김일성종합대학학보
정보과학

주체110(2021)년
제67권 제1호

응집성에 기초한 조선어본문분할의 성능개선방법

조성영, 박련금

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《과학기술정보사업을 강화하여야 합니다. 과학기술정보사업을 잘하여야 적은 밑천과 품을 들여 과학기술발전에 절실히 요구되는 귀중한 자료들을 얻을수 있습니다.》(《**김정일**선집》 증보판 제15권 501페지)

본문분할(text segmentation)은 문서를 서로 다른 주제(topic)들을 가지는 여러개의 부분본문(토막)들로 분할하는 처리로서 정보검색과 정보추출, 본문요약, 질문응답 등에 널리 응용되고있다.

TextTiling본문분할방법은 류의어목록과 사전 등의 외부의 자원을 리용하지 않고 본문에서 단어들의 중복출현에 의해 분할을 진행하므로 계산량이 적으며 따라서 영어를 비롯한 많은 언어의 본문분할에 널리 적용되고있다. 그러나 본문에 출현하는 실지의 문장을 단위로 하지 않고 일정한 개수의 형태부들의 순서렬을 문장으로 간주하므로 분할적중률이 낮아질 가능성이 있다.[1－3]

론문에서는 본문의 응집성에 기초하여 TextTiling에 의한 조선어본문분할의 성능을 개선하기 위한 방법을 제안하였다.

1. 본문분할과 응집성

토막의 부분주제는 본문에서 론의되는 사건의 어느 한 부분으로서 여러개의 문장들로 구성된다.

응집성(Cohesion)은 본문의 표층구성요소인 문장들이 호상 련결되는 정도를 특징짓는다.

본문분할시 응집성관계에 있는 문장들은 같은 토막으로 분할되여야 한다.

본문의 응집성을 보장하는 요소들중에서 본문분할을 위한 전처리와 분할점의 선택에서 중요한것은 지시적맞물림과 생략, 대용, 결합이다.

① 지시적맞물림

지시적맞물림은 본문에서 이미 언급된 단어들을 반복하지 않고 그것을 가리키는 단어를 사용하는 언어적현상이다.

실례로《책의 임자는 영수였다. 그는 그것을 나에게 빌려주었다.》라는 문장들에서 뒤문장의 대명사《그》나《그것》은 각기 앞의 문장에서 쓰인《영수》와《책》을 거슬러 참조하기 위해 사용된 단어들이다.

본문자동분할시 대명사가 지시하는 대상을 고려하지 않으면 우의 두 문장에 중복되는 단어가 없으므로 련결의 정도가 약해져 서로 다른 토막으로 갈라질수 있다.

지시적맞물림을 나타내는 단어들을 AR로 표시할 때 대표적으로 대명사단어들이 속한다.

$$AR = \{그,\ \ 그것,\ 이,\ 이것\}$$

② 생략

생략은 본문에서 이미 언급된 대상을 다른 단어로 표현하지 않고 뒤의 문장에서 아예 없애버리는것이다.

문장에서 주어나 보어는 앞의 문장들로부터 쉽게 알수 있는 범위에서 생략되는 경우가 많다.

실례 1《책의 임자는 영수였다. 그가 (책을) 빌려주었다.》

형태단어에 접속하여 주어로 되게 하는 토들을 *TS*, 보어로 되게 하는 토들을 *TO*로 표시한다.

$$TS = \{\text{가, 이, 는, 은, 랑, 이랑}\}$$
$$TO = \{\text{을, 를}\}$$

③ 대용

대용은 본문의 앞부분에서 표현된 내용을 그대로 대신하는 말을 써서 문장을 앞의 문장에 련결시키는것이다.

실례 2《나도 그렇게 생각합니다.》

대용관계를 나타내는 단어 *SUB*는 다음과 같다.

$$SUB = \{\text{이렇게, 저렇게, 그렇게, …}\}$$

④ 결합

언어적인 접속수단을 리용하여 두 문장을 의미－론리적으로 련결시키는것이다.

실례로 조선어의 《왜냐하면》, 《때문에》와 같은 이음부사들은 앞문장의 내용과 뒤문장의 내용이 각각 《리유》와 《인과》관계로 긴밀하게 련결되여 본문으로 결속되게 하는 기능을 수행한다.

결합관계를 나타내는 단어 *COJ*는 다음과 같다.

$$COJ = \{\text{왜냐하면, 때문에, 따라서, 그러므로, 이리하여, …}\}$$

2. 응집성에 기초한 본문분할

본문분할은 전처리, 류사도계산, 분할점선택의 세단계를 통하여 진행된다.

① 조선어본문에 대한 전처리

문장들의 순서렬로서의 본문을 $T = S_1 S_2 \cdots S_N$이라고 하자.

N을 본문에 포함된 문장의 수라고 할 때

$$S_i = F_{i,1} F_{i,2} \cdots F_{i,L_i}$$
$$M(F_{i,j}) = M_{i,j}^1 M_{i,j}^2 \cdots M_{i,j}^{K_{i,j}}$$
$$MS_i = M(F_{i,1}) M(F_{i,2}) \cdots M(F_{i,L_i})$$

이다. 여기서 L_i는 i번째 문장의 형태단어수, $F_{i,j}$는 문장 S_i의 j번째 형태단어, $M(F_{i,j})$는 형태단어 $F_{i,j}$에 대한 형태부해석결과, $K_{i,j}$는 형태단어 $F_{i,j}$의 형태부수, $M_{i,j}^k$는 $F_{i,j}$의 k번째 형태부, MS_i는 S_i의 형태부해석결과이다.

실례로 2개의 문장으로 이루어진 본문《책의 임자는 영수였다. 그는 그것을 나에게 빌려주었다.》에 대한 형태부해석결과는 다음과 같다.

$M_{1,1}^1 =$책, $M_{1,1}^2 =$의, $M_{1,2}^1 =$임자, $M_{1,2}^2 =$는 $M_{1,3}^1 =$영수, $M_{1,3}^2 =$였, $M_{1,3}^3 =$다

$M_{2,1}^{1}=$그, $M_{2,1}^{2}=$는, $M_{2,2}^{1}=$그것, $M_{2,2}^{2}=$을, $M_{2,3}^{1}=$나, $M_{2,3}^{2}=$에게, $M_{2,4}^{1}=$빌려주, $M_{2,4}^{2}=$었, $M_{2,4}^{3}=$다

본문 T에 대하여 응집성을 고려한 전처리과정은 다음과 같다.

걸음 1 $i\leftarrow 0$

걸음 2 $i\leftarrow i+1$

걸음 3 $i>N$이면 모든 MS_i에 대한 금지어처리를 진행하고 처리를 끝낸다.

걸음 4 S_i의 모든 $F_{i,j}(j=\overline{1,\ L_i})$에 대한 형태부해석을 진행한다.

걸음 5 $i<2$이면 걸음 2로 이행한다.

걸음 6 모든 $M(F_{i,j})\ (1\le j\le L_i)$에 대하여 다음의 처리를 진행한다.

ㄱ) $A\leftarrow AR\bigcap\{M_{i,j}^{1},\ M_{i,j}^{2},\ \cdots,\ M_{i,j}^{K_{i,j}}\}$

ㄴ) $A\neq\phi$이면 ㅅ)에로 이행한다.(지시적맞물림판정)

ㄷ) $A\leftarrow SUB\bigcap\{M_{i,j}^{1},\ M_{i,j}^{2},\ \cdots,\ M_{i,j}^{K_{i,j}}\}$

ㄹ) $A\neq\phi$이면 ㅅ)에로 이행한다.(대용의 판정)

ㅁ) $A\leftarrow COJ\bigcap\{M_{i,j}^{1},\ M_{i,j}^{2},\ \cdots,\ M_{i,j}^{K_{i,j}}\}$

ㅂ) $A\neq\phi$이면 ㅅ)에로 이행한다.(결합의 판정)

ㅅ) MS_{i-1}의 뒤에 A의 모든 요소들을 추가한다.

걸음 7 MS_i에 포함된 용언형태부가 형용사이거나 자동사이고

$$MS_i\bigcap TS=\phi$$

이면 MS_{i-1}에서 TS의 요소를 찾고 그앞의 형태부를 MS_i의 뒤에 추가한 다음 걸음 2로 이행한다.(자동사의 주어생략)

걸음 8 $MS_i\bigcap TS=\phi$이면 MS_{i-1}에서 TS의 요소를 찾고 그앞의 형태부를 MS_i의 뒤에 추가한 다음 걸음 2로 이행한다.(타동사의 주어생략)

걸음 9 $MS_i\bigcap TO=\phi$이면 MS_{i-1}에서 TO의 요소를 찾고 그앞의 형태부를 MS_i의 뒤에 추가한 다음 걸음 2로 이행한다.(타동사의 보어생략)

걸음 3에서 금지어처리는 MS_i의 형태부들중에서 관형사, 부사(SUB, COJ의 요소들은 제외), 감동사, 토의 형태부들을 제거하는 방법으로 진행된다. 이와 함께 《명사+보조동사》의 형태부들을 합하여 하나의 형태부로 만드는 처리도 진행한다.

② 류사도계산

전체 MS_i들을 련결하여 형태부들의 순서렬을 $MT=w_1w_2\cdots w_L$이라고 하자.

본문분할에서는 분할되는 량쪽의 문장들이 서로 같은 정보량을 가지도록 하기 위하여 본래의 문장을 리용하지 않고 형태부순서렬 MT에서 SL개씩의 형태부들을 순서대로 취하여 1개의 문장으로 간주한다. 이때 두 문장사이의 경계점은 본문분할을 위한 분할점 후보로 된다. SL이 너무 짧으면 문장이 충분한 정보량을 포함하지 못하며 너무 길면 필요한 분할점들을 놓칠수 있다.

한편 일반적으로 하나의 주제(토막)는 여러개의 문장들로 구성되므로 이러한 문장들을 하나의 블로크들로 묶고 블로크들을 단위로 하여 류사도를 계산한다. 블로크의 길이 즉 블로크로 묶는 단어들의 수 BL은 보통 본문의 평균단락길이로 설정하거나 실험적으로 결정한다.

류사도계산은 경계점을 기준으로 각각 왼쪽과 오른쪽의 문장을 포함하는 두 블로크를 창문으로 하여 다음의 식에 따라 계산한다.

$$sim(B_L,\ B_R)=\frac{\sum_t w_{L,t}w_{R,t}}{\sum_t w_{L,t}^2\sum_t w_{R,t}^2}$$

여기서 t는 창문범위내에 출현하는 전체 단어, $w_{L,t}$, $w_{R,t}$는 각각 왼쪽블로크와 오른쪽블로크에서 t의 무게들이다. 무게는 블로크에서 단어의 출현회수로 정의한다. 두 블로크에 동시에 출현하는 단어가 많으면 블로크들사이의 류사도가 높게 된다.

이와 같은 방법으로 모든 경계점들에 대한 류사도를 구하여 류사도순서렬을 얻는다.

③ 분할점의 선택

분할점선택은 류사도순서렬에서의 류사도변화에 기초하여 진행한다.

먼저 류사도순서렬상에서의 작은 국부적변화를 억제하고 비교적 큰 류사도변화를 추출할수 있도록 하기 위하여 류사도순서렬에 대한 평활화를 진행한다. 매 경계점에 대하여 경계점의 류사도와 앞쪽으로 $W/2$개, 뒤쪽으로 $W/2$개 경계점들의 류사도들을 모두 합한 다음 평균값을 구한것을 평활화된 류사도값으로 한다. 여기서 W는 평활화창문의 길이로서 분할대상으로 하는 본문의 크기에 따라 정해진다. W값이 작으면 보다 많은 분할점들이 생성되게 된다.

다음 경계점순서렬에 기초하여 심도값순서렬을 얻는다. 매 경계점에 대하여 앞쪽으로 첫번째 국부최소점을 찾고 두 경계점의 류사도값의 차를 계산하여 d_L로, 뒤쪽으로 첫번째 국부최대점을 찾고 류사도값차를 계산하여 d_R로 하였을 때 $|d_L|+|d_R|$를 심도값으로 한다.

심도값은 경계점의 좌우부분토막의 차이의 정도를 표시하며 심도값이 클수록 토막들사이의 차이가 크고 경계점이 분할점으로 될 가능성이 높다.

경계점의 심도값이 $d=s-\sigma/2$보다 크면 분할점으로 확정한다. 여기서 s는 심도값순서렬의 평균값, σ는 표준편차이다.

d값이 커짐에 따라 분할점의 수는 작아진다.

분할점들을 찾은 다음 매 분할점에 대응하는 실제문장을 찾아 최종적인 분할위치들을 얻는다.

3. 실험 및 결과분석

제안된 방법에 기초한 본문분할에 대한 실험평가는 20개의 소론문본문들에 대하여 선행방법[1]과 적중률과 완전률을 비교하는 방식으로 진행하였다.(표)

$$\text{적중률}=\frac{\text{체계가 찾은 정확한 분할점의 수}}{\text{체계가 찾은 전체 분할점의 수}}$$

$$\text{완전률}=\frac{\text{체계가 찾은 정확한 분할점의 수}}{\text{수동적으로 결정한 전체 분할점의 수}}$$

실험결과 제안된 방법은 $SL=16$, $BL=4$, $W=5$일 때 성능이 가장 높으며 선행한 방법[1]보다 성능이 개선되였다.

표. 본문분할에 대한 분석

본문분할방법	완전률	적중률	*F*-척도
선행방법[1]	0.55	0.62	0.58
제안방법	0.56	0.65	0.60

맺 는 말

하나의 주제에 대응하는 부분본문의 응집성을 일정한 개수의 형태부들로 된 문장들에 반영하는 방법을 제기함으로써 조선어본문분할의 성능을 개선할수 있게 하였다.

참 고 문 헌

[1] 조흥일; 정보과학과 기술 1, 3, 주체109(2020).

[2] Aqil M. Azmi et al.; Information Processing and Management, 54, 903, 2018.

[3] Anja Habacha Chaibi et al.; Procedia Computer Science, 35, 437, 2014.

주체109(2020)년 11월 5일 원고접수

A Cohesion-Based Approach to Improve Performance of Korean Text Segmentation

Jo Song Yong, *Pak Ryon Gum*

In this paper, we proposed a method of improving performance of Korean text segmentation by applying text cohesion.

Keywords: information retrieval, text segmentation, text cohesion