단락 자동 구분을 통한 중요 문장 추출

김계성°, 이현주, 정영규, 서연경, 손기준, 이상조
경 북 대 학 교 컴 퓨 터 공 학 과
{kskim, hyunju}@comeng.ce.knu.ac.kr sjlee@bh.knu.ac.kr

Setences Extraction System using Automatic Division of Paragraph

Kye-Sung Kim°, Hyun-Ju Lee, Young-Giu Jung, Youn-Kyoung Seo, Ki-Jun Son, Sang-Jo Lee
Dept. of Computer Engineering, Kyungpook Nat'l University

요 약

본 논문은 단락의 자동 구분을 통한 중요 문장 추출 시스템을 제안한다. 먼저 어휘의 재출현 여부와 어휘의 일치도, 어휘의 역할 변화를 파악하여 재출현 어휘에 대한 양상을 분석하고 이를 통하여 문장 간의 긴밀도를 정량적으로 계산한다. 다음으로 측정된 문장 간 긴밀도를 이용하여 사용자의 추출 범위에 따라 단락을 구분하고, 각 단락의 대표 문장을 선정하여 최종 요약문을 생성한다. 제안한 방법은 문서 제목, 문장의 위치, 수사 구조 등의 정보를 이용하지 않으며, 단순히 어휘의 출현 빈도만을 이용하던 기존의 통계적인 방법보다 질높은 요약문을 생성할 수 있다. 또한 제안한 방법론은 본 논문이 대상으로 삼고 있는 신문기사의 영역뿐만 아니라 다른 영역으로의 적용이 가능하다.

1. 서 론

인터넷의 발달과 함께 사회는 빠르게 변화하고 있다. 회사나 가정에서 컴퓨터를 통하여 정보를 찾고 상품을 구입하며 회의나 거래를 하고, 자택근무도 가능하게 되었다. 이러한 인터넷의 발달과 더불어 정보에 대한 개념이 달라지고 있으며, 좀더 빠르고 정확하게 정보를 검색할 수 있는 여러 방법이 모색되고 있다.

특히 웹 문서의 분산화와 대용량화는 정보 검색 시스템의 필요성을 증가시켰으며, 이에 따라 자동 요약 시스템의 중요성도 함께 부각되고 있다.

자동 문서 요약에 관한 연구는 이미 1960년대부터 시작되었다. 하지만 최근들어 인터넷의 사용이 급격히 증가하고 문서의 디지털화가 급속화되기 시작하면서부터 그 관심이 크게 대두되고 있을 뿐만 아니라, 많은 연구들이 활성화되고 있다. 최근 들어 국내에서도 문서 요약에 대한 관심이 높아지고 있다[4,5,6]. 현재 진행되고 있는 국내의 연구들은 대부분이 문서에 나타난 어휘의 출현 빈도를 이용하여 중요 문장을 추출하고 있다.

본 논문에서는 어휘의 재출현 양상을 분석하고 단락의 자동 구분이라는 관점에서 중요 문장을 추출하고자 한다. 어휘의 재출현 양상을 분석함으로써 어휘의 출현빈도와 문서 내의 분포 정도만을 이용하는 기존 방법보다 질높은 요약문을 생성할 수 있을 뿐 아니라 문장 간의 긴밀도를 파악함으로써 수사 구조가 자주 발견되지 않는 문서에도 효과적으로 적용할 수 있다.

2장에서는 문서 요약에 관한 기존 연구를 살펴 보고 3장에서 단락 구분을 통한 중요 문장의 추출 방안에 대해 알아본다. 그리고 4장에서 실험을 하며, 5장에서 결론을 맺는다.

2. 연구 동향

자동 문서 요약에 대한 기존 연구는 크게 통계 정보에

기반한 방법과 지식에 기반한 방법으로 나누어 진다.

먼저 통계 정보에 기반한 방법은 문서로부터 추출된 명사의 출현 빈도수를 계산하여 문서를 대표하는 어휘 집합을 설정한 후 문장의 중요도 가중치에 반영하는 방법이다[1]. 이 방법은 비교적 적은 비용으로 빠르게 어느 정도 신뢰할 수 있는 결과를 얻을 수 있다는 장점이 있는 반면, 문장이나 단어를 단순히 열거하는 수준에 그치기 때문에 추출된 요약문의 내용이 부자연스럽다.

다음으로 지식에 기반한 방법은 문서에 포함된 각 문장의 의미와 문장 간의 관계 분석 등을 통한 문맥 구조의 파악을 바탕으로 이루어진다. 이 방법은 해당 분야의 지식과 문장의 문법적 구조를 기반으로 고품질의 요약문을 생성하나, 복잡한 자연어 처리 과정을 요구하기 때문에 구현하는 데에 많은 어려움이 따른다. 현재 이러한 지식에 기반한 연구는 대부분 수사 구조(Rhetorical Structure)에 그 기반을 두고 있다[2,3]. 이 방법은 수사 구조에 의해 문장의 중요도를 평가하고, 이를 통해 요약문을 생성한다.

3. 단락의 자동 구분을 통한 중요 문장 추출

텍스트는 문맥 위에서 의미적으로 연결된 문장들이 서로 관계를 이루며 나열되어 있다.

본 연구는 문장과 문장 사이의 내용이 얼마나 긴밀한가를 파악하려는 데 초점을 맞춘다. 실제 이러한 연구를 위해서는 통사적인 접근과 의미적인 접근이 동시에 이루어져야 하지만, 언어 처리에 대한 의미 연구는 아직까지 계속해서 풀어나가야 하는 문제이므로, 본 연구에서는 통사적인 접근만을 가지고 문장 사이의 긴밀도를 파악하고자 한다. 문장 간 긴밀도는 단락을 자동 구분하는데 이용되며, 이를 통하여 각 단락의 대표 문장을 추출함으로써 요약문을 생성하게 된다.

본 논문에서 제안한 단락의 구분을 통한 중요 문장 추출기의 전체 구성은 다음과 같다.

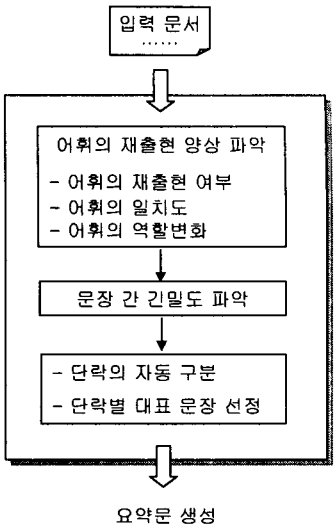

그림 1. 시스템 구성도

다음의 예를 살펴보자.

LG화학 독자신용으로 1억불 차입 <이동주>
ＬＧ화학(www.lgchem.co.kr)은 정보전자소재와 생명과학사업의 신규 설비투자자금융으로 1억달러의 외화를 차입했다.
ＬＧ화학은 최근 ＩＮＧ베어링과 산업은행 공동주간사로 국내외 19개 금융기관이 참여한 가운데 홍콩에서 차입계약을 체결했다.
이번 외화차입은 외환위기 이후 국제신용평가기관의 신용등급 판정없이 개별기업의 독자신용으로 이뤄진 첫 사례다.
차입조건은 리보(ＬＩＢＯ)+1.5%의 금리에 만기 2년이다.
ＬＧ화학은 지난해말 한국신용평가정보 등 국내 신용평가기관들로부터 투자등급을 상향조정 받았으나 외국기관들에게는 금년중으로 신용평가를 의뢰한다는 계획이다.
ＬＧ관계자는 "이번 차입계획은 당초 7500만달러를 목표로 한 것인데 지난 2월말 마감한 결과 국내외 금융기관들이 1억5000만달러를 신청해 옴에 따라 차입규모를 다소 늘리게 됐다"고 말했다.
ＬＧ화학은 이 자금을 주력사업인 정보전자소재와 생명과학분야에 집중 투자한다는 계획이다.

위의 문서에서 전체 내용의 응집력(cohesion)을 높이는 어휘는 '차입'과 'LG화학'이다. 특히 '차입'은 "외화를 차입하다->차입계약->외화차입->차입조건->차입계획->차입규모" 등의 어휘로 재출현하면서 문장과 문장 사이의 긴밀도를 높이는 동시에 문서 전체의 응집력(cohesion)의 형성에 크게 기여하고 있음을 볼 수 있다.

따라서 본 논문에서는 이러한 어휘의 재출현 양상을 세 가지 기준에서 분석하고, 이를 문장 간 긴밀도 계산

에 적용한다.

3.1 어휘의 재출현 양상

본 연구에서 설정한 어휘의 재출현 양상은 다음과 같다.

어휘의 재출현 양상
(1) 재출현 여부
(2) 어휘의 일치도
(3) 어휘의 역할 변화

문장 간 긴밀도 계산을 위해서 어떤 어휘가 다음 문장에서 재출현하고 있는지의 여부를 살핀다(1). 만약 재출현 어휘가 존재한다면, 두 어휘간의 모습이 얼마나 일치하는지(2), 그리고 재출현한 어휘의 역할에 변화가 일어났는지를 분석한다(3). 여기서 (2)와 (3)을 분석하는 것은 어휘의 일치도와 어휘의 역할 변화가 문장 사이에 나타난 내용의 긴밀도에 많은 차이를 주기 때문이다.

어휘의 일치도는 재출현한 어휘가 이미 앞에서 출현한 어휘와 완전히 일치하는지 혹은 부분 일치하는지에 따라 다시 다음과 같이 나눈다.

완전일치	예. LG화학 -> LG화학	
부분일치	중심어로 이동	예. 차입하다-> 외화차입
	수식어로 이동	예. 차입하다-> 차입조건

여기서 '부분 일치'는 그 모습을 달리 하면서 다음 문장으로 옮겨간 것을 말하는 것으로, 새로운 복합 명사를 형성하거나 접사가 결합되어 재출현한 경우가 대부분 여기에 해당한다. 위의 경우에 '외화 차입'은 '외화의 차입'으로, '차입조건'은 '차입에 대한 조건'으로 파악할 수 있고, 두 경우에서 보여지는 '차입'에 대한 역할이 서로 다르기 때문에 이를 중심어로의 이동과 수식어로의 이동으로 구분하여 다음에서 설명할 문장간 긴밀도 분석에서 달리 적용한다. 특히 본 논문에서 대상으로 하고 있는 신문 기사의 경우는 새로운 복합 명사가 생성되거나 한자어 접사들이 많이 출현하여 문장 사이의 긴밀도를 높이는 데 중요한 역할을 하고 있다.

그리고 재출현 어휘의 일치도를 분석함과 동시에 어휘의 역할 변화를 수집하게 되는데, 이것은 구문분석(parsing)에 그 기반을 둔다. 다시 말해서, 서술어의 역할을 하던 어떤 어휘가 다음 문장의 주어로 출현한 경우와 수의 성분으로 출현한 경우를 구분하여 문장 사이의 긴밀도에 달리 적용하고자 하는 것이다. 본 연구에서는 재출현한 어휘의 역할 변화를 다음과 같이 구분한다.

어휘의 역할 변화
a. 필수 성분 -> 수의 성분
b. 수의 성분 -> 필수 성분
c. 역할 변화 없음

세가지 관점에서 어휘의 재출현 양상이 파악되면, 이를 바탕으로 문장 간 내용의 긴밀도를 정량적으로 계산하고, 이를 통해 단락을 구분한다.

3.2 문장 간 긴밀도

문장 간의 긴밀도는 두 문장의 내용이 얼마나 긴밀한지를 살피는 척도이다. 문장 간 긴밀도 파악을 위해 이용하는 출현 어휘의 각 양상에 대한 가중치는 다음의 조건을 고려하여 달리 설정한다.

- 어휘의 일치도에 대한 가중치
  - 완전일치 > 중심어 부분일치 > 수식어 부분일치
  - 복합명사 일치 > 단일명사 일치
- 어휘의 역할 변화에 대한 가중치
  - 필수성분 > 수의 성분 > 역할 변화 없음

특히, 어휘의 일치도를 살필 때 복합 명사가 단일 명사보다 어휘의 특정성이 높다는 점을 이용하여 복합 명사의 일치를 단일 명사의 일치보다 가중치를 높게 부여한다.

두 문장 간의 긴밀도는 재출현 어휘가 많을수록 높아지며, 그 형태가 완전히 일치하면서 문장의 필수 성분으로 이동한 경우에 가장 높은 점수를 받게 된다. 두 문장의 긴밀도 CS는 다음과 같이 계산한다.

$$CS(Si,Sj) = \sum_{k}^{n} Lex[k]$$

여기서 n은 문장 Si와 Sj에서 재출현한 어휘의 개수이고, Lex는 재출현한 어휘의 양상 점수이다. Lex는 다음과 같이 계산한다.

$$Lex = (\alpha \times AG(wi,wj)) + (\beta \times RL(wi,wj))$$

여기서 AG(wi,wj)는 재출현 어휘에 대한 일치도 점수이고, RL(wi,wj)는 재출현 어휘의 역할 변화에 대한 값이다. α, β 는 각각 어휘의 일치도와 어휘의 역할 변화에 대한 가중치이다. 측정된 CS(Si,Sj)의 값이 클수록 두 문장 간의 내용이 서로 긴밀함을 의미한다.

본 논문에서는 글의 흐름을 전제로 문장 간의 긴밀도를 파악한다. 즉 문장 Sn->Sm(n<m)사이의 긴밀도만을 분석 대상으로 설정한다. (여기서 n, m은 문장 번호이다.) 다시 말해, 문장 S1->S2, S1->S3, .., S2->S3 등의 관계는 긴밀도 분석의 대상이 되지만, 문장 S2->S1, S3->S2, S4->S1 등의 관계는 긴밀도 분석에서 제외한다는 것이다. 왜냐하면, 일반적으로 글은 내용의 흐름에 따라 일관성있게 전개되기 때문이다.

3.3 단락 구분 및 중요 문장 추출

단락의 자동 구분은 문맥의 개념을 중요 문장 추출에 최대한 반영하기 위한 것이다. 사용자가 원하는 요약문의 추출 범위에 따라 단락의 개수가 동적으로 나누어 진다. 예를 들어, 7개 문장으로 구성된 문서에서 30%의 중요 문장을 추출하고자 할 때, 두 개의 문장이 추출되어야 하므로 긴밀도가 가장 느슨한 두 문장 사이에서 단락을 구분하고, 각 단락의 대표 문장을 중요 문장으로 추출한다.

여기서, 각 단락을 대표하는 문장을 선출하여 요약문을 생성하는 방법과 대표 단락을 이용하여 요약문을 생성하는 방법을 고려해 볼 수 있다. 이는 입력되는 문서의 구조와 요약의 정도에 따라 달라질 수 있는데, 정보의 손실을 최소화하고자 한다면 각 단락의 대표 문장을 선출하는 것이 바람직하며, 비교적 한 주제와 관련된 문장들만을 요약으로 생성하고자 한다면 대표 단락을 선출하는 방법이 바람직하다[7]. 본 논문의 경우는 비교적 하나의 주제에 대해 이야기하고 있는 신문 기사를 대상으로 하고 있으므로, 각 단락의 대표 문장을 선출하여 요약문을 생성하도록 함으로써 비슷한 문장들이 중복 추출되는 것을 피하도록 하였다.

먼저 사용자가 원하는 요약문의 추출 범위에 따라 단락을 구분하기 위해서 본 연구에서는 인접한 문장 사이의 긴밀도만을 이용한다. 문서의 내용이 전개되어 나가는 방향에서 가장 낮은 점수를 가진, 즉 내용상으로 가장 느슨한 긴밀도를 가지는 두 문장 사이에서 단락이 1차 분리된다. 계속해서 사용자가 원하는 추출 범위에 따라 2, 3차 분리점이 선택된다.

단락을 동적으로 구분한 후, 각 단락의 대표 문장을 추출한다. 이 때는 인접한 두 문장 사이의 긴밀도뿐만 아니라 모든 문장 간의 긴밀도를 이용한다. 본 시스템은 단락 내에서는 가장 높은 긴밀도를 가지며, 타 단락과는 가장 낮은 긴밀도를 가지는 문장을 선택하여 각 단락의 대표 문장으로 추출한다.

4. 실험

본 연구는 신문 기사를 대상으로 한다. 실험 문서는 매일 경제, 한겨레 신문에서 증권, 금융, 부동산, 정치, 사회·문화 분야별로 발췌하여 준비하였다.

다음은 증권 분야에서 발췌한 문서이다. 요약문 추출 과정을 예로 보이면 다음과 같다.

> 재경부, 국채인수자금 2조원 저리지원
>
> <노영우>
>
> 1. 앞으로 은행 증권 증금 등 국채전문딜러로 책정된 금융기관은 국채를 인수할 때 저리의 국고자금을 지원받을 수 있게 된다.
> 2. 재정경제부는 2일 국채인수·매매 활성화를 위해 국고와 증권금융 등을 통해 2조원의 자금을 조성해 국채전문기관이 3월부터 국채를 인수하거나 유통할 때 지원하기로 했다고 밝혔다.
> 3. 국채인수 자금 지원을 위해 정부는 국고에서 1조원, 증권금융 자체자금 1조원등 총 2조원의 자금을 조성했다.
> 4. 이 자금은 3년이상 만기의 중장기 국채를 인수할 때 국채인수기관이 신청한 범위 내에서 대출 형식으로 전액 지원된다.
> 5. 금리는 시중의 콜금리-1%로 결정되고 만기는 30일 이내다.
> 6. 다만 증권 금융 자금에서 지원되는 경우에는 증권사에 대한 대출금리(현재 연 5.5%)가 적용된다.
> 7. 국채 인수 자금 지원후 남는 자금은 국채 전문 딜러별로 지원 한도를 책정해 별도로 지원하기로 했다.
> 8. 정부는 자금 지원시 은행이 인수한 국채를 담보로 책정한다.

문장 사이에 나타난 재출현 어휘의 양상을 분석하여 계산된 문장 간 긴밀도는 다음과 같다.

	1	2	3	4	5	6	7	8
1		3.4	3.2	1.6	0.0	0.85	2.5	1.4
2			3.5	1.7	0.0	0.85	2.1	1.0
3				1.3	0.0	0.9	1.33	1.3
4					0.2	1.4	1.73	1.0
5						0.7	0.0	0.0
6							1.35	0.5
7								1.4
8								

만약 추출 범위 30%의 요약문을 생성하고자 한다면 2개의 대표 문장을 추출해야 하므로, 문서를 크게 두 개의 단락으로 나눈다. 이 문서의 경우에는 가장 낮은 긴밀도를 보이는 문장 4와 5사이에서 단락이 나누어진다. 그리고 두 단락의 대표 문장으로 문장 2와 5가 각각 선택되어 다음과 같은 최종 요약문을 생성한다.

> 재정경제부는 2일 국채인수·매매 활성화를 위해 국고와 증권금융 등을 통해 2조원의 자금을 조성해 국채전문기관이 3월부터 국채를 인수하거나 유통할 때 지원하기로 했다.
> 금리는 시중의 콜금리-1%로 결정되고 만기는 30일 이내다.

5. 결 론

본 논문에서는 단락의 자동 구분을 통한 중요 문장 추출시스템을 제안한다. 먼저 재출현 어휘의 양상을 어휘의 재출현 여부, 어휘의 일치도, 어휘의 역할 변화로 구분하여 분석하고 이를 문장 사이의 내용이 얼마나 긴밀한가를 파악하기 위한 척도로 이용하였다. 문장 사이의 긴밀도 분석을 통해서 사용자의 원하는 요약문의 추출 범위에 따라 단락을 나누고 각 단락의 대표 문장을 선출함으로써 최종 요약문을 생성하였다.

제안한 방법은 어휘 빈도만을 이용하던 기존의 통계적인 방법보다 질높은 요약문을 생성할 수 있다. 뿐만 아니라 본 연구의 방법론은 도메인에 독립적이라 할 수 있으며, 중요 문장 추출시에 문맥을 고려하므로 요약문 생성에 효과적으로 이용할 수 있다.

현재 연구중인 대용어 처리 부분과 고유 명사 처리에 관한 연구를 결합시킨다면 본 연구에서 제안한 문서 요약기의 성능을 보다 향상시킬 수 있을 것이다.

참고문헌

[1] K.McKeown, J.Robin, and K.Kukich, "Generating Concise Natural Language Summaries", Advances in Automatic Text Summarization, MIT press, pp. 233-264, 1999.

[2] Daniel Marcu, "The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts", Ph.D dissertation, University of Toronto, Canada, 1997.

[3] 정준호, "수사구조를 이용한 문서 요약 시스템", 경북대학교 대학원 컴퓨터공학과 석사학위 논문, 1999.

[4] 한경수, 백대호, 임해창, "질의 확장을 이용한 자동 문서 요약", 정보과학회 봄 학술발표논문집(B) 제 27권 1호, pp. 339-341, 2000.

[5] 류동원, 이종혁, "단어 공기 정보를 이용한 자동화 문서 요약", 정보과학회 봄 학술발표 논문집(B), 제 27권 1호, pp. 345-347, 2000.

[6] 강상배, "한국어 문서의 통계적 정보를 이용한 문서 요약 시스템 구현", 부산대학교 대학원 전자계산학과 석사학위 논문, 1998.

[7] 박혁로, 신중호, 김태희, "검색/요약/필터링을 위한 텍스트 이해 모형 및 처리기술 개발", 연구 개발 정보센터 연구 보고서, 1999.

[8] 김재봉, "텍스트 요약 전략에 대한 국어교육학적 연구", 조선대학교 대학원 국어국문학과 박사학위논문, 1997.

[9] Inderjeet Mani and Mark T.Maybury, "Advances in Automatic Text Summarization", MIT Press, 1999.