VNA 집합을 이용한 뉴스기사의 중요문장 추출

나종열[0], 신지애, 최기선
한국과학기술원 전자전산학과 전산학전공
시맨틱웹첨단연구센터

Unsupervised News Article Summarization Using VNA Sets

Division of Computer Science, Semantic Web Research Center, KAIST

요 약

본 연구에서는 문서의 문장들을 순위화하여 추출하는 일반적인 문서 요약 방법론을 소개한다. 첫 번째 단계는 주제와 관련되는 동사, 명사, 형용사(VNA) 단어들의 집합을 구하여 각 문장의 주제 관련성 정도를 결정하며, 두 번째 단계는 단어들의 의존관계를 통해 각 문장의 정보 함유량을 판단한다. 두 개의 방법은 모두 주제와 관련된 정보를 많이 내포하는 문장에 중요도를 부여하고 있다. 이러한 방법은 주제와 연관성이 높고 정보전달성이 높은 문서요약을 만들기 위함이다. 생성된 문서요약본의 성능평가는 문서요약의 결과로 추출된 문장들과 설문에 의해 추출된 문장들의 일치율에 의해 시행되었으며 68%의 일치율을 보였다.

1. 서 론

본 연구에서는 동사-명사-형용사의 묶음인 VNA 집합을 이용하여 뉴스기사를 요약하는 방법을 소개한다. VNA 집합은 문서 주제에 대한 단어들의 삼항식(triplet) 표현이며 KAIST SWRC에서 처음 사용되었다[1]. 이 방법의 장점은 3개 이하의 단어로 이루어진 VNA 집합의 구성으로 뉴스기사를 효율적으로 표현할 수 있으며 자동적으로 만들어질 수 있다. 뉴스기사 요약의 기본 아이디어는 뉴스제목에서 생성된 VNA 집합들을 특정 알고리즘을 이용하여 중요도에 따라 점수를 부여하고 각 문장에 존재하는 VNA 집합에 따라 문장에 점수를 부여한다. 최종적으로 문장들에 부여된 점수의 평균값 이상의 문장들을 추출하여 뉴스기사의 최종 요약본이 만들어진다. 다음 절에서는 관련연구와 본 연구를 비교하겠으며 3절에서는 VNA 집합을 생성하기 위한 의존문법 트리에 대한 인접행렬 표현을 소개한다. 4절에서는 VNA집합들을 생성하는 방법 및 문장 점수부여 방법 그리고 문서요약 방법을 설명한다. 5절에서는 본 연구의 문서요약 성능평가 결과를 보이며 결론을 6절에서 다루겠다.

2. 관련 연구

Goldstein[2]은 각 문장이 문서 요약에 관련이 있는지 판별하기 위하여 통계적 특성(statistical features)와 언어학적 특성(linguistic features)에 근거한 방법을 제시하였다. 통계적 요소에는 코사인 유사도, TF-IDF 가중치 및 문장의 위치 등이 있고 언어학적 요소에는 인용문, 경어적 표현, 품사 등이 있다. 이와 같은 요소들을 정량화 시켜서 다음 수식을 이용하여 문장 벡터(sentence vector)를 구하였다.

$$Score(\mathcal{S}_i) = \lambda \sum_{s \in \mathcal{S}} \omega_s * (Q_s \cdot \mathcal{S}_i) + (1 - \lambda) * \sum_{\ell \in L} \omega_\ell * (L_\ell \cdot \mathcal{S}_i)$$

여기서 S_i는 i번째 문장의 문장벡터, Q_s는 통계적 요소 집합을 위한 쿼리 벡터 그리고 L_i는 언어학적 요소 벡터이다. w_s와 지은 쿼리의 종류에 의해 정해진다. λ (lambda) 값은 통계적 요소와 언어학적 요소의 가중치 조절에 따라 정해진다.

Hovy[3]는 의미의 다양성이 제한되는 단어들의 쌍(전치사, 동사 또는 명사)을 이용한 방법을 제시 하였다. 도메인에 따라 다른 의미를 가질 수 있는 단어단위로 일치하는 빈도를 측정하기보다는 의미의 폭이 제한이 되는 최소의미단위(minimal semantic units)의 기본요소 단위를 이용해 일치 빈도를 구함으로서 요약문서에 들어갈 문장의 유효성을 더 정확하게 측정할 수 있다고 했다.

Metzler[4]는 의존관계에 있는 명사들이나 형용사로 수식이 되는 명사 등의 독립적인 명사구 단위들을 사용하여 문장의 단위를 나누는 방법을 제시하였다. 예를 들어, 'fire engine'에 대한 문장은 'engine'으로 표현하기에는 너무 일반적이며 'skyscrapers in Seattle'에 대한 문장에서 'Seattle' 또는 'skyscrapers'를 따로 두는 것은 의미가 없다. 이를 극복하기 위해 의존관계에 있는 명사들이나 형용사로 수식이 되는 명사구 또는 두개 이상의 명사가 의존하는 동사 중심으로 문장의 단위를 나누는 방법을 제시하였다.

본 연구에서는 의존문법으로 분석된 제목의 구문트리를 인접행렬(adjacency matrix)을 이용하여 단어 집합의 주제와의 연관성을 판단한다. 또한 행렬의 정보로부터 4W(when, where, who, what)와 관련된 의존성을 띄는 단어가 연관성 판단에 반영한다.

3. 의존문법 트리에 대한 인접행렬 표현

본 연구에서는 기존의 의존문법 분석이 트리의 너비우선탐색(BFS)으로 되어 온 것을 시간적으로 개선하는 방법을 제시한다. 의존문법에 의하여 방향성 트리 그래프(digraph tree)로 표현된 문장의 구조를 인접 행렬(adjacency matrix)로 표현함으로써 단어간 의존관계를 직관적이고 실시간으로 파악할 수 있다는 이점이 있다.

3.1. 의존문법

의존문법 구조의 적격성(well-formedness)를 판단하기 위해 Robinson[5]이 제시한 네 가지 공리는 다음과 같다.

- 한 문장 안에서 오직 한 단어(루트)만 독립적이다.
- 하나의 문장 안에서 한 단어(루트)를 제외한 모든 단어는 다른 단어와 의존관계에 있다.
- 하나 이상의 단어와 의존관계에 있는 단어는 없다.
- 단어 A가 단어 B와 의존관계에 있으면 A와 B 사이에 있는 단어 C는 A와 B 사이의 단어에 의존하게 된다.

이 공리들을 단어의 집합 W={w_1, w_2, …, w_{k-1}, w_k}와 이진관계 R로 재정의를 하면,

- $R \subset W \times W$
- $\forall w_1 w_2 \cdots w_{k-1} w_k \in W: \langle w_1, w_2 \rangle \cdots \langle w_{k-1}, w_k \rangle \in R: w_1 \neq w_2$ (Acyclicity)
- $\exists! w_1 \in W: \forall w_2 \in W: \langle w_1, w_2 \rangle \notin R$ (Rootedness)
- $\forall w_1 w_2 w_3 \in W: \langle w_1, w_2 \rangle \in R \wedge \langle w_1, w_3 \rangle \in R \rightarrow w_2 = w_3$ (Single-headedness)

따라서 acyclic, rootedness, single-headedness의 특성들을 통해 의존문법 파싱트리는 방향성을 갖는다.

3.2. 의존문법 트리의 인접행렬 표현

하나의 문장을 의존문법으로 파싱을 하면 각 단어의 토큰번호, 원형, 기본(base)형, 의존함수 및 형태소 정보가 나온다. 그림1(a)에서 박스 안의 의존함수들을 근거로 파싱트리의 인접행렬을 구한 그림1(b)에 표현된 내용은 다음과 같다.

- $A := (a_{i,j})_{n \times n}$, n 은 제목의 토큰 갯수
- $a_{i,j} = k$, 토큰 i가 토큰 j에게 k의 의존성을 갖음
- 토큰 m이 root이면 $a_{m,m} = *$
- 토큰 m이 문장부호 이면 $a_{m,m}$ = 문장부호

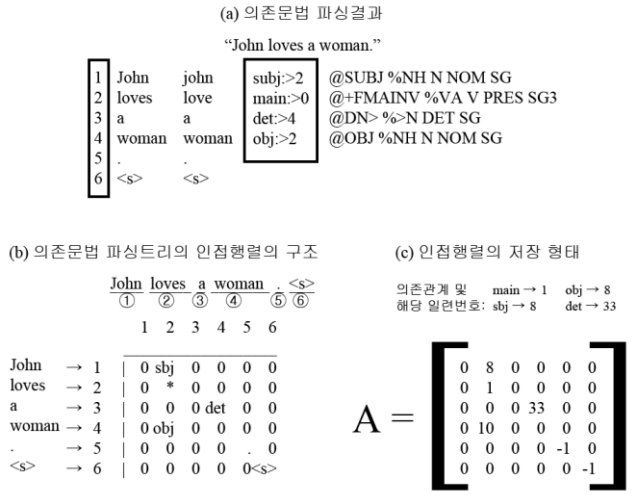

그림 1. 의존문법 트리의 인접행렬 표현 (a) 각 단어의 토큰번호, 원형, 의존함수, 형태소 정보, (b) 의존함수에 의해 생성된 인접행렬(adjacency matrix), (c) 실제 시스템에 저장되는 행렬

행렬의 가로와 세로의 색인은 각 단어의 토큰 값이고 각 행렬의 (i,j) 위치에 있는 행렬값 n는 "i는 j와 n의 관계"라는 뜻이다. 예를 들어, 문장의 첫 토큰인 'John'은 'loves'에 subject의 관계로 의존하기 때문에 그림1(b)의 행렬 (1,2)위치에 'sbj'값이 부여된다. 의존 함수에서 main으로 지정된 단어가 트리의 루트이며 또한 이 문장에서 가장 중요한 동사 (토큰 2번 'loves')이다. 루트는 '*'로 인접행렬에 표시되고 '마침표와 문장의 끝을 알리는 <s>태그는 그대로 표시를 한다.

그림(b)의 행렬표현은 본 논문에서 설명을 위해 쓰이고 실제 시스템에서 인접행렬의 원소들은 그림1(c)와 같이 루트는 1, 의존관계는 해당 일련번호 그리고 문장부호는 -1로 표시된다.

4. 뉴스기사 중요 문장 추출

각 문장의 주제와의 관련성의 정도의 결정은 주제와 관련되는 단어들의 집합을 구하는 VNA 집합 생성 단계와 문장들의 4W[1] 정보 함유량을 판단하는 두 단계를 거친다.

4.1 VNA 집합들의 생성

VNA 집합은 동사, 명사 및 형용사로 이루어진 집합이며 하나의 뉴스기사에 존재하는 여러 VNA 집합들은 중요도에 따라 레벨이 나뉘어 진다.

"Samsung produces world's fastest DRAM"

위의 제목의 핵심 주제는 '삼성 DRAM 개발'이다. 개발을 한 것은 '삼성의 빠른 DRAM'이며 그 빠르기는 '세계에서 가장 빠른' 이다. 이와 동일한 과정으로 VNA 집합들이 만들어진다.

[1] 4W: Who, When, Where, What

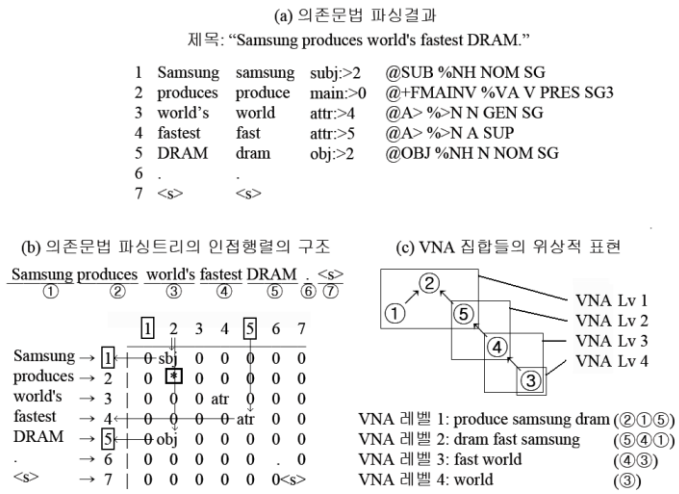

그림 2. VNA 집합 생성과정 (a) 각 단어의 토큰번호, 원형, 의존함수, 형태소 정보, (b) 의존함수에 의해 생성된 인접행렬(adjacency matrix), (c) VNA 집합들의 위상적 표현

VNA집합을 구하기 위해 뉴스 제목을 의존문법 파서로 분석하여 얻은 의존함수들을 근거로 인접행렬을 생성한다. 그림 2(a)는 제목의 파싱 결과이며 이를 통해 의존함수들이 인접행렬로 표현된다 (그림 2(b)). 이 제목의 VNA 집합들을 구하는 과정은 다음과 같다.

- 파싱 결과의 의존함수 중 main으로 지정된 단어가 루트이므로 ‘produces’가 루트가 되며 인접행렬에서 ‘*’로 표시된다.
- 루트로 지정된 단어(그림 2(b)의 토큰번호②)는 제목문장의 중심 동사이므로 루트에 의존하는 단어들과 함께 핵심주제를 표현한다.
- 그림 2(b)의 인접행렬에서 루트의 토큰번호 ②번 열에 의존관계가 존재하는 단어(토큰번호 ①, ⑤)들을 찾으면 루트와 함께 VNA Level 1이 되며 제목의 주요 내용을 나타낸다 (‘①Samsung ②produces ⑤DRAM’).
- 뉴스기사의 모든 내용이 핵심주제만 다루는 것이 아니므로 핵심주제에 관련된 내용을 찾기 위해 루트에 의존하는 단어들에게 의존하는 단어들을 찾는다.
- 핵심주제와 관련된 내용에 대한 단어들을 찾기 위해 ①번 열과 ⑤번 열에 의존관계가 존재하는 단어들의 토큰번호를 찾는다.
- 제 1열은 행렬의 원소값이 모두 0이므로 토큰번호 ①번(‘Samsung’)에 의존하는 단어는 없다.
- 제 5열에서 의존관계가 존재하는 토큰번호 ④번를 찾음으로써 토큰번호 ①번 및 ⑤번과 함께 VNA Level 2를 구성한다 (‘①Samsung ④fastest ⑤DRAM’).
- 다시 토큰번호 ④부터 의존관계가 끝날 때까지 재귀적으로 주제와 관련된 단어들을 찾으면 모든 VNA집합들을 구하게 된다. (그림 2(c))

따라서 다음과 같은 VNA 집합들을 구하게 된다.

VNA 레벨 1: “produce samsung dram”
VNA 레벨 2: “dram fast samsung”
VNA 레벨 3: “fast world”
VNA 레벨 4: “world”

VNA 집합을 구성하는 단어들은 제목에 나타나는 단어들의 원형(base form)이므로 시제 및 인칭에 종속되지 않는다. VNA 집합을 모두 구한 다음 각 문장마다 주제 관련성 가중치를 부여한다. VNA Level 1의 단어들이 모두 들어가 있는 문장에 가장 높은 점수를 부여하고 레벨 번호가 증가하면서 점수를 낮추어 가중치를 부여하면 모든 문장의 중요도를 결정 할 수 있다.

이로써 뉴스의 주제에 충실한 문장은 복수의 VNA 집합이 포함되므로 높은 점수를 받게 되고 주제에 관련된 문장들까지 점수를 부여하면서 뉴스 요약에 문장들을 추출 할 수 있게 된다.

4.2 4W 정보의 추출

뉴스본문의 문장에서 when, where, why, who와 관련된 의존형태를 나타내는 단어들을 찾아서 해당 점수를 문장의 주제 연관성의 판단에 반영한다. 뉴스는 시간, 장소, 인물, 사건을 중심으로 어떻게 또는 왜 일어났는지에 대한 설명을 하고 있다. 그러므로 시간, 장소, 인물, 사건에 대한 단어가 많은 문장이 정보 응집력이 크다.

다음은 어느 뉴스 기사 중 한 부분이다.

“Park plans to complete the development of stem-cell technologies by 2011, with the aim of offering therapeutic stem cells from 2012.”

Who에 관련된 단어는 Park, what에 관련된 단어는 complete, development 및 stem-cell 그리고 when에 관련된 단어는 2011, 2012이다. 위 문장을 의존문법으로 파싱하여 돌아온 결과는 그림 3(a)와 같다. 결과에서 위 단어들의 의존관계를 살펴보면 who에 해당하는 단어들은 subj, what에 해당되는 단어들은 obj 그리고 when에 해당되는 단어들은 pcomp의 의존관계를 보임을 알 수 있다.

Korean Herald 1023개 기사의 15432개의 문장에 대한 실험결과 위 현상은 다른 문장에서도 보이는 것을 확인하였다. 실험결과 발견된 현상은 그림 3(b)와 같다.

한 단어가 전치사구(prepositional complement)의 의존성을 보이면 주체에 대한 보완이 대부분이므로 when, where의 정보가 따를 확률이 높아진다. 주어(subject)의 의존성을 보이면 동사에 연관된 주어이므로 who의 정보를, 목적어(object)의 의존성을 보이는 단어는 타동사와 연관이 있으므로 what의 내용을 담고 있다.

(a) 단어간 의존관계

"Park plans to complete the development of stem-cell technologies by 2011, with the aim of offering therapeutic stem cells from 2012."

1	Park	subj:>2	13	with	ha:>4
2	plans	main:>0	14	the	det:>15
3	to	pm:>4	15	aim	pcomp:>13
4	complete	obj:>2	16	of	mod:>15
5	the	det:>6	17	offering	pcomp:>16
6	development	obj:>4	18	therapeutic	attr:>19
7	of	mod:>6	19	stem	attr:>20
8	stem-cell	attr:>9	20	cells	obj:>17
9	technologies	pcomp:>7	21	from	sou:>17
10	by	tmp:>4	22	2012	pcomp:>21
11	2011	pcomp:>10	23	.	
12	,		24	<p>	

(b) 4W 형태별 의존관계

- Who: subject의 의존성을 보이는 단어
  1번 토큰: Park
- What: object의 의존성을 보이는 단어
  4번 토큰: complete 6번 토큰: development
  20번 토큰: cells
- When / Where: prepositional complement의 의존성을 보이는 단어
  9번 토큰: technologies 11번 토큰: 2011
  15번 토큰: aim 22번 토큰: 2012

그림 3 4W 정보의 특성 (a) 뉴스 기사 내용 파싱 결과, (b) 4W 형태별 의존관계

따라서 위의 최소핵심의미 단위를 이용한 문장 가중치 부여와 4W 정보 추출에 의한 문장 가중치 부여를 통해 뉴스 기사의 중요 문장이 추출 가능해진다.

5. 실험결과 및 분석

총 73개의 Korean Herald 기사를 무작위로 11~12개씩 32명에게 제공하고 제목과 관련이 높은 문장을 선택하도록 설문조사를 실시했다. 각 뉴스 기사는 6~10개의 문장으로 되어 있었고 문장을 2~4개 선택 하도록 하였다. 설문지에 제공된 기사마다 설문참여자가 선택한 각 문장들의 분산을 다음과 같이 구했다.

$$\sigma_m^2 = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}(x_{mi} - \bar{x}_m)^2 \text{ , m은 기사 번호; i는 문장 번호}$$

설문조사 결과 각 기사마다 문장의 선택 빈도수에 대한 분산이 각 기사의 평균 빈도수 이상이 되는 기사들을 선택하여 VNA 집합에 의한 선택문장들을 비교한 결과 평균 68%의 일치도를 보였다. 임의의 기사에 대해 설문에 의해 선택된 문장들중 평균 이상인 문장들과 VNA 집합에 의해 추출된 문장들에 대한 백분율에 의해 평가를 하였다.

$$\text{추출 문장 일치도}(\%) = \frac{\mathrm{H}_{sur}}{\mathrm{H}_{vna}}$$

관측 결과 VNA 집합에 의한 문장선택 방법은 사람의 중요 문장 추출 방법과 유사한 메커니즘을 보인다. 발견된 취약점은 VNA집합 단어들이 동의어나 축약어 및 머리글자에 의한 단어를 인식 못하기 때문에 특정 뉴스 기사에서는 낮은 성능을 보였다. 향후 워드넷을 이용한 VNA 집합들의 동의어와 및 약어를 찾아냄으로써 VNA 집합을 확장시키면 더욱 나은 결과를 얻을 수 있을 것이다.

6. 결 론

본 논문에서는 각 문장의 주제와의 관련성의 정도의 결정을 위해 VNA 집합 생성 및 각 문장 일치 여부 판단 단계와 문장들의 4W 정보 함유량을 판단하는 두 단계를 거친다. 이 방법의 특징 및 이점은 다음과 같다

- 도메인에 따른 종속 없이 일반적인 뉴스 기사 요약이 가능하다.
- 주제를 중심으로 관련이 높은 문장들을 찾기 때문에 사람의 문서요약 방법과 비슷하다.
- 기존의 문법트리 탐색보다는 트리에 대한 인접행렬 분석을 통하여 시간적으로 개선된 방법을 제시한다.

References

[1] 최주원, 나종열, 최동현, 여명숙, 신지애, 최기선, "*ONASys: 온라인 뉴스기사의 재구성*", Semantic Web Research Center, KAIST, pp. 3-4, 2007.

[2] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonell, "*Summarizing Text Documents: Sentence Selection and Evaluation Metrics*", Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 121-128, 1999.

[3] Eduard Hovy, Chin-Yew Lin, Liang Zhou, "*Evaluating DUC 2005 using Basic Elements*", In Proceedings of the Fifth Document Understanding Conference (DUC '05) , 2005.

[4] D. P. Metzler, T. Noreault, L. Richey, B. Heidorn, "*Dependency Parsing for Information Retrieval*", Proc. of the third joint BCS and ACM symposium on Research and development in information retrieval, pp. 313-324, 1984.

[5] Ralph Debusmann, "*An Introduction to Dependency Grammar*", Hausarbeit fur das Hauptseminar SoSe 99. Univeristat des Saarlandes, 2000.