다중문서 요약에서 적응 기법을 이용한 문장 추출

임정민[O] 강인수 배재학[+] 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터,
[+]울산대학교 컴퓨터정보통신공학부
{beuuettO, dbaisk, jhlee}@postech.ac.kr, [+]jhjbae@ulsan.ac.kr

Sentence Extraction Using Adapting Method in Multi-Document Summarization

Jung-Min Lim[O] In-Su Kang Jae-Hak J. Bae Jong-Hyeok Lee
Dept. of Computer Science and Engineering, Division of Electrical and Computer Engineering
Pohang University of Science and Technology
and Advanced Information Technology Research(AITrc)
[+]School of Computer Engineering and Information Technology, University of Ulsan

요 약

기존의 다중 문서요약은 전체 대상문서에 대해서 한번에 요약문을 생산하지만, 본 논문은 요약 대상문서 집합에서 핵심내용을 갖는 문서를 기본 문서로 선택, 임시 요약문장을 추출하고 대상문서 집합에서 순차적으로 문서를 입력 받아 중요문장을 추출, 이전에 구축된 요약문장과 현재 추출된 문장을 비교하면서 요약에 필요한 문장을 선택하는 적응 기법을 제안한다. 제안한 방법으로 구현한 시스템은 NTCIR TSC 3에서 사용된 29개의 다중 문서집합을 통해서 성능을 평가하였다. 적응 기법 시스템은 TSC3의 baseline시스템인 Lead 방법보다는 높은 성능을나타냈지만, TSC 3에 참가한 시스템들과의 비교에서는 월등한 성능 우위를 나타내지 못했다.

1. 서 론

자동 문서요약에 대한 연구는 1950년에 단일문서 요약에서부터 시작하였다. 인터넷의 급속한 발전에 의해서 문서가 폭발적인 증가하면서 1990년대 후반부터 관련된 주제의 문서들을 요약하는 다중문서 요약이 등장하였다. 현재 다중 문서요약은 동일 주제별로 이루어진 문서들을 대상으로 요약하고, 미국의 DUC (Document Understanding Conference), 일본의NTCIR (NII-NACSIS Test Collection for IR Systems) 등의 국제 학회에서, 다중 문서요약 시스템에 대한 새로운 기술 및 평가 방법에 대해서 연구하고 있다.

다중문서 요약에 필요한 3가지 중요 기술로는 여러 문서에서 중요 문장을 추출하는 기술, 추출된 내용간에 중복된 내용 검사기술, 추출된 문장을 압축하고 교정하는 기술이 필요하다. 이 3가지 기술 중, 중요문장을 추출하는 기술과 중복된 내용을 검사하는 기술은 단일 문서요약과 달리 다중문서 요약에서 필요한 기술로써 기존의 단일 문서요약의 연구에서는 수행되지 않았다. 이 중에서 여러 문서에서 중요문장을 추출하는 기술은 다중 문서요약에서 가장 중요한 방법으로써 다중 문서 요약을 연구하기 시작한 이후로 많은 방법들이 제안되었다.

본 논문은 다중 문서요약에서 중요문장 추출을 위한 방법으로 적응 기법을 제안한다. 적응 기법은 대상문서 집합에서 핵심내용이 있는 문서를 기본 문서로 선택, 요약문장을 추출하고 순차적으로 문서를 입력 받아 중요문장을 추출, 이전에 구축된 요약 문장과 현재 추출된 문장을 비교하면서 요약에 필요한 문장을 선택하는 문장 추출 방법이다.

2. 관련 연구

현재 다중 문서요약에서 중요 내용을 추출하는 방법으로 Mani와 Bloedern은 문서 집합으로 text span을 추출하기 위해서 관련된 lexical item의 네트워크를 구축하는 방법[1]과, 관련된 문서집합에서 유사성와 차이성을 이용한 추출방법[2]을 사용하였다. Radev와 McKeown은 전문화된 분야별 지식과 Template을 이용한 지식 기반(knowledge-based)방법[3]으로 다중 문서를 요약하였다. 다중 문서집합에서 중요 내용을 찾기 위해서 개별 문서보다는 전체 문서들에 중심이 되는 내

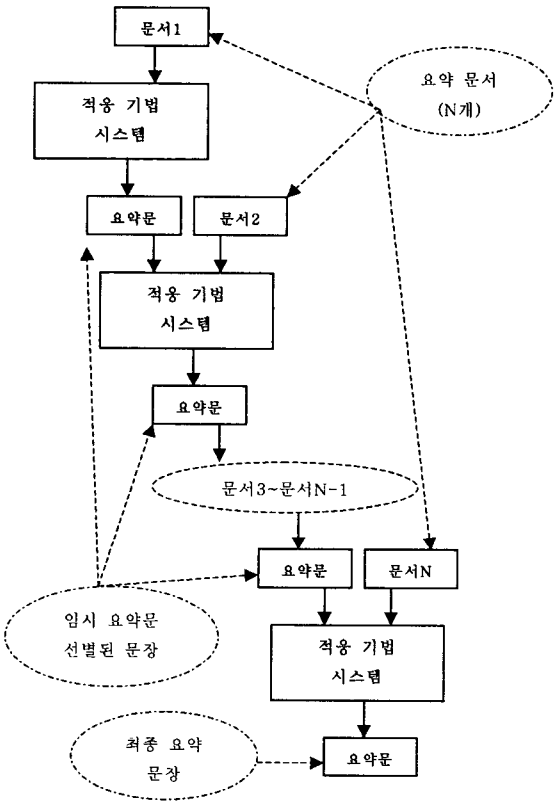
용을 찾는 방법으로 Centroid-based 방법[4]이 있다. Query-based 인 요약의 경우 Query와 문서간의 유사도, 문장이 속한 클러스터의 범위와 크기, 최근 문서에 가중치 부여, 문장의 통계적 언어학적 특성을 조합해서 가중치를 부여해서 문장간에 유사성과 상이점을 최대화하는 Maximal Marginal Relevance Multi-Document (MMR-MD) 방법[5]이 있다. Columbia University는 기계 학습방법과 통계적 특성을 이용해서 문장을 추출하고 융합하는 방법[6,7]을 사용한다. 문장에서 unigram, bigram, trigram으로 단어를 조합하고, 각각의 단어 조합에 주제와 관련된 문서집합, 주제와 관련 없는 문서집합을 통해서 확률을 구하고, 이를 이용해서 중요 내용을 선택하는 방법[8]이 있다. 또한, 개별 단어에 의존하지 않고 중요 단어들의 개념을 사용하여 동의어, 상위어, 하위어로 확장, 핵심 주제어의 개념집합을 이용해서, 중요 문장 및 내용을 추출하는 방법[9]이 있고, 파서를 이용해서 요약 대상문서의 모든 문장에서 구문정보 및 주어 와 동작, 시간, 장소, 위치 등의 색인정보를 구축하고 문장 추출에 이용하는 방법이 있다[10].

단일 문서요약에서 Discourse-level정보를 이용한 Text cohesion, Text coherence을 바탕으로 Lexical Chaining 이나, RST를 이용한 방법이 사용되는 것처럼 Radev는 다중 문서요약을 위한 교차 문서구조 이론 (Theory of Cross-Document Structure, CST)[11]의 필요성을 제기하였고, 다중 문서집합을 나타내기 위한 기본 구조로 육면체 형태와 그래프 형태를 제안하고, CST를 문장 수준에서 연구하여 subsumption, update, elaboration 등 24개의 CST 관계[12,13]를 발견했다. Radev가 제안한 CST는 서로 다른 문서에 존재하는 문장간에 관계를 찾았다. 그러나 문서집합은 관련된 주제를 갖는 문서이지만, 각 문서마다 저자가 다르기 때문에, 교차 문서의 문장간에 관계를 설정하고 시스템을 통해서 찾기가 어렵다. 또한, 다중 문서집합의 특성상 모든 문서간에 관계를 찾기 위해서는 다대다(M : N) 관계가 존재하기 때문에 모든 문장간에 관계를 부여하기 난해하고, 시스템을 통해서 발견하기 어렵다. 현재 Radev는 CST를 정형화하지 못했다[14].

3. 적응 기법을 이용한 중요문장 추출

3.1 적응 기법 개요

기존의 다중 문서요약은 모든 대상문서를 통해서 한 번 에 요약문을 생산하지만, 적응 기법을 이용한 방법은 요약 대상문서 집합에서 순차적으로 문서를 입력 받아 중요문장을 추출하고, 이전에 구축된 요약 문장과 추출된 문장을 비교하면서 요약에 필요한 문장을 선별한다. 적응 기법을 이용한 문장추출은 첫 문서의 내용과 문서의 입력순서에 따라서 성능이 변할 수 있고, 문장 비교에 사용되는 관계설정에 의해서도 성능이 변할 수 있다. 따라서 문서 집합에서 가장 중요한 "기본(baseline)문서의 선택", "문서의 입력순서", 추출문장과 현재 요약문장간의 비교를 위한 "문장간의 관계 설정"의 3가지 분야에 대한 연구가 병행되어야 한다.



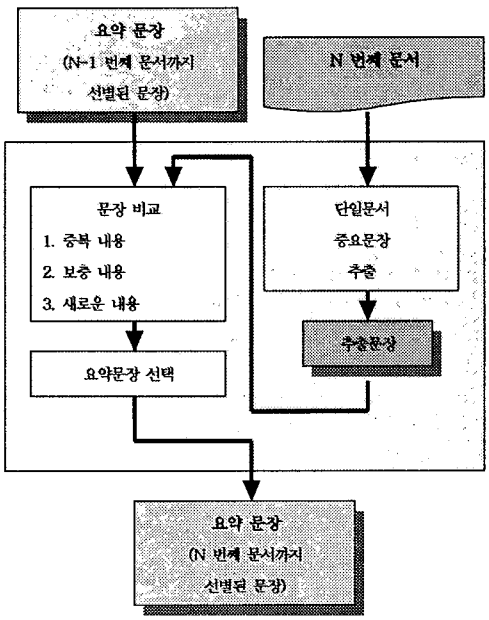
[그림 1] 적응 기법을 이용한 문장 추출

Mani 의 Automatic Summarization[14]에 의하면 사람이 구축하는 다중 문서요약 방법은 기본(baseline)이 되는 문서의 내용에 보충할 수 있는 내용을 다른문서에서 찾고, 전체 요약문을 교정한다. 적응 기법은 첫 문서를 기본(baseline)문서로 선택, 가중치를 부여하고, 그 외의 문서들에 보충 내용을 선택해서 전체 요약문을 작성한다면, 사람이 수행하는 다중 요약전략을 쉽게 적용할 수 있다. 또한, 적응 기법은 Radev가 제안한 CST 관계를 항상 임시 요약문과 현재 고려하는 문서에서만 찾기 때문에 다대다(M : N) 관계를 일대일(1 : 1) 관계로 바꿀 수 있는 장점이 있다. 모든 문장에서 구문정보 및 색인 정보를 구축하는 방법[10]에 비해서 적응 기법은 임시 요약문장을 선별하고, 선별된 문장에 대해서만 색인 정보 및 구문 정보를 구축하기 때문에, 중요 문장

추출 시 불필요한 문장에서 발생되는 정보를 제거할 수 있을 것이다.

3.2 적응 기법 시스템

적응 기법은 첫 문서로써 문서집합에서 기본이 되는 문서를 선택하고 문서들의 입력 순서를 결정하는 부분이 필요하지만, 현재 연구 중에 있어서 본 논문에서는 제외한다. 적응 기법 시스템은 단일문서 중요문장 추출 부분과 현재 요약문장과 추출된 문장간의 비교부분, 요약문 재구성을 위한 중요문장 선택부분으로 구성된다.



[그림 2] 적응 기법 시스템 구조

적응 기법 시스템에서 첫번째 문서의 입력시 구축된 요약 문장이 없기 때문에 첫 문서의 추출된 문장을 요약문장으로 하고, 두번째 문서부터 적응 기법 시스템의 문장 비교부분과 중요문장 선택부분이 이용된다. 각 부분의 세부 내용은 다음 항목에서 설명한다.

3.3 단일문서 중요문장 추출

단일문서에서 중요문장을 추출하기 위해서 기존의 전통적 자질인 문장의 위치, 길이, 제목의 단어, 감점단어 리스트들을 이용하였다. i번째 문장의 점수는 다음과 같이 계산된다 :

$$S_{ext}(S_i) = S_{pos}(S_i) \times w_{position} - Pen\ (S_i)$$
$$- |S_i \cap P| \times w_{stigma} \times + |S_i \cap L| \times w_{lead}$$

- S_{pos}(Si) : 문장 Si의 위치별 점수
- Pen(Si) : 문장 Si의 길이에 따른 점수
- L : 제목에 있는 단어집합
- P : 감점단어 리스트

각 문장의 부여된 점수에 따라서 높은 점수의 문장을 추출하였다. 사용된 자질들의 세부사항은 다음 항목에서 설명한다.

3.3.1 문장의 위치

문장의 위치는 자동 문서 요약이 연구되기 시작한 60년대부터 사용된 전통적인 자질이다. 문두의 문장일수록 문서에서 중요한 내용을 포함한다는 가정으로 문장에 점수를 부여한다. 그러나 몇몇 중요문장이 문서의 끝에 존재할 수 있고[15], 문서의 중간에 존재하는 문장에 부여된 점수는 위치에 따른 특성을 잘 반영하지 못한다는 가정을 하고 점수를 부여하였다. i번째 문장의 점수는 다음과 같이 계산된다 :

$$S_{pos}(S_i) = \max(1 - \frac{(i-1)}{n}, 1 - \frac{(n-i+1)}{n}, 0.3)$$

문서 중간에 위치한 문장의 점수가 0.3 이하로 낮아지는 것을 막기 위해서 0.3으로 고정하였다.

3.3.2 문장의 길이

중요문장을 선택할 때 너무 길거나 짧은문장은 음의 가중치(penalty)를 부과하여 중요한 문장에서 배제하는 방법이다. 한국어 신문기사에서 중요한 문장이라고 추출되는 문장의 길이는 10-30 어절 사이에 존재한다[16]. 따라서 문장 길이가 10-30어절 사이를 벗어날 경우 문장에 감정을 부여하였다.

3.3.3 감점 단어리스트

인용문을 갖는 문장이 요약 결과에 포함될 경우, 중복된 내용을 포함할 수 있는 가능성이 있기 때문에[8], 따옴표(")를 갖는 문장에 대해서 음의 가중치(penalty)를 부여하였다.

3.3.4 제목의 단어

문서의 제목는 문서가 전달하려는 정보를 가장 간단한 형태로 표현한 문장이다. 따라서 제목에 포함된 단어를 많이 갖는 문장일수록, 문서가 전달하려는 핵심 내용을 포함할 가능성이 크므로, 제목의 단어에 따른 가중치를 문장에 부여하였다.

3.4 문장 비교

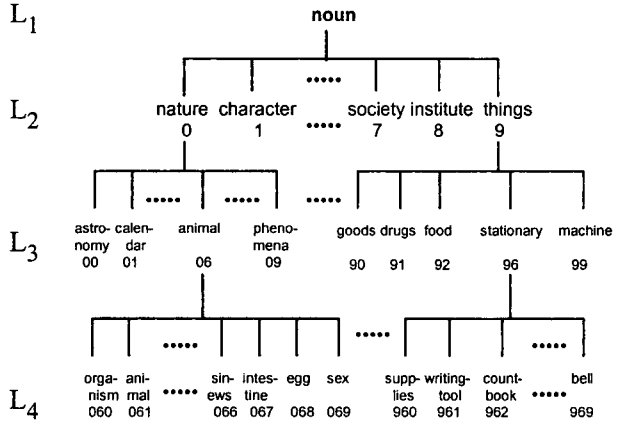
단일문서에서 추출된 문장과 현재까지 구축된 요약문장간에 존재하는 관계를 구분하고, 추출된 문장이 앞으로 생성될 요약문장에 포함될 가능성이 있는지 확인하는 부분이다. 문장간의 관계에 Radev가 제안한 24개의 CST의 관계를 부여할 수 있지만, CST는 아직 정형화되지 않았고, 관계를 찾기 어렵기 때문에, 문장간의 관

계를 3가지로 제한하였다. 기존의 요약문장에 중복된 내용을 갖는 문장, 보충하는 내용을 갖는 문장, 새로운 내용을 갖는 문장으로 관계를 구분하였다. 보충 내용을 갖는 문장과, 새로운 내용을 갖는 문장에 대해서는 앞으로 생성될 요약문에 포함될 기회를 부여하고, 중복된 내용을 갖는 문장과, 위의 3가지 관계 모두에 해당되지 않는 문장은 제거하였다.

3.4.1 중복 문장 구별방법

추출된 문장과 현재 요약문장 사이에 중복된 내용의 구별방법은 두 문장간의 중복단어 수에 기반한 Dice coefficient를 이용하였고, 구문정보를 이용해서 단어 가중치를 추가하는 방법으로 개선하였다[17]. 기존의 방법은 중복도 계산시 문장에 포함된 모든 단어를 대상으로 측정하지만[4], 문장에서 의미를 갖는 핵심 단어에 대해서는 가중치를 부여하여 중복도 측정을 개선하였다. 직관적으로 복문인 문장에서 주절은 종속절보다 상대적으로 중요한 내용을 갖고 있을 것이고, 문장의 구성요소 중에서, 주어, 목적어, 동사 등은 구문단계에 다른 요소에 비해서 상대적으로 중요할 것이다. 그러므로 주절과 종속절에 따른 가중치와 단어의 구문정보에 따른 가중치를 부여하였다.

주제별로 연관된 문서집합에서는 동일한 내용을 포함하는 문장이라도 유사한 의미를 갖는 다른 단어의 형태로 표현될 수 있기 때문에, 문장간의 단어 비교에서는 단어의 형태만을 고려하지 않고 단어가 갖는 개념코드간에 비교를 추가하였다. 시스템구현을 위해서 사용한 개념코드는 Kadokawa 시소러스의 체계를 사용하였다.



[그림 3] Kadokawa 시소러스의 구조

Kadokawa 시소러스는 4단계(L1, L2, L3, L4) 구조의 1110개 의미적 클래스(semantic class)를 갖고 있는다. L1, L2, L3 단계는 10개의 하위클래스를 갖고 L4 단계는 000과 999사이의 3자리 코드를 갖는다[18]. 두 문장 S_1과 S_2의 중복도는 다음과 같이계산된다 :

$$Sim\ (S_1, S_2) = \frac{\sum_{(t_i,t_j)\in S_1 \overset{*}{\cap} S_2} W(t_i,S_1)+W(t_j,S_2)}{\sum_{t\in S_1} W(t,S_1)+\sum_{t\in S_2} W(t,S_2)}$$

- $S = \{t_1, \Lambda\ , t_n\}$
- $W(t,S) = W_{st}(t,S) + W_{gr}(t,S)$
- $W_{st}(t,S)$: 단어 t 의 문장종류에 따른 가중치
- $W_{gr}(t,S)$: 단어 t 의 구문정보에 따른 가중치
- $\overset{*}{\cap} = \overset{L}{\cap} + \overset{C}{\cap}$
- $S_1 \overset{L}{\cap} S_2 = \{(t_i,t_j) \mid t_i \in S_1, t_j \in S_2, t_i = t_j\}$
- $S_1 \overset{C}{\cap} S_2 = \{(t_i,t_j) \mid t_i \in S_1, t_j \in S_2, C(t_i) \cap C(t_j) \neq \phi\}$
- $C(t)$: 단어 t 의 개념코드 집합

중복도 점수가 주어진 임계값(T_R)을 초과할 경우, 두 문장은 중복된 내용을 갖는다고 간주 하였고, 두 문장 중에서 최근 문서에 포함된 문장을 현재 요약문에 남긴다. 이 단계에서 얻은 단어들의 구문정보 및 문장의 종류에 따른 가중치 정보는 보충 내용 문장 구별 및 새로운 내용 포함 문장 구별과정에도 사용된다.

3.4.2 보충 내용포함 문장 구별방법

보충내용 문장을 구별하는 방법은 추출된 문장과 현재 요약문장간에 중요단어의 문장종류에 따라서 구별하였다. 복문의 경우, 요약문장의 종속절에 나타나는 단어들이 추출된 문장의 주절에 나타날 경우, 추출된 문장이 요약문장의 내용을 보충한다고 가정하였고, 단문일 경우는 주절의 단어와 종속절의 단어 구분 없이 전체 단어에 대해서 검사하였다. 현재 요약문장에 포함된 문장 S의 주절 단어집합을 $Sm=\{st_1, st_2, .st_n\}$, 추출된 문장 E의 종속절 단어집합을 $Es=\{et_1, et_2.. et_n\}$, 이라고 할 때, 두 문장은 다음과 같이 계산된다 :

$$Sm_{rate} = \frac{\sum_{st\in S\cap E} W(st,Sm)}{\sum_{st\in S} W(st,Sm)} \times w_{main\_term}$$

$$Es_{rate} = \frac{\sum_{et\in S\cap E} W(et,Es)}{\sum_{et\in E} W(et,Es)} \times w_{sub\_term}$$

S_{mrate}과 E_{srate} 값이 각각 주어진 임계값(T_S ,T_E)을 초

과할 경우 추출된 문장은 요약문장 S에 대해서 보충하는 내용을 포함한다고 간주하고, 앞으로 생성될 요약문장 후보 목록에 저장한다.

3.4.3 새로운 내용포함 문장 구별방법

중복내용 및 보충내용 구별방법에서 선별되지 않는 문장들에 새로운 내용을 포함하고 있는지 검사한다. 단일 문서입력에 따른 요약문장을 생성하면서 각 문서의 제목을 저장하고, 제목에 포함되는 단어들을 이용해서 전체문서의 핵심 공통단어 집합을 형성한다. 핵심 공통단어 집합을 G, G에 포함된 단어 k의 빈도수를 f_k, 추출된 문장 E의 단어 집합을 $\{t_1, t_2, .. t_n\}$ 이라고 할 때, 다음과 같이 계산된다 :

$$N_E = \sum_{t \in E \cap G} W(t,E) \cdot f_t \times w_{weight}$$

계산된 NE의 값이 주어진 임계값(T_N)보다 크면 생성문장 후보목록에 추가한다.

3.5 요약 문장 선택

추출된 모든 문장들과 현재 요약문장들을 비교해서 요약문장 후보 목록을 작성하면, 후보 목록과 현재 요약문장들 중에서 앞으로 생성될 요약문장을 선택하여야 한다. 일반적으로 DUC, NTCIR의 문서요약 분야에서는 문서 집합 별로 요구하는 요약문의 길이 및 문장 개수가 제공된다. 문장을 선택하기 위해서 사용한 자질은 문장 별로 중복문장 포함개수, 단일 문서 문장추출에 사용된 점수, 및 핵심 공통단어 집합을 사용하였다. 여러 문서에서 서로 동일한 내용을 포함하고 있다면 중복되는 내용은 전체 문서집합에서 핵심내용이 될 가능성을 갖기 때문에[19], 중복 문장 개수를 요약 문장 선택의 자질로서 선택하였다. 요약 후보 문장들과 현재 요약문장들의 집합을 $P=\{S_1, S_2, .S_n\}$라고 할 때, 모든 문장들은 다음과 같이 점수가 부과 된다 :

$$S_{select}(S_i) = R(S_i) \times w_{redundnat} + S_{ext}(S_i) \times w_{ext\_score} + \sum_{t \in S_i \cap G} W(t,S_i) \cdot f_t \times w_{global}$$

$-R(S_i)$: 문장 S_i에 중복된 문장 개수

각 문장 별로 점수가 부여되면, 문서집합이 요구하는 문장의 길이나 크기에 따라서 요약 문장을 선택한다. 선택된 문장들은 새로운 단일 문서의 입력시 적응 기법 시스템의 현재 요약문장으로 제공되고, 새로운 입력 문서가 없을 경우, 최종 요약문장이 된다.

4. 실험에 사용한 문서집합 및 평가 방법

4.1 실험 문서집합

제안한 방법을 평가하기 위해서 주제별 연관된 29개의 문서집합에 대해서 중요문장 추출의 성능 측정하였다. 실험에 사용한 문서집합은 2003년 NTCIR TSC3 NII-NACSIS Test Collection for IR System, Text Summarization Challenge 3) 에서 사용한 문서집합을 사용하였다. TSC3의 문서집합은 1998년, 1999년 마이니치, 요미우리 신문 기사들로 구성되었다. TSC3 에서는 참가자들에게 문장 추출 성능을 평가할 수 있는 프로그램과 정답 문장 집합을 제공 하였고, 제안한 방법을 평가하기 위해서 이를 사용 하였다. NTCIR TSC 3는 일본어 문서집합이기 때문에, 한국어 다중 문서요약 시스템에 적용하기 위해서, 일한기계번역 시스템(COBALT-JK) [20]을 사용해서 일본어 문서를 한국어 문서로 변역 하였다. 기계번역에서 발생되는 오류 때문에 번역된 문서를 사용한 실험은 시스템의 성능 판단에 영향을 줄 수 있다. 하지만 현재 한국어 문서로 구성된 다중 문서 집합 및 정답 요약문이 없기 때문에, 본 연구의 시스템 성능을 평가 하기 위해서 TSC 3의 일본어 문서집합을 번역해서 사용하였다.

4.2 평가방법

평가에 사용한 프로그램은 NTCIR TSC3에 사용한 프로그램이기 때문에, Hirao, Okumura의 TSC 3 Overview 논문[21]의 내용을 통해서 평가 방법을 간단하게 설명한다.

4.2.1 정확도(Precision)

정확도는 시스템이 추출한 문장들이 정답 문장집합에 얼마나 많이 포함되어 있는지를 평가하는 방법이다. 정확도를 구하는 수식은 다음과 같다.

$$precision = \frac{m}{h}$$

h는 Abstract 요약문을 구성하기 위해서 필요로 하는 최소한의 문장 개수고, m은 시스템이 추출한 문장 중에 정답 문장집합에 포함되는 문장 개수이다.

Abstract 문장 번호	일치하는 문장집합
1	$\{S_1\}$__$\{S_{10}, S_{11}\}$
2	$\{S_3, S_5, S_6\}$
3	$\{S_{20}, S_{21}, S_{23}\}$__$\{S_1, S_{30}, S_{60}\}$

[표 1] 정답 문장 집합 예

4.2.2 적용률(Coverage)

적용률은 시스템이 추출한 문장간의 중복도을 고려하여 시스템의 결과가 사람이 작성한 정답 요약문에 얼마나 가까운지를 평가하는 방법이다.

요약 대상 문서들에서 사람이 작성한 Aabstact 요약문 i번째 문장과 일치하는 문장 집합을 $A_{i,1}$, $A_{i,2}$,..., $A_{i,m}$. 으로 표시한다. 이때 m개의 일치하는 문장을 갖는다. 여기서, $A_{i,j}$ 는 요약 대상문서들의 문장에 일치하는 원소의 집합을 나타낸다. $A_{i,j}=\{G_{i,j,1}, G_{i,j,2}$., $G_{i,j,k}\}$. 예를 들어 [표 1.]의 $A_{1,2}=G_{1,2,1}$, $G_{1,2,2}$이고 $G_{1,2,1,}=s_{10}$, $G_{1,2,2}=s_{11}$. 이다. Abstract 요약문의 i 번째 문장을 평가하기 위해서 $e(i)$ 함수를 다음과 같이 정의한다.

$$e(i) = \max_{i \le j \le m} \left(\frac{\sum_{k=1}^{|A_{i,j}|} v(G_{i,j,k})}{|A_{i,j}|} \right)$$

식의 v()는 시스템이 $\acute{\alpha}$를 생산하면 1, 그렇지 않으면 0을 부과한다. $e(i)$ 함수는 어떤 $A_{i,j}$ 라도 완전하게 만족하면 1, 그렇지 않으면 $A_{i,j}$ 에 일치하는 문장의 개수만큼 부분 점수를 부여한다. n개의 문장을 갖는 Abstract 요약과 $e(i)$ 함수를 통해서 적용률은 다음과 같이 정의된다.

$$Coverage = \frac{\sum_{i=i}^{n} e(i)}{n}$$

시스템이 추출한 문장이 "S_{10}, S_{11}, S_5, S_{17}, S_{60}, S_{61}"라고 한다면, $e(i)$ 값은 다음과 같이 계산된다.

$$e(1) = \max(1,1) = 1$$
$$e(2) = \max(0.67) = 0.67$$
$$e(3) = \max(0,0.67) = 0.67$$

각 $e(i)$ 합과 3인 문장 개수를 통해서 적용률은 0.553이 된다.

5. 실험 및 결과 분석

제안한 방법을 평가하기 위해서 3절에서 제안한 적응 기법을 이용한 문장 추출 시스템을 구축하고, NTCIR TSC 3에서 사용된 29개의 문서집합을 번역해서 실험 대상 문서로 사용하였다.

적응 기법을 이용하기 위해서는 3.1에서 언급한 3가지 선행 연구 "기본 문서의 선택", "문서의 입력순서", "문장간의 관계설정"이 필요하지만, 현재 연구 중에 있기 때문에 각 부분에 대한 상세구현 없이 3.2의 내용을 바탕으로 기본 시스템을 구현하였다. 첫 번째 문서로는 가장 오래된 날짜에 작성된 문서들 중에서 임의의 문서를 선택하였고, 가중치는 부여하지 않았다. 문서의 입력 순서로는 각각의 문서가 갖고 있는 작성 날짜를 이용해서, 오래된 문서부터 입력하고 최근 문서를 마지막에 입력했다. 문장간의 관계는 3.4절에서 언급한 3가지 관계만을 사용하였다.

단일문서 문장추출에서는 문서마다 70-80% 비율로 문장을 추출하였고, 실험 대상문서 집합이 신문기사로 구성되었기 때문에 문장 위치에 따른 점수에 상대적으로 높은 가중치를 부여 하였다. 추출문장과 요약문장의 비교에서는 중복도 임계값 T_R은 0.8, 보충내용 선택 임계값 Tc는 0.4, T_E는 0.3, 새로운 내용을 선택하는 임계값 T_N는 0.5로 설정하였다. NTCIR TSC 3에서는 문서 집합마다 짧은(Short) 요약과, 긴(Long) 요약을 평가하였다. 구현한 적응 기법 시스템은 요약하려는 문장의 개수를 긴 요약에 필요한 만큼으로 설정하고 요약 문장을 추출, 추출한 문장 중에서 높은 점수를 갖는 문장을 선택하여 짧은 요약에 사용하였다.

시스템	Short		Long	
	Cov.	Prec.	Cov.	Prec.
적응 기법	**0.285**	**0.485**	**0.323**	**0.544**
LEAD	0.212	0.426	0.259	0.539

[표 2] 적응 기법 시스템 실험 결과

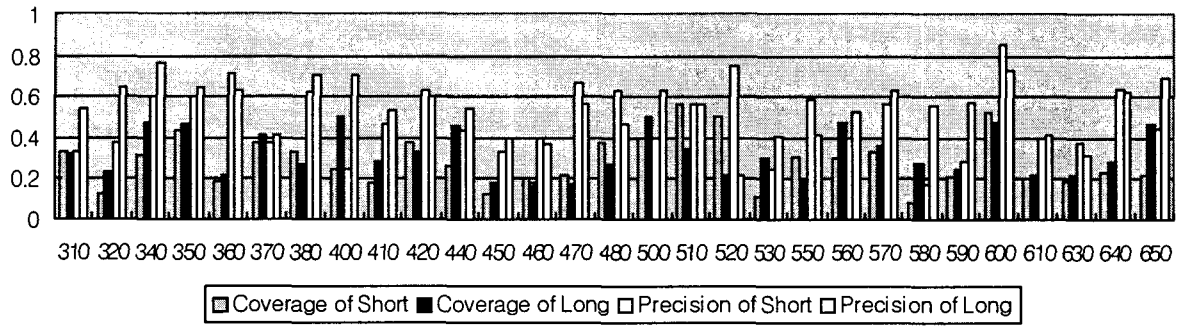
실험 결과 전체적인 성능은 TSC3에서 baseline 시스템으로 사용된 LEAD 방법보다는 높은 성능을 나타냈지만, Long 형태의 요약에서 정확도 차이는 다소 낮았다. 적응 기법 시스템의 성능을 NTCIR TSC3에 참가한 시스템의 문장추출과 비교한 결과, 전체 11개 시스템(제안한 방법 포함) 중에서 6-7번째의 순위를 보였다. 실험에 사용한 문서가 번역 오류를 갖고 있지만, 제안한 방법이 다른 시스템에 비해 좋은 성능을 나타내지는 못했다.

System ID	Short		Long	
	Cov.	Prec.	Cov.	Prec.
F0301(a)	0.315	0.494	0.355	0.554
F0301(b)	0.372	0.591	0.363	0.587
F0303(a)	0.222	0.314	0.313	0.432
F0303(b)	0.293	0.378	0.295	0.416
F0304	0.328	0.496	0.327	0.535
F0306	0.283	0.406	0.341	0.528
F0307	0.329	0.567	0.391	0.680
F0309	0.308	0.505	0.339	0.585

F0310	0.181	0.275	0.218	0.421
F0311	0.251	0.476	0.247	0.547
적응 기법	0.285	0.485	0.323	0.544
Ranking	7/11	6/11	6/11	6/11

[표 3] NTCIR TSC3 참가 시스템과의 성능 비교

적인 실험이 필요하다. "기본문서의 선택" 부분은 문서 분류를 이용한 방법[15,22]를 적용하고, "문서의 입력순서" 부분은 기존의 시간 자질에 대한 연구[23]를 적용할 계획이다.



[그림 4] 문서 집합 별 성능 결과

문서 집합 별로는짧은 요약문과 긴 요약문 간에 정확도 및 적용률이 유사하였지만, 몇몇 문서 집합 310, 320, 340, 520, 580, 650 에서는 상대적으로 큰 차이를 보였다. 추출된 문장에서 짧은 요약에 포함되지 않지만, 긴 약에 포함되는 문장에 대한 분석이 필요할 것으로 생각된다.

6. 결론 및 향후 연구

본 논문에서는 다중 문서 요약에서 중요 문장 추출을 위한 방법으로 적응 기법을 제안하였다. 적응 기법은 문서집합에서 기본이 되는 문서를 바탕으로 첫 요약문장을 구성하고, 순차적으로 문서를 입력 받아 문장을 추출하고 현재 요약문장과 비교를 통해서 최종 요약문장을 추출하는 방법이다. 실험을 통해서 적응 기법이 Lead 방법에 비해서 높은 성능을 나타냈지만, TSC 3에 참가한 시스템에 비해서는 상대적으로 만족할 만한 성능을 보이지 못했다.

제안한 적응 기법은 "기본문서의 선택", "문서의 입력순서", "문장간의 관계 설정"을 필수 연구 요소로 갖는다. 본 논문의 실험에 사용한 시스템은 "기본문서의 선택", "문서의 입력순서" 부분을 간단한 방법으로 구현하였으며, "문장간의 관계 설정"에 사용한 3가지 관계 및 이에 따른 문장의 중요도 설정 역시 검증하지 못했다. 향후 적응 기법의 성능 향상을 위해서는 본 논문에서다루지 못했던 "기본문서의 선택", "문서의 입력순서" 부분의 연구 와 "문장간의 관계 설정" 부분의 개별

감사의 글

본 연구는 첨단정보기술 연구센터(AITrc)를 통하여 과학재단의 지원을 받았습니다.

참고 문헌

[1] I. Mani and E. Bloedorn, 1997. *Multi-document summarization by graph search and merging.* In Proceedings of AAI-97, pages 622-628

[2] I. Mani and E. Bloedorn, 1999. *Summarizing similarities and differences among related documents.* Information Retrieval, 1 : 35-67

[3] D. R. Radevand K. McKeown, 1998. *Generating natural language summaries from multiple on-line sources.* Computational Linguistics, 24 (3), pages 469-500, 1998

[4] D. R. Radev, H. Jing, M. Budzikowska, 2000. *Centroid-Based Summarization of Multiple Documents.* ANLP/NAACL Workshop, 2000

[5] J. Goldstein, V.Mittal, J.Carbonell, M.Kantrowitz, 2000. *Multi-Document Summarization By Sentence Extraction.* ANLP/NAACL Workshop, 2000

[6] K. McKewon, J. L. Klavans, V. Hatzivassiloglou, R. Bazilay, E. Eskin, 1999. *Towards Multi-document Summarization by Reformulation : Progress and Prospects.* Proceedings of the

AAAI, 1999

[7] R. Barzilay, K. McKeown, M. Elhadad, 1999. *Information Fusion in the Context of Multi-Document Summarization.* Proceedings of the 38th Annual Meeting of the ACL, 1999

[8] C. Y. Lin and E. Hovy, 2002. *Form Single to Multi-document Summarization : A Prototype System and its Evaluation.* Proceedings of the 40th Annual Meeting of the ACL, p.457-464, 2002

[9] B. Schiffman, A. Nenkova, K. McKeown, 2002. *Experiments in Multidocument Summarization*, HLT conference, 2002

[10] Y. Guo and G. Stylios, 2003. *A New Multi-document Summarization System.* DUC 2003

[11] D. R. Radev, 2자동 구축된 문맥 패턴과 개체명 사전에 기반한 제목 개체명 인식000. *A Common Theory of Information Fusion from Multiple Text Sources Step One : Cross- Document Structure.* ACL SIGDIAL Workshop, 2000

[12] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, 2001. *Experiments in Single and Multi-Document Summarization Using MEAD.* DUC, 2001

[13] D. R. Radev, S. J. Otterbacher, H. Qi, D. Tam, 2003. *MEAD ReDUCs : Michigan at DUC2003*, DUC, 2003

[14] I. Mani, *Automatic Summarization* John Benjamins Publishing Company

[15] C. Nobata, S. Sekine, K, Uchimoto, H. Isahara, 2002, *A Summarization system with categorization of document sets*, NTCIR workshop 3 meeting TSC2

[16] J. M. Yoon. 2002. *Automatic summarization of newspaper articlesusing activation degree of 5W 1H. Master's thesis*, POSTECH

[17] J. M. Lim, I. S. Kang, J. H. Bae, J. H. Lee, 2003. *Measuring Improvement of Sentence-Redundancy in Multi-Document Summarization.* Proceedings of the 30th KISS fall conference, 2003

[18] S. Ohno and M. Hamanishi, *New synonyms Dictioinary, Kadokawa Shoten*, Tokyo, 1981.

[19] D. R. Radev. 1999 *Topic Shift Detection-Finding new information in threaded news.* Technical Report CUCS-026-99, Columbia University Department of Computer Science. 1999

[20]] C. J. Park, J. H. Lee, G. B. Lee, and K. Kakechi, *Collocation-Based Transfer Method in Japanese -Korean Machine Translation.*Transaction of Information Processing Society of Japan, 1997, 38(4), page 707-718.

[21] T. Hirao and M. Okumura, 2004. *Text Summarization Challenge 3-Text summarization evaluation at NTCIR Workshop4*, NTCIR Workshop 4, 2004

[22] K. R. McKeown, R Barzilay, D. Evans, V. Hatzivassilogou, M. Y. Kan, F. Schiffman and S. Teufel. 2001. *Columbia Multi-Document Summarization : Approach and Evaluation.* In Online Proceedings of DUC 2001.

[23] J. M. Lim, I. S. Kang, J. H. Bae, J. H. Lee, 2004. *Multi-Document Summarization Using Time Feature.* Proceedings of the 31th KISS spring conference, 2004