단어간의 연관성을 고려한 어휘 체인 기반 자동 요약

송영인[U] 한경수 임해창
고려대학교 컴퓨터학과
{sprabbit, kshan, rim}@nlp.korea.ac.kr

Automatic Summarization based on Lexical Chains considering Word Assocication

Young-In Song[U] Kyoung-Soo Han Hae-Chang Rim
Dept. of Computer Science & Engineering, Korea University

요 약

자동 문서 요약 분야에서 대상 문서를 컴퓨터가 이해할 수 있는 형태로 어떻게 파악하고 구조화 할 것인가는 중요한 이슈가 되어 왔다. 문서에 출현한 단어들은 Bag of Words 가정처럼 서로 독립적으로 존재하는 것이 아니라 문서가 쓰여진 의도에 따라 서로 간의 의미적, 혹은 지시적으로 연관되어 있다. 이러한 단어간의 연관성은 결속성(cohesion)이라고 표현하며, 이를 이용한 자동 요약 방법으로 Barzilay의 어휘 체인(lexical chain)을 사용한 자동 요약 방법이 대표적이다. 본 연구에서는 단어간의 연관성과 영문 시소러스인 워드넷(wordnet)에서 단어의 위치 정보를 사용하여 어휘 체인의 성능을 개선하였고, 요약 대상 문서의 개념을 어휘 체인에 기반해 표현하여 자동 요약의 성능을 개선하는 방안을 제시한다.

1. 서론

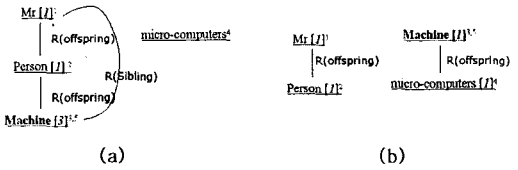
정보 검색, 문서 분류와 같이 문서 집합을 다루는 분야와는 달리, 자동 요약에서는 한 문서에서 핵심적인 내용을 추려내는 것을 목표로 한다. 이러한 특성 때문에 자동 요약의 성능 개선을 위해서는 대상 문서에서 문서 주제에 관련된 더 많은 양질의 정보가 필요하게 된다. 이런 측면에서 본다면, 문서 주제에 따라 형성되는 단어나 문장간의 연관성과 같은 텍스트 고유의 자질을 요약에서 사용하는 것은 상당한 이점이 있다.

단어나 문장 같은 텍스트 구성요소간의 연관성을 텍스트 언어학에서는 일관성(Coherence)과 결속성(Cohesion)으로 구분하여 정의하고 있다. 결속성은 텍스트의 표층구조의 연결성 혹은 언어학적 의미의 연결성에 대한 개념을, 일관성은 심층구조의 연결성 혹은 수사학적 내용의 논리성에 대한 개념을 의미한다.[1]

일관성은 주로 문장 간의 논리적 관계에 초점을 맞추고 있는 반면, 결속성은 텍스트의 구성요소 중 단어나 구 등의 하위 요소 사이의 관계에 관련되어 있으므로 자연어처리에서는 상대적으로 적용하기 용이한 개념으로 볼 수 있다.

자연어 처리의 여러 다양한 분야에서 텍스트의 결속성을 사용한 연구가 시도해왔으며, 그 중 요약에 결속성 개념을 적용한 연구로 Barzilay가 요약에 사용한 어휘 체인(lexical chain)을 들 수 있다[2].어휘 체인은 텍스트에서 문법적인 결속 장치를 제외한 어휘적 결집(lexical cohesion)만을 사용하여 텍스트를 분석하고 구조화한다. Barzilay의 어휘 체인은 텍스트에서 나타나는 어휘 사이의 의미적 관계(동의어, 반의어, 상/하위어등의 출현)에 초점을 맞추고 있으며, 이를 위한 리소스로 워드넷(Wordnet)을 사용하여 단어사이의 관계를 정의한다. [예1]은 [그림1-b]와 같이 사슬형태로 표현할 수 있다.

Mr[1] Kenny is the person[2] that invented an anaestheitc machine[3] witch uses micro-computers[4] to control the rate at which an anaesthetic is pumped in the blood. Such machines[5] are nothing new. [예1]



(a) (b)

[그림1] [예1]에서 생성가능한 어휘 체인의 두가지 예

[예1]의 핵심적인 내용은 [발명자]와 [발명된 기계]에 대한 개념이다. 각 개념에 대해 상이한 형태의 단어들이 사용되고 있지만, 단어들의 의미적 연관 관계를 사용하여 [그림 1-(b)]처럼 사슬형태로 표현할 수 있음을 볼 수 있다. 저자가 어떤 주제를 표현하기 위해 의미적으로 연관된 단어를 선택하여 사용하기 때문이다.

본 연구에서는 이 같은 직관에 바탕을 두고 새로운 어휘 체인 생성 방법을 제안하며, 생성된 체인이 요약 성능 개선에 도움을 줄 수 있음을 밝히려 한다. 본 연구의 초점은 다음과 같다.

문서에서 어휘체인을 생성하는 작업, 즉 단어와 단어간의 의미적 연관성을 밝혀내는 작업은 원칙적으로 대상이 되는 개별 단어들의 의미에 대해 고려해야만 한다.

[예1]에서 Machine의 의미가 [기계적으로 일하는 사람, 기계적 인간]으로 사용되었다면 [Person, Machine] 사이의 의미적 연관성이 성립하며 [그림1-a]의 어휘체인이 형성된다. Machine이 [기계, 도구]의 의미로 사용되었다면 [Person, Machine] 사이에는 의미적 연관성이 존재하지 않으며, 대신 [Machine, Micro-computer] 사이에서 의미적 연관성이 존재하게 된다 [그림1-b]. 즉 어휘체인을 생성하는 과정에서 단어의 다의적 특성 때문에 필연적으로 중의성이 발생하게 되며, 어휘 체인의 중의성 해소 과정이 필요하게 된다.

Barzilay의 경우, 단어와 단어사이에 성립 가능한 의미 관계에 대해 [표1]과 같이 경험적 가중치를 부여했으며, 이를 사용하여 어휘 체인의 중의성을 해소했다.

[표1] Barzilay의 의미 관계에 따른 가중치

의미 관계	가중치
동의어관계 (Synonym)	10
상하위관계 (Hyper/Hyponym)	8
반의어 관계 (Antonym)	7
전체/부분어 관계 (mero/holonym)	4
인접어 관계 (Sibling)	2

이러한 어휘체인의 중의성 해소 방법은 Barzilay 이후 많은 어휘 체인 생성방법에 대부분 적용되어 왔으나 이러한 방법은 지나친 단순화로 인한 오류를 포함할 수 밖에 없다. [2,4,5]

모든 후보 단어에 대한 단어 의미 결정은 현재의 기술 수준으로 불가능하지만 원시 말뭉치등을 통해 개별 후보 단어 혹은 단어간에 획득할 수 있는 공기 정보, 워드넷에서의 단어의 위치 정보등 여러 유용한 정보를 사용한다면 보다 정확한 어휘 체인을 획득할 수 있다. 본 논문의 2장에서는 이와 같은 체인을 이루는 개별 단어에 대한 정보를 활용하여 보다 정확한 어휘 체인을 생성하는 방법에 대해 언급한다.

다음 3장에서는 기존 Barzilay의 자동 문장 추출 방법의 단점과 이를 개선한 새로운 자동 문장 추출 알고리즘에 대해 제안하며, 4장에서는 어휘 체인 생성 결과와 문장 추출 결과에 대한 평가 및 결과 분석에 대해 다룬다.

2. 어휘 체인 생성

단어와 단어사이의 의미적 관계는 각각의 단어 의미에 종속적이기 때문에, 모든 단어의 의미를 결정할 수 있다면 어휘 체인의 중의성 해소 작업을 별도로 행할 필요가 없다. 하지만 문서에 나타난 모든 후보 단어들에 대해 단어 의미 결정을 해주는 것은 현재 자연어처리 수준으로는 어려운 일이므로 다른 문제 해결 방식이 필요하게 된다.

단어와 단어 사이에 의미적 관계를 찾아내는 문제는 두 단어 w_1, w_2가 주어졌을 때, 두 단어 사이에서 성립 가능한 의미적 관계 중 올바른 관계 r을 찾는 작업으로 정의할 수 있으며 이에 대한 불리언 함수 f_b를 사용해 $f_b(r(w_1,w_2))$ = true인 r을 찾는 일로 생각할 수 있다. 하지만 명확한 불리언 값을 가지는 함수 f_b를 찾아내는 것은 현실적으로 어려우므로, 결과값이 높을수록 $f_b(r(w_1,w_2))$이 참이 될 확률이 증가하는 스코어 함수 $f_s(r(w_1,w_2))$를 찾아내는 것으로 단어와 단어 사이에 의미적 관계 중의성 해소 문제를 정의한다.

이와 같은 함수 f_s를 찾아내기 위해, 본 연구에서는 소량의 의미 부착된 말뭉치(Semcor 1.6)와 대량의 원시 말뭉치(TREC Data 중 AP와 WSJ문서 중 10만건)를 사용하여 획득한 공기 정보와 영어 시소러스인 워드넷에 근거한 단어의 의미정보를 이용한다.

최종 유도되어 사용되는 스코어 함수는 (단어간 연관성 스코어) × (단어 사이의 의미 관계 가중치) 의

형식을 갖는다.

2.1 어휘 공기 정보를 이용한 단어간 연관성

두 단어가 원시 말뭉치에서 공기하는 빈도가 높을수록 두 단어 사이에서는 의미적 연관성이 발생할 확률이 일반적으로 증가하며, 원시 말뭉치에서 추출한 단어의 공기정보는 체인의 중의성을 해소하는데 중요한 실마리를 제공할 수 있다. 예컨대, Machine은 Person과 대량의 원시 말뭉치에서 42회 공기하는 반면 Computer와는 192회 공기한다. 이는 Machine이 Computer와 의미적 연관 관계가 성립할 기회가 Person보다 많음을 암시한다.

이런 가정하에 제안하는 중의성 해소 알고리즘은 다음과 같은 단어간 연관성 스코어를 사용한다.

$$Assoc_1(w_1, w_2) = \log(p(w_1, w_2)+1) = \log(\frac{count(w_1, w_2)}{segmentsize}+1)$$

w_1, w_2 는 워드넷에서 의미적 관계가 성립할 수 있는 후보 단어이며, $Segmentsize$는 대량의 원시 말뭉치에서 공기 정보를 획득하기 위해 분할된 텍스트 조각(Segment)의 수를 지칭한다. $Count(w_1, w_2)$는 두단어가 함께 출현한 조각 수를 뜻한다.

$p(w_1,w_2)$의 확률을 바로 사용하면 말뭉치에서 낮은 빈도로 출현한 단어와 높은 빈도로 출현한 단어사이의 가중치가 과도한 차이를 보인다. 이런 현상을 막기 위해 log를 취한 값을 사용한다.

정의된 $Assoc_1$은 의미 태깅되지 않은 원시 말뭉치에서 의미에 대한 고려 없이 추출하였기 때문에 자주 사용되는 다의어에 대해 편향된 스코어를 부여하는 경향이 있으므로 다음과 같이 정규화(normalization)하여 값을 보정한다.

$$Assoc_2(w_1, w_2) = \frac{Assoc_1(w_1, w_2)}{N_s(w_1) \times N_s(w_2)}$$

여기서, $N_s(w)$는 단어 w가 가질 수 있는 의미의 가지수를 뜻한다.

단어의 의미별 출현 빈도를 알 수 있다면 더 정확한 스코어 부여가 가능하지만, 의미 태깅된 말뭉치의 크기가 모든 단어의 의미별 출현 빈도를 계산하기에는 충분치 않으므로 간단하게 $Assoc_2$처럼 추정하여 사용한다.

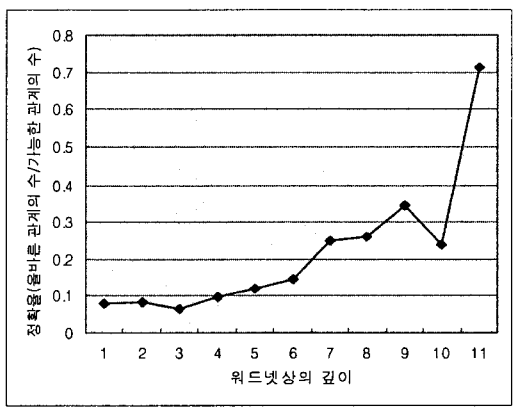
2.2 워드넷에서의 위치정보를 이용한 단어간 연관성

워드넷은 그 구조가 트리형태로 되어 있다는 특징을 가지고 있다. 실제로 문서에서 보다 의미있는 단어는 상위 동의어 집합(Synset)에 소속되는 상위 개념의 단어이기보다는 어느정도 구체적인 의미를 지니기 시작하는 중/하위 개념어이며, 문서의 구체적인 주제와 연관되어 사용되는 단어는 더 구체적인 의미를 지니는 하위어인 경우가 많다.

[표2], [그림2]는 의미 부착 말뭉치인 Semcor 1.6 문서 집합에서 선택한 임의의 10여개 문서에서 출현한 단어 w에 대해, 가능한 의미관계와 올바른 의미관계의 비율(정확율)이 w의 워드넷 상의 깊이(개념 루트와의 거리)와 어떻게 연관되는지를 보여준다. 이를 보면 하위 개념어로 갈수록 관계의 정확율이 증가하는 추세를 보임을 알 수 있다. 이런 경향은 문서가 모호한 상위어를 사용하여 주제를 설명하기 보다는 보다 구체적인 하위어들을 사용하여 주제를 전개해 나가기 때문으로 보인다.

[표2] 단어 w의 워드넷 상의 깊이(개념 루트와의 거리)에 따른 의미 관계 발생

단어 w의 깊이	단어 출현 빈도	가능한 의미 관계 수	올바른 의미 관계 수	정확율
0	2232	1584	59	0.0372
1	339	226	18	0.0796
2	1040	868	71	0.0817
3	4570	2711	176	0.0649
4	6853	5761	566	0.0982
5	4311	3234	390	0.120
6	2240	1505	234	0.143
7	2261	550	378	0.251
8	751	297	143	0.26
9	460	42	102	0.343
10	57	35	10	0.238
11	42	35	25	0.714



[그림2] 깊이와 의미 관계 정확율

이같은 경향을 반영하여, 연관성 스코어 함수가 [그림2]에 근사하도록 다음과 같이 변형한다.

$$Assoc_3(w_1, w_2) = Assoc_2(w_1, w_2) \times (2 + Depth(w_1))^2 \times (2 + Depth(w_2))^2$$

$Depth(w)$ 는 단어 w의 개념 루트로부터의 거리, 즉 워드넷 상의 단어의 깊이를 뜻하며, $(2+Depth(w))^2$는 [그림2]와 유사한 곡선을 생성하기 위해 선택된 함수이다. 단어 w의 워드넷 상의 깊이와 의미 관계의 정확도는 이 함수를 따른다고 가정한다.

2.3 의미 관계 가중치

단어와 단어사이의 의미 관계가 어떤 종류인지에 따라서도 가능한 의미 관계와 올바른 의미 관계의 비율, 즉 의미 관계의 정확율이 [표3]에서처럼 차이를 보이는 현상을 관찰할 수 있다.

[표3] Semcor 1.6 의미 부착 말뭉치에서 임의로 선택한 90문서에서의 의미 관계별 정확율

의미 관계	정확율
동의어관계 (Synonym)	0.19
상하위관계 (Hyper/Hyponym)	0.17
반의어 관계 (Antonym)	0.36
전체/부분어 관계 (mero/holonym)	0.40
인접어 관계 (Sibling)	0.08

[표3]은 Barzilay의 [표1]과 다소 상이한 결과이다. 반의어나, 전체/부분어 관계는 낮은 빈도로 발생하며 높은 빈도로 발생하는 상하위, 인접어 관계보다 높은 정확율을 보인다. 이는 워드넷의 의미 관계 분포의 특성 때문인데, 반의어, 전체/부분어는 다른 고빈도 의미 관계에 비해 단어당 의미 관계가 1/10 미만의 빈도를 가진다는 사실에 기인한다. 즉 한 단어가 가질 수 있는 반의어, 전체/부분어 관계는 상당히 한정되어있다.

[표3]의 정확률값으로 의미관계 가중치 weight(r(w_1,w_2))을 정의한다.

단어 w_1, w_2 에서 의미 관계 r 에 대한 스코어 함수 $f_s(r(w_1,w_2))$ 는 최종적으로 다음과 같이 정의된다.

$$f_s(r(w_1, w_2)) = Assoc_3(w_1, w_2) \times weight(r(w_1, w_2))$$

2.4 어휘 체인 스코어

어휘 체인 스코어는 어휘 체인C_i를 이루는 의미 관계 r_j와 각각의 의미 관계를 구성하는 두 단어 w_{j1}, w_{j2}에 대해 다음과 같이 부여된다.

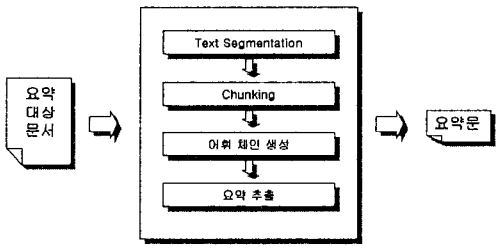
$$Score(C_i) = \sum_{r_j \in C_i} f_s(r_j(w_{j1}, w_{j2}))$$

어휘 체인의 중의성이 발견되었을 때, 생성 가능한 경우에 대해 어휘체인 스코어를 계산하여 가장 높은 점수를 얻는 체인을 올바른 것으로 선택한다.

3. 어휘 체인을 사용한 자동 요약

3.1. 어휘 체인을 사용한 자동 문서요약시스템 개요

본 논문의 자동 문서 요약 시스템의 흐름은 [그림3]과 같으며, 기존의 어휘 체인을 사용한 자동 요약 시스템과 동일한 흐름을 거친다.



[그림3] 어휘 체인을 사용한 자동 요약 시스템의 구조

문서를 주제 단위로 분할하여 처리하기 위해 Text Segmentation을 사용하며, 기존 Barzilay의 요약 시스템과 동일하게 Hearst의 text tiling 방법을 사용했다[2,3]. 또한 복합 명사를 처리하기 위해 분할된 문서 Segment[6]를 Chunking한 후, 어휘 체인을 생성하여 중요 문장을 추출한다.

3.2. 중요 문장 추출

기존 Barzilay의 어휘 체인을 사용한 중요 문장 추출 시스템은 요약문의 길이 조절이 불가능하고 문장 위치나 길이, 실마리 단어등 통계기반 접근방법에서 유용함이 입증된 자질들을 사용할 수 없다는 단점이 있다[2]. 또한 문장의 중요도 랭킹이 불가능한 접근방법을 사용하기 때문에 다른 모델과의 혼용이 어렵다.

하나의 어휘 체인을 이루는 단어들은 문서에서 표현하고자 하는 하나의 개념을 상이한 형태로 표현하고 있다

고 가정한다면, 어휘 체인의 구성 단어들은 이형동의어, 혹은 유사 개념어가 된다. 어휘 체인을 사용하면 단어의 동일한 형태만을 반영하는 문서의 단어빈도를 문서의 중요 개념을 더 잘 반영하도록 조정할 수 있을 것이다.

제안하는 중요 문장 추출 알고리즘은 어휘 체인을 사용하여 문서의 여러 개념을 찾아낸 후, 각 단어 빈도를 개념에 대한 빈도로 재평가한다. 그 후 조정된 빈도를 사용해 문장을 중요도에 따라 랭킹하여 요구되는 요약율로 문장을 추출한다.

단어의 빈도를 재평가하여 문장을 추출하는 방법은 다음과 같다.

1) 요약 대상 문서에서 주제 단어 집합 *1* 를 생성한다. 주제 단어는 문서에서 다음 *Threshold* 이상의 빈도로 출현하는 단어로 정의한다.

$$Threshold = avg(tf(w_i,d)) + 2 \times stddev(tf(w_i,d))$$

여기서 $tf(w_i,d)$는 단어 wi가 문서 d에서 출현한 횟수를 의미한다.

2) 다음의 $Freq_{Seg}$를 사용하여 문서의 각 세그먼트에서의 단어 빈도를 재평가한다.

$$Freq_{seg}(w_i, s_j) = Freq_{doc}(w_i, d) + Freq_{chain}(w_i, c_{jk})$$

$$Freq_{doc}(w_i, d) = \begin{cases} tf(w_i, d), & w_i \in T \\ 0, & otherwise \end{cases}$$

$$Freq_{chain}(w_i, c_{jk}) = \begin{cases} tf(w_i, s_j) + \alpha \times weight_{chain}(C_{jk}) \\ \quad w_i \in c_{jk},\ w_i \in T \\ tf(w_i, s_j) + \beta \times weight_{chain}(C_{jk}) \\ \quad w_i \in c_{jk},\ w_i \in T \\ otherwise\ 0 \end{cases}$$

$$weight_{chain}(c_{jk}) = \sum_{w_i \in c_{jk}} tf(w_i, s_j)$$

w_i : 문서에서 출현한 i번째 단어

s_j : 문서 d의 j번째 Segment

c_{jk} : 문서 d, j번째 Segment의 k 번째 어휘 체인

1 : 주제 단어 집합

$tf(w, d)$: 단어 w가 문서 d에서 나타난 빈도

$tf(w, s_j)$: 단어 w가 세그먼트 s_j에서 나타난 빈도

α, β : 주제 단어 포함 여부에 따른 $Weight_{Chain}$ 의 반영정도를 결정하는 상수값. 실험적으로 결정한다.

3) 요약 대상 문서의 각 문장에 대해, 문장을 구성하는 각 단어의 재평가된 빈도의 합을 다음과 같이 스코어로 할당한 후, 요구되는 요약률에 따라 상위 문장을 추출한다.

n개의 단어 $w_1..w_n$로 구성된 j번째 *Segment* s_j에 속한 문장$Sentence_{ji}$이 존재할 때,

$$Score_{sen}(Sentence_{ji}) = \sum_{i=1}^{n} Freq_{seg}(w_i, s_j)$$

4. 실험 및 평가

4.1. 어휘 체인 정확성 평가

이상적인 어휘체인은 어휘 체인을 이루는 각 단어의 의미와 구성 단어들간의 의미적 관계가 원문의 것들과 일치해야하며, 원문에서 발생하는 모든 단어간 의미적 관계를 모두 반영해야한다. 이러한 관점에서 어휘 체인을 위한 평가 척도로 어휘 체인 생성 과정에서 결정된 단어 의미와 의미 관계의 정확률과 재현율을 사용한다.

단어 의미와 의미 관계의 정확률과 재현율을 측정하기 위해 의미 부착 말뭉치인 Semcor 1.6 문서 집합의 A, B ,C, D 섹션 16문서를 실험집합으로 사용했으며, 어휘 체인 생성을 위한 리소스로 워드넷 1.6을 사용했다.

비교 대상은 이 논문에서 제안하는 어휘 체인 중의성 해소 방법을 사용한 시스템과 [Barzilay 97]의 방법을 사용한 시스템이다.

제안하는 시스템이 단어 의미 부착 정확율/재현율, 의미 관계 정확율/재현율에서 기존 시스템보다 모두 개선이 이루어졌음을 [표4]과 [표5]에서 확인할 수 있다.

[표4] 기존 방법을 이용한 시스템의 체인 생성 평가

	단어 의미 부착	의미 관계
정확률	0.4138	0.3526
재현율	0.1667	0.3479
F-measure	0.2376	0.3503

[표5] 제안하는 시스템의 체인 생성 평가 [1]

	단어 의미 부착	의미 관계
정확률	0.4943	0.4483
재현율	0.1884	0.4229
F-measure	0.2728	0.4352

[1] [표5]의 제안하는 시스템에서는 과생성을 막기 위해 $f_s(r(w_1,w_2))$의 스코어가 일정 이하일 경우는 고려대상에서 제외했다.

4.2. 문장 추출 시스템 평가

문장 추출 시스템의 평가 실험에서는 Daniel Marcu의 Ziff-davis 요약 실험 문서 집합 중 임의로 선택한 50건을 사용하였으며, 문서길이의 40% 만큼을 요약문으로 추출하였다. 실험에 사용한 Daniel Marcu의 Ziff-davis 요약 실험 문서 집합은 특정 요약 비율이 정해져 있지 않고, 문서마다 가변적인 수의 중요 문장이 태깅되어 있는 특성이 있다. 그래서 실험에 사용한 50개 문서의 평균 중요 문장 수 / 문서 길이를 요약 비율로 사용하였다.

평가 방법으로는 50개 문서의 정확률/재현율을 평균한 평균 정확율과 평균 재현율, F-measure를 사용하였다.

실험 비교 대상은 제안 하는 요약 시스템(LC)과 단어 빈도, 제목, 문장 길이, 문장 위치, 실마리 단어를 자질로 사용한 Naive-bayes classifier 통계 기반 요약 시스템(NB), 문서의 앞부분 40%를 요약으로 제시하는 시스템(HEAD)이다. 기존 [Barzilay 97] 시스템은 요약문 길이 조절이 불가능한 문제로 인해 비교평가가 어려워 실험 대상에서 제외했다.요약 시스템(LC)는 상수 α를 0.5, β를 0.3으로 사용하였다.

[표6] 요약 성능 평가

	평균 정확율	평균 재현율	F-measure
LC	0.52	0.47	0.50
NB	0.53	0.38	0.47
HEAD	0.50	0.45	0.44

제안하는 시스템은 분류기를 사용한 통계 기반 요약 시스템과 비슷한 평균 정확율과 23%의 평균 재현율 향상도를 보이며, 전체적으로 다른 시스템에 비해 좋은 성능을 보여주고 있다.

5. 결론 및 향후 연구

본 논문에서는 단어간의 연관성을 고려한 어휘 체인의 중의성 해소 방법과, 기존 어휘 체인을 이용한 요약 기법의 단점을 개선한 요약 방법을 제안하였다.

어휘 체인 중의성 해소를 위해 원시 말뭉치에서의 공기정보와 워드넷 상의 단어 위치 정보를 기존의 의미 관계 가중치와 병행사용하였고, 실험 결과 개선된 성능을 보였다.

기존의 [Barzilay 97]의 요약시스템이 가지는 문제점을 개선하기 위해 어휘 체인을 사용하여 문서에 출현한 각 단어의 빈도를 문서의 중요 개념에 충실하도록 조정해주는 방법을 사용한 어휘 체인 기반 요약 시스템을 제안하였다.

실험 결과, 다른 요약 시스템에 비해 큰 폭의 성능 향상은 보이지 않았으나, 3개의 비교 대상 중 안정적인 성능을 보였다.

또 제안한 요약 시스템은 재평가된 단어의 문서 출현 빈도를 사용한다는 면에서 단어의 문서 출현 빈도를 자질로 사용하는 기존의 분류 기반 통계적 요약 모델이나, 학습 기반 요약 모델과 혼용 사용하기 용이하다는 장점이 있다.

반면 제안하는 요약 시스템은 문서 주제에 관련된 단어들과 그렇지 않은 주변 단어들간에 차별성 부여가 기대보다 좋지 않았다. 따라서 향후 연구로는 체인의 정확성에 대한 성능 개선과 보다 정교한 단어 선택 및 빈도 조정이 가능한 요약 모델에 대한 연구가 필요하다.

6. 참고 문헌

[1] 김용도, "텍스트 결속이론", 부산외국어대학교 출판부, 1996

[2] Barzilay, Regina and Michael Elhadad, "Lexical Chains for Text Summarization", Master's thesis, Ben-Gurion University, 1997

[3] Hearst, Marti A, "Multi-paragraph segmentation of expository text" In Procdeeings of the 32nd Annual Meeting of the ACL, pp9-16, 1994

[4] Siber, H.G and McCoy, K.F 2000. "Efficient Text Summariztion using Lexical Chains". In Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000), 2000

[5] Meru Brunn, Yllias Chali, christopher j. Pinchak, "Text Summarization Using Lexical Chains", In Proceedings of the Document Understanding Conference (DUC-2001), 2001

[6] Young-Sook Hwang, So-Young Park, Hoo-Jung Chung, Yong-Jae Kwak and Hae-Chang Rim, "Shallow Parsing By Weighted Probabilistic Sum" In Proceedings of the 2001 International Conference on Computer Processing of Oriental Languages, pp. 236-241, 2001.