다중 문서요약에서 문장의 중복도 측정방법 개선

임정민 ○ 강인수 배재학+ 이종혁
포항공과대학교 전자컴퓨터공학부 컴퓨터공학과, 첨단정보기술 연구센터,
+울산대학교 컴퓨터 정보통신 공학부
{beuuett○, dbaisk, jhlee}@postech.ac.kr, +jhjbae@mail.ulsan.ac.kr

Measuring Improvement of Sentence-Redundancy in Multi-Document Summarization

Jung-Min Lim○ In-Su Kang Jae-Hak 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

다중문서요약에서는 단일문서요약과 달리 문장간의 중복도를 측정하는 방법이 요구된다. 기존에는 중복된 단어의 빈도수를 이용하거나, 구문트리 구조를 이용한 방법이 있으나, 중복도를 측정하는데 도움이 되지 못하는 단어와, 구문분석기 성능에 따라서 중복도 측정에 오류를 발생시킨다. 본 논문은 주절 종속절의 구분, 문장성분, 주절 용언의 의미를 이용하는 문장간 중복도 측정방법을 제안한다. 위의 방법으로 구현된 시스템은 기존의 중복된 단어 빈도수 방식에 비해 정확율에서 56%의 성능 향상이 있었다.

1. 서론

유사한 정보를 갖는 문서들이 많이 생산되면서 문서요약의 대상을 단일문서로만 국한하는 것을 벗어나 유사한 내용을 갖는 여러 문서에 대한 요약이 요구되어졌다. 현재 영어권과 일본에서는 Document Understanding Conference (DUC)와 NTCIR 등을 통해서 단일 문서요약뿐만 아니라, 다중문서(Multi-Document)요약을 Task 에 추가하고 다중문서요약 시스템에 대한 성능평가를 하고 있다.

단일 문서요약은 일반적으로 문서에서 중요한 문장을 추출하는 방법과 추출된 문장을 간결화하고 재구성하는 방법으로 나눌 수 있다. 이에 비해 다중문서요약은 관련된 문서들에서 중요 문장을 추출하기 때문에 중복된 의미를 갖는 문장이 추출될 수 있고, 중복된 문장이 요약에 포함되는 것을 배제하기 위해서 문장간의 중복성을 측정하는 기술이 추가로 요구된다. 중복성을 측정하는데 쉽게 사용되는 방법은 두 문장에 중복되는 단어의 수를 계산하여 문장의 중복여부를 측정하는 것이다. 그러나 이 방법은 문장의 의미를 나타내는 중요 단어들이 문장전체를 이루는 단어들에 비해서 적기 때문에 문장간의 중복도 판별에 오류를 발생시킨다.

본 논문에서는 문장 전체에서 추출된 단어를 이용하지 않고, 문장의 의미를 나타내는 중요 단어를 추출, 가중치를 부여하고, 단어들의 의미코드를 이용해서, 문장간의 중복도 측정방법을 개선하려고 한다.

2. 관련 연구 및 문제점

다중문서요약에서 문장간의 중복도 측정 방법은 가장 기본적인 방법은 앞에서 언급한 중복되는 단어의 수를 계산하여 문장의 중복도를 측정하는 것이다[1], 또한 cosine similarity metric 와 문장이 속한 클러스터간의 비교를 이용해서 중복도 측정을 하는 MMR-MD[2] 방법이 있다. 위의 방법들은 단어수준에서 중복도를 측정하기 때문에 간단하다는 장점이 있다.

하지만 단어의 중복을 이용한 측정 방법은 문장들이 포함하고 있는 불필요한 단어와 문장전체의 의미를 나타내는데 도움이 되지 못하는 많은 단어들의 중복 때문에 잘못된 결과를 초래한다. 예를 들면

예문 1. 30 일 새벽 경기도 화성군 씨랜드 청소년수련원에서 발생한 화재는 수련원생과 인솔 교사들이 대부분 잠든 사이에 발생, 피해가 컸던것으로 알려졌다.
중요단어:새벽,경기도,화성군,씨랜드,청소년수련원,발생, 화재,수련원생,인솔,교사,사이,발생,피해

예문 2. 경기도 화성군 「씨랜드」청소년수련원에서 화재가 발생해 수련회에 참가해 잠을 자고있던 유치원생 23명이 숨지고 3명이 부상하는 참사가 발생했다.
중요단어:경기도,화성군,씨랜드,청소년수련원,화재,발생 수련회,참가,잠,유치원생,부상,참사,발생

위의 예문1은 " 화재가 잠든 사이에 발생 피해가 컸다는 사실" 을 말하고, 예문2는 " 유치원생 23명이 숨지고

3 명이 다쳤다" 라는 것을 말하고 있다. 두 문장간의 의미는 다르지만, 문장의 의미를 분석하는데 도움이 되지 못하는 부분에 중복된 단어가 존재하기 때문에 수식 1 에(k값은 2, $R(S_1,S_2)$의 값이 0.5 이상일 때 유사하다고 판별)의해 두 문장은 중복한 것으로 잘못 판단된다.

$$R(S_1, S_2) = k \times \frac{|S_1 \cap S_2|}{|S_1| + |S_2|}$$

[수식 1] 중복된 단어 수를 이용한 문장의 중복도 측정법

중복된 단어 수에 의존하지 않는 방법으로는 어느 한 문장을 요약에서 배제하는 방법대신에 두 문장을 융합해서 하나의 문장으로 만드는 방법이 있다.[3][4] 이 방법은 각 문장을 구문분석하고, 기능어를 배제하고 핵심어만으로 의존구조(Dependency Structure)를 형성한다. 이후 문장간의 중복도 측정시, 의존구조를 이루는 각각의 노드들이 일치하는 경우 두 문장을 융합한다. 이 방법은 기존의 단어 일치 방법을 이용하지 않고 구문분석의 결과인 트리구조를 이용하여 단어 일치 방법이 갖는 문제점을 해결하였다.

그러나 이 방법은 구문분석기의 성능에 좌우될 수 있는데, 현재 영어의 경우 89.5%(100 단어 이하)[5]이고 한국어의 경우 89%[6]의 성능을 보이고 있어서, 구문분석의 결과를 전적으로 문서요약에 적용하기에 어려움이 있다. 또한 한국어의 구문분석은 복문처럼 문장의 길이가 길어질수록, 구문분석에서 의존관계에 많은 오류가 존재하고, 미지격 문제[7]가 있다.

3. 제안한 문장 중복도 측정방법

현재 한국어 다중문서요약에서 문장의 중복도 측정에 의존구조를 이용한 방법이나 중복된 단어의 빈도수를 이용한 방법은 위에서 언급했듯이 문제점이 있다. 특히 한국어 구문분석은 의존관계를 나타내는데 오류가 많아서 의존구조를 이용한 방법은 적절하지 못하다. 하지만 의존관계를 제외한 구문분석 정보는 문장의 중복도 측정에 충분히 이용할 수 있다. 따라서 중복된 단어의 빈도수를 이용한 방법에 부분적인 구문분석 정보를 이용한 방법을 제안한다.

문장의 중복도 측정을 위해서는 문장전체에서 단어를 추출하기 보다, 문장의 의미를 갖고 있는 부분에서 단어를 추출하는 것이 바람직하다. 복문의 경우, 문장을 구성하는 모든 부분이 전체 문장의 의미를 나타내는데 도움을 주지 않고, 주절의 내용이 전체 문장의 의미를 주로 드러낸다. 예를 들어 "**북상중인 제 7 호 태풍 '올가'(OLGA)가 3 일 낮 한반도 내륙을 정면으로 관통할 것으로 보여 전국적으로 큰 피해가 예상된다.** 라는 문장에서 종속절은 " 북상중인 제 7 호 태풍 '올가'(OLGA)가 3 일 낮 한반도 내륙을 정면으로 관통할 것으로 보여" 이고 주절은 " 전국적으로 비 피해가 예상된다." 이다. 위의 예문은 주절만을 읽어보아도 전체문장의 의미를 파악할 수가 있다. 즉 의미분석에는 주절의 역할이 크다고 할 수 있다.

또한, 문장의 중복도 판별에 있어서 주절의 동사는 중요한 역할을 한다. 문장을 구성하는 다른 단어들과 달리 같은 의미를 나타내면서도 다양한 형태로 문장에서 존재할 수 있기 때문이다. " 큰 피해가 예상된다" 라는 문장은 " 피해는 더욱 커질 전망이다" 이라는 문장과 비슷한 의미를 같고 동사들 또한 유사한 의미를 내포하지만 " 예상" 과 " 전망" 이라는 다른 형태로 문장에서 존재할 수 있다. 그러므로 문장간의 동사를 비교할 때는 단어 형태뿐만 아니라 의미까지도 비교를 하는 것이 바람직하다.

따라서 전체문장에서 주절의 의미적 특성과 주절에 있는 동사의 형태적 특성을 이용하면 문장전체에서 중요 단어 추출 및 문장간의 중복도 측정을 개선시킬 수 있어 기존의 단어 중복을 기반으로 한 중복도 측정방법의 단점을 보완 할 수 있다. 이 방법은 주절을 찾기 위해서 구문분석의 정보를 이용하지만, 제한적인 정보만을 이용하기 때문에 구문분석 전체에서 발생되는 오류에 영향을 덜 받는다. 이를 바탕으로 기존의 수식을 개선하면 다음과 같다.

$$R(S_1, S_2) = \frac{k \times \sum_{(t_i,t_j)\in S_1 \overset{*}{\cap} S_2} (W(t_i,S_1) + W(t_j,S_2)) / 2}{\sum_{t\in S_1} W(t,S_1) + \sum_{t\in S_2} W(t,S_2)}$$

$$S = \{t_1, \cdots, t_n\}, \quad \overset{*}{\cap} = \overset{L}{\cap} + \overset{C}{\cap}$$

$$W(t,S) = W_{st}(t,S) + W_{gr}(t,S)$$

$$S_1 \overset{L}{\cap} S_2 = \{(t_i,t_j) \mid t_i \in S_1, t_j \in S_2, t_i = t_j\}$$

$$S_1 \overset{C}{\cap} S_2 = \{(t_i,t_j) \mid t_i \in S_1, t_j \in S_2, C(t_i) \cap C(t_j) \neq \phi\}$$

$C(t) = t$ 의 의미코드 집합

[수식 2] 개선된 문장 중복도 계산 식

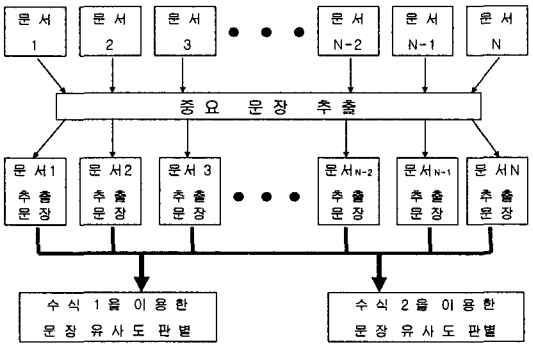
S 는 문장이고, t 는 문장에 속한 단어이다. $W_{st}(t,S)$는 문장 S 에서 t 가 주절에 있는지 종속절에 있는지에 따라서 다른 값을 갖는다. $W_{gr}(t,S)$는 문장 S 에서 t 의 문법적 기능에 따라서 값을 갖는다. $S_1 \overset{L}{\cap} S_2$ 는 단어가 같을 때, $S_1 \overset{C}{\cap} S_2$ 는 단어의 의미코드가 같을 때 t_i와 t_j 을 찾는다.

4. 실험 시스템 및 결과 분석

실험에 사용한 다중문서요약 시스템은 문장추출을 기반으로 하였다. 문장별 점수를 측정하는 방법으로는 실험 대상 문서인 신문기사의 특성에 맞추어서 Lead method 방식을 이용하되, 기사의 끝부분에도 가중치를 부여하는 방식을 채택하였고, 너무 긴 문장과 짧은 문장에 대해서는 감점을 주었다. 또한 인용문을 포함하는 문장은, 요약에서 불일치를 유발할 가능성이 있기 때문에 따옴표를 갖고 있는 문장에 대해서는 감점을 주었다.[8] 제목에 있는 단어들에 대해서는 가중치를 부여해서 해당 단어를 포함하는 문장에 대해서는 점수를 더해 주었다.

문서집합은 1999년 한국일보 신문기사를 사용하였다. '태풍 올가' '씨랜드 참사' 등 4개의 주제별로 10개 18개 사이의 관련 기사를 문장 비교대상의 집합으로 설정하였다. 각 신문기사는 전처리 과정을 통해서 문장 단위로 나누었다.

유사한 문장에 대한 비교를 하기 위해서 각 신문기사별로 일정량(전체문장의 40%)의 문장을 추출하였다. 추출된 문장들에 대해서 수식 1을 사용한 방법과, 수식 2를 사용한 방법을 이용해서 문장간의 중복도 판별을 하였다.



[그림 1] 실험에 사용한 다중문서요약 시스템

수식 1과 2에서 k를 2로 설정하고, R(S1,S2)를 0.50이상일 경우 유사하다고 설정하여 중복된 단어의 수가 전체 단어수의 절반을 넘을 경우 중복된 것으로 간주하였다. 수식 2에서는 단어가 주절에 위치한 경우 0.9, 종속절에 위치한 경우 0.1의 값을 주었고, 단어의 문법적 기능이 '주어' 일 경우 0.7의 가중치를 주었다. 또한 '미등록어' 로 처리되는 단어에 대해서도 중요하다고 판단하여 0.7의 가중치를 부여하였다. 단어 중복의 판별에서 의미코드의 사용은 동사로만 제한을 하였다. 사용된 의미코드는 Kadokawa 시소러스 의미코드를 기반으로 구축된 사전을 이용하였고, Single-ilnk 방법으로 문장집합을 형성하였다.

	시스템이 생성한 총집합 수	정답을 포함하는 집합 수
수식 1	13	6
수식 2	10	7

[표 1] 중복 문장집합 형성 실험 결과

	Precision	Recall
수식 1	53.3%	39.3%
수식 2	83.3%	47.7%
Improvement	56%	21%

[표 2] 신문기사 주제별 수식1과 수식2의 성능 비교

실험은 시스템에서 추출된 문장 중에서 2개 이상의 유사한 문장을 포함하는 정답집합을 사람이 만들고, 시스템에서 생성한 문장집합과 비교하였다. 사람이 만든 정답집합의 수는 총 15개 였고, 한 개의 집합 당 평균 2.9개의 문장을 포함했다. 시스템이 생성한 문장집합과 정답집합의 비교시에는 2개 이상의 문장이 같을 경우 정답으로 판단하였다.

[표 2]를 통해서 문장간의 중복도 판별에는 문장에 나타나는 모든 단어보다 문장의 의미를 나타내는 중요 단어위주로 중복도 구별을 하는 것이 좋다는 것을 알 수 있다. 하지만 Precision 에 비해 Recall 은 상대적으로 낮았다.

5. 결론 및 향후 연구

다중문서요약에서 요구되는 문장의 중복성 측정은, 기존의 중복된 단어의 수를 이용하는 방법에 구문분석 결과를 제한적으로 이용하면 성능을 향상 시킬 수 있었다.

현재 실험에 사용한 다중문서요약시스템은 중요 문장 추출방법에서 단일문서요약에 사용된 방법을 변형하지 않고 적용했다. 하지만 단일문서요약에서 사용되는 문장 추출 방법은 관련된 여러 문서들의 특성을 반영하지 않기 때문에, 다중문서요약에 적용을 위해서 반드시 개선할 필요가 있다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았습니다.

6. 참고 문헌

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

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

[3] R. Barzilay, K. R. McKeown, M. Elhadad, Information Fusion in the Context of Multi-Document Summarization, ACL, 1999

[4] K. R. McKewon, J.L. Klavans, V. Hatzivassiloglou, R. Bazilay, E. Eskin, Towards Multidocument Summarization by Reformulation, AAAI, 1999

[5] E. Charniak. A maximum entropy-inspired parser, Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 132-139, 2000

[6] M. Y. Kim and J. H. Lee, S-clause Segmentation for Efficient Syntactic Analysis of Long Sentences, Proceedings of the 20th International Conference on Computer Processing of Oriental Languages (ICCPOL), p.147-153, 2003

[7] 이용훈, 김미영, 이종혁, 대등접속구문과 미지격 명사구의 문법기능 결정, 한국정보과학회 논문지 제30권 제1호, 2003

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