학습에 의한 실마리어추출

감독학습에서 기본은 학습자료의 준비이다.

대량의 요약문을 구축하는데 많은 비용이 드는것과 관련하여 우리는 잡지기사들에 수동으로 부여한 실마리어들을 리용하였다.

실마리어와 뜻같은말관계, 뜻비슷한말관계에 있는 단어들도 실마리어로 보고 리용한다. 즉 이 실마리어를 포함한 문장들을 요약문으로 보고 학습자료를 구축한다.

모든 잡지기사문서들의 그라프들에서 매 마디점들에는 1 혹은 0을 표시한다. 즉 그 마디점의 단어가 실마리어이면 1, 그렇지 않으면 0을 준다. 이렇게 2개의 클라스표식을 리용한다.

또한 본문내용을 특징짓는 통계특징들과 그라프구조를 특징짓는 그라프에 기초한 특징(실례로 차수)들을 리용하여 문서의 구조적특징을 반영하려고 한다.

그라프의 매 마디에 부여하는 특징들은 다음과 같다.

① 입력차수: 마디로 들어오는 릉들의 개수

② 출력차수: 마디에서 나가는 릉들의 개수

③ 차수: 모든 릉들의 개수

④ 빈도수(TF): 마디로 표현된 단어의 빈도수로서 문서에 자주 출현하는 단어의 수를 문서의 총 단어수로 나눈 값을 취한다.

⑤ 빈도단어분포(분산된 마디빈도수): 0과 1의 값을 취한다. 빈도수가 턱값보다 크면 1을 취하고 작으면 0을 취한다. 우리는 턱값을 0.05로 설정한다.

⑥ 위치득점: 마디로 표현되는 단어 N을 포함하는 모든 문장들의 위치값에 대한 평균값을 취한다.

$$Score(N)=\frac{\sum_{S_i\in S(N)} Score(S_i)}{|S(N)|}$$

론문에서는 문장위치득점을 본문에서 문장위치의 거꿀수로 계산한다.

$$Score(S_i)=\frac{1}{i}$$

⑦ TF-IDF값: 마디로 표현되는 단어의 TF-IDF값을 계산한다.

$$\frac{tf}{tf+1}\log_2\frac{|D|}{df}$$

여기서 tf는 항목빈도수, $|D|$는 코퍼스에서의 문서의 총수, df는 항목이 나타나는 문서의 총수이다.

⑧ 표제득점: 문서표제가 마디로 표현되는 단어를 포함하면 1, 그렇지 않으면 0을 취한다.

우와 같은 특징들과 분류표식을 가진 문서들의 학습자료우에서 분류알고리듬을 적용하여 실마리어를 추출한다. 분류알고리듬으로 베이즈분류기를 리용한다.

분류과정은 다음과 같다.

첫째로, 문서를 그라프로 변환한다.

둘째로, 그라프의 매 마디로 표현되는 단어의 특징벡토르를 만들고 벡토르의 조건부확률을 계산한다.

셋째로, 분류기를 창조한다.

2) 무감독학습에 의한 실마리어추출

여기서는 추출된 실마리어들에 대한 순위화방법에 대하여 서술한다.

본문의 적합성순위를 결정하는 TextRank모형을 요약을 위한 가장 중요한 실마리어를 순위화하는데 리용한다.

TextRank모형은 단어그라프에 토대한다.

$$G=(V, E)$$

여기서 V는 마디점모임, E는 릉모임($V\times V$의 부분모임)이다.

w_{ij}: 임의의 두 마디점 v_i, v_j의 릉의 무게

$In(v_i)$: 마디점 v_i로 들어오는 점들의 모임

$Out(v_i)$: 마디점 v_i에서 나가는 점들의 모임

$TR(v_i)$: TextRank모형에 의해 얻어진 마디점 v_i의 득점

$$TR(v_i) = (1-d) + d \sum_{v_i \in In(v_j)} \frac{w_{ji}}{\sum_{v_k \in Out(v_j)} w_{jk}} TR(v_j)$$

d는 우연실마리어에 대한 사용자의 확실성을 나타내는데 0과 1사이의 값이다.

TextRank알고리듬에서 그라프의 매 마디점에 대한 득점계산은 그라프의 매개 점이 수렴할 때까지 그라프상에서 반복된다.

실마리어로서 가장 높은 득점값을 가지는 웃준위 m개의 실마리어를 선택한다.

2. 실마리어추출에 대한 실험결과

50개의 기사문서로 이루어진 문서모임과 실마리어들을 학습자료로 리용한다.

TextRank의 d값은 0.85로 설정한다.

우리가 사용한 평가척도로는 적중률, 완전률, F-척도이다.

TF-IDF방법과 제안한 방법을 비교하였다.

적중률, 완전률, F-척도는 다음과 같이 계산된다.

적중률: $P = \frac{|S_s \cap S_h|}{|S_s|}$

완전률: $R = \frac{|S_s \cap S_h|}{|S_h|}$

F-척도: $F = \frac{(\beta^2+1) \times P \times R}{\beta^2 \times P \times R}$

여기서 S_h는 정답요약문장모임이고 S_s는 체계가 출력한 추출요약문장모임이다. 그리고 β는 적중률과 완전률의 중요도를 조절하는 상수로서 $\beta = 2$로 설정하였다.

실마리어추출알고리듬의 성능비교결과는 표와 같다.

표. 실마리어추출알고리듬의 성능비교

실험방법	P	R	F
TF-IDF법	0.408	0.310	0.325 644
제안방법	0.520	0.358	0.381 788

표에서 보여주는바와 같이 제안한 방법이 전통적인 TextRank알고리듬에 비하여 우월하다는것을 알수 있다.

맺 는 말

문서요약을 위한 실마리어를 추출하기 위하여 문서의 그라프표현으로부터 학습자료의 특징들을 결정하고 감독학습에 의해 실마리어들을 추출하였으며 TextRank모형을 리용하여 추출된 실마리어들에 대한 순위화를 진행하고 그 성능을 평가하였다.

참 고 문 헌

[1] **김일성**종합대학학보 정보과학, 65, 4, 12, 주체108(2019).

[2] X. Tian; New Technology of Library and Information Service, 9, 30, 2013.

[3] T. Mikolov et al.; Advances in Neural Information Processing Systems, 26, 3111, 2013.

주체109(2020)년 11월 5일 원고접수

Keyword Extraction for Single-Document Summarization

Kim Ye Hwa, *Jong Man Hung*

In this paper, we proposed a keyword extraction method for single-document summarization. The text document was graphically represented and final keyword was selected by identifying keywords given by supervised learning and ranking of keywords identified by unsupervised learning.

Keywords: TextRank algorithm, word graph, keyword extraction