신문기사의 감정추출 방법에 관한 연구

백 선 경° 김 판 구
조선대학교 전자계산학과
{zamilla100°, pkkim}@chosun.ac.kr

A Study on Method for Extraction of Emotion in Newspaper

Sunkyoung Baek° , Pankoo Kim
Dept. of Computer Science, Chosun University

요 약

정보검색에서의 사용자의 다양한 질의어는 객관적인 키워드에서 인간이 주관적으로 생각하고 느끼는 감정요소를 동반한 어휘들로 범위가 넓어지고 있다. 이에 본 논문에서는 감정에 기반한 신문기사 검색을 위하여 기사의 구문 분석 및 품사 태깅 절차를 거쳐 동사를 추출하고 그 중 감정을 내포하는 동사들의 관계를 이용하여 신문기사의 감정을 추출한다. 감정동사의 관계를 참조하기 위하여 감정동사들을 OWL/RDF(S)를 이용해서 온톨로지를 구축하였고 에지(Edge)기반의 유사도 측정방법을 제안하였다. 제안한 방법은 여러 가지 감정을 추출하고 감정 정도를 측정할 수 있기 때문에 이는 향후 감정기반 신문기사 검색에 효과적으로 사용될 수 있을 것이다.

1. 서 론

컴퓨터는 미래의 인간 시대를 더욱 윤택하고 편리한 도구로 자리매김하면서 컴퓨터 과학 분야의 연구들은 인간 생활과 아주 밀접하게 거리를 좁혀오고 있다. 현재 우리는 감성 정보화 시대를 살고 있고, 인간의 주관적이고 애매모호한 감성들의 정보를 컴퓨팅할 수 있는 연구들이 활발히 진행되고 있다. 또한 인간이 컴퓨터의 기능이나 형식의 제한에 구애받지 않고, 컴퓨터 존재의 인식 없이 컴퓨터로부터 유익한 서비스를 받을 수 있는 유비쿼터스 컴퓨팅 시대가 도래했다[1]. 그러므로 인간의 느낌이나 인상, 그리고 감성이나 감정에 대한 정보들은 좋은 입력정보가 되며 프로세서 부분에서도 당연히 인간의 주관적인 생각과 느낌을 고려하여 처리할 수 있어야 한다는 것이다.

인간의 감성을 시스템에 효과적으로 전달하여 적용하고 개개인의 사용자가 원하는 정확한 정보를 만들 수 있는 감성 정보에 대한 연구 중 하나로 본 논문에서 우리는 감성에 포함된 인간의 감정을 사용하여 신문기사를 검색하고자 신문기사의 감정추출 방법을 제안하고자 한다.

기존의 정보검색에서의 사용자의 다양한 질의에 대한 처리들은 객관적인 키워드를 찾거나 사용자의 질의어를 포함하고 있는 단순 문서를 사용자에게 돌려주는 검색 방법들이 대부분이다. 하지만 인간의 주관적인 생각이나 감정요소를 동반한 질의어를 받았을 때는 의미적으로 감정에 적합한 정보를 검색할 수 있는 방법이 필요하다.

이러한 감정 기반의 신문기사 검색을 위하여 기사의 구문 분석 및 품사 태깅 절차를 거쳐 동사를 추출하고 추출된 동사들 중 감정을 표현하는 동사들의 관계를 이용하였다. 감정동사의 관계를 참조하기 위하여 본 논문에서는 인간의 여러 가지의 감정들 중 대표적인 10가지(Happiness, Sadness, Fear, Anger, Desire, Calm, Craze, Disgust, Attraction, Despair)의 감정을 최상위 클래스로 갖는 감정동사들을 OWL/RDF(S)를 이용해서 온톨로지를 구축하였고 개념간의 거리 측정을 위하여 의미적 유사성을 평가하는 에지(Edge)기반의 유사도 측정방법을 제안하였다. 그리고 제안된 방법을 신문기사에 적용하여 감정을 추출하고 감정 정도를 측정함으로써 감정기반 검색에 유리하게 적용될 수 있음을 확인하였다.

본 논문의 구성은 다음과 같다.

2장에서는 기존의 클러스터링 방법을 이용한 감정 정보 분류 방법을 소개하고 3장에서는 본 논문에서 제안하고자하는 감정 추출 방법에 대하여 전체적인 시스템 구성을 전개한 후 이에 따른 신문기사의 감정추출과 감정 정도 측정의 단계별 과정과 결과를 제시한다. 마지막으로 4장에서는 결론 및 향후 연구에 대하여 논하였다.

2. 관련 연구

감정정보를 기반으로 한 연구는 과거의 인공지능 분야에서 가장 많이 다뤄졌으며 대표적인 시스템에는 '공포', '분노', '불신' 세가지 감정요소를 지원하는 정신병 환자와의 대화를 목적으로 설계된 Colby의 PARRY, 유명한 인공지능 사례인 에피소드를 인식하는 심층인식 소프트웨어였던 Dyer의 BORIS, 그리고 가상현실과 대화형 소설을 작성하기 위해 OCC(Ortony, Collins, Clore)들의 모델을 기반으로 하는 시스템인 CMU의 Oz, Wright의 감정에이전트 등이 있다[3][4][5]. 앞에서 제시한 시스템들은 주로 추론과 문장 이해를 목표로 구현되어졌다.

최근 감정요소를 정보검색에 사용하는 비슷한 연구 중 김명관의 감정요소추출시스템이 있다[6]. 이것은 감정에 기반한 정보검색을 수행하기 위해 감정시소러스를 구성하였고 기본 5가지 감정요소를 해당 문서에서 추출하여 문서를 벡터화시키고 이를 K-NN기법, 단순베이지안 및 상관계수기법을 사용한 2단계 투표방식을 통해 학습하고

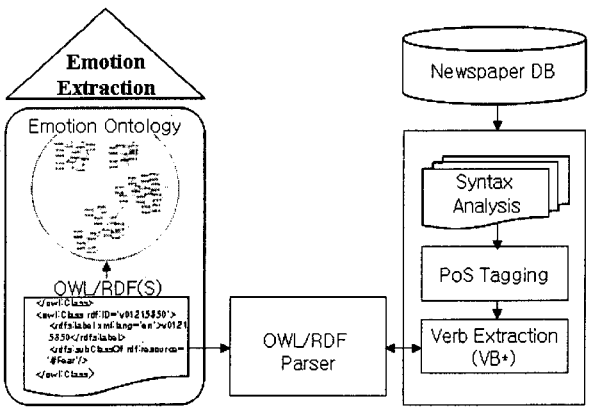
분류하는 연구이다. 이는 외부에서 들어오는 Web 및 각종 문서들은 감정요소 추출기에서 5가지 감정요소가 추출되어 해당 문서의 감정백터를 구성한다. 감정 요소를 추출하는 과정은 해당 문서의 형태소 분석을 통하여 감정시소러스에 등록된 단어들의 감정요소들의 수를 누적하여 그 빈도수에 따라 키워드를 얻어낸다. 이렇게 얻어진 감정 요소 벡터화일과 키워드 자동색인파일에 의해 1차와 2차 검색 과정을 사용한다. 이는 감정시소러스에 등록된 단어들을 이용하여 감정요소를 추출해 내는 방법이다.

이 연구의 감정성분 추출방법은 다음과 같다.

① 입력된 문서에서 하나의 단어를 추출한다.
② 불용어인 경우 이를 제거하고 아니면 시소러스를 검색한다.
③ 시소러스를 사용하여 해단 감정성분 값을 누적한다.
④ 값을 계산할 때 시소러스의 상위어 하위어 유사어 등에 값을 차등화 하여 누적한다.
⑤ 만들어진 5개의 벡터 값을 정규화한다.

3. 신문기사의 감정추출 시스템

본 논문은 감정기반 신문기사 검색 시스템 구현을 위해 기사별 감정의 정도 측정과 추출 방법을 제시한다. 감정추출의 전체적인 시스템 구성은 [그림 1]과 같다.



[그림 1] 감정추출 시스템

감정추출을 위한 첫 번째 단계는 신문기사 DB의 각 문서들의 구문 분석(Syntax Analysis) 과 품사 태깅(PoS Tagging)이다. 이는 기존의 구축되어진 FreeLing 1.2 PoS Tagging을 사용한다[7]. 그리고 두 번째 단계에서 실제 태깅된 품사 중 동사를 추출(Verb Extraction)한다. [그림 2]는 실제 신문기사 문서의 품사태깅과 동사추출의 예를 보여준다.

A blast from a suicide car bomb Tuesday near Baghdad's Adhamiya Palace killed four people and wounded 38 others, authorities said.	
A	*DT*
blast	*NN*
from	*IN*
a	*DT*
suicide	*NN*
car	*NN*
bomb	*NN*
Tuesday	*NP*
near	*IN*
Baghdad_'s_Adhamiya_Palace	*NP*
killed	*VBD*
four	*JJ*
people	*NNS*
and	*CC*
wounded	*NN*
38	*Z*
others	*NNS*
,	*Fc*
authorities	*NNS*
said	*VBD*
.	*Fp*

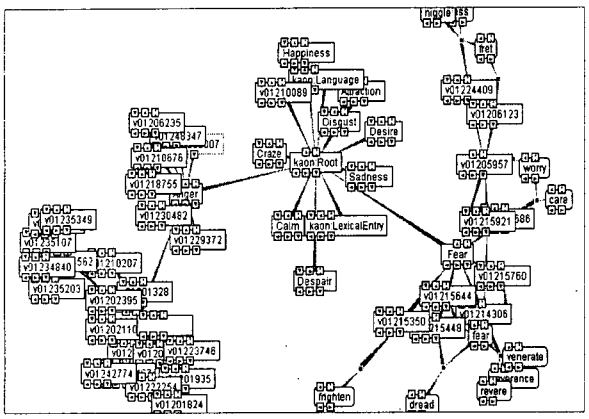
[그림 2] 품사태깅과 동사추출

위와 같이 추출된 동사들을 이용하여 신문기사의 감정정보를 분석하기 위해 본 연구에서는 감정 온톨로지(Emotion Ontology)를 참조한다. 이를 위하여 어휘 온톨로지의 일종인 워드넷(WordNet)의 동사중 *Emotion or Psych Verbs*들을 OWL/RDF(S)를 기반으로 한 KAON툴을 이용해서 온톨로지를 구축하였다.

```
<owl:Class rdf:ID="Fear">
    <rdfs:label xml:lang="en">Fear</rdfs:label>
</owl:Class>
<owl:Class rdf:ID="v01215644">
    <rdfs:label xml:lang="en">v01215644</rdfs:label>
    <rdfs:subClassOf rdf:resource="#Fear"/>
</owl:Class>
<owl:Class rdf:ID="v01205957">
    <rdfs:label xml:lang="en">v01205957</rdfs:label>
    <rdfs:subClassOf rdf:resource="#Fear"/>
</owl:Class>
<owl:Class rdf:ID="v01215448">
    <rdfs:label xml:lang="en">v01215448</rdfs:label>
    <rdfs:subClassOf rdf:resource="#Fear"/>
</owl:Class>
.....
<rdf:Description rdf:ID="fear">
    <rdfs:label xml:lang="en">fear</rdfs:label>
    <rdf:type rdf:resource="#v01215448"/>
    <rdf:type rdf:resource="#v01214306"/>
    <rdf:type rdf:resource="#v01215644"/>
    <rdf:type rdf:resource="#v01215921"/>
    <rdf:type rdf:resource="#v01215760"/>
</rdf:Description>
<rdf:Description rdf:ID="worry">
    <rdfs:label xml:lang="en">worry</rdfs:label>
    <rdf:type rdf:resource="#v01205686"/>
    <rdf:type rdf:resource="#v01205957"/>
</rdf:Description>
.....
```

[그림 3] 감정 온톨로지를 위한 OWL/RDF(S) 표현

위의 [그림 3]을 이용하여 [그림 4]의 감정 온톨로지를 구축하였다. 워드넷에서 *Emotion or Psych Verbs*들은 대표적인 5가지의 감정(Happiness, Sadness, Fear, Anger, Disgust)의 종속되는 관계를 이용하였다. 그러나 우리는 동사의 sense들을 분석 및 분류하여 10개의 감정(Happiness, Sadness, Fear, Anger, Desire, Calm, Craze, Disgust, Attraction, Despair)으로 최상위 클래스를 확장하였다.



[그림 4] 감정 온톨로지

마지막 단계로, 추출된 감정동사들은 감정 온톨로지를 참조하여 자신의 위치를 찾고 최상위 클래스와의 에지를 기반으로 유사도를 측정하여 최상위 클래스인 감정을 추출, 감정의 정도를 계산한다. 수식 [1]은 본 논문에서 제안하는 에지 기반의 감정 정도를 측정하는 식이다. 이를 이용하여 감정별 백분율을 결과로 제시한다.

$$E_{T_i} = \sum_{k=1}^{n} [\, D_{\max}(T_i) - MIN(d_k)] \qquad [1]$$

- E_{T_i} : 추출된 감정 정도
- $D_{\max}(T_i)$: 추출된 감정 서브클래스의 최대 깊이
- $MIN(d_k)$: 추출된 감정과 감정동사간의 최단거리

감정 정도는 추출된 감정의 서브클래스들의 깊이 중 가장 최대의 값과 감정동사가 갖는 클래스 중 최상위 클래스와의 최단거리의 차를 감정동사의 클래스 수만큼 합하여 측정한다.

에지 기반 측정 방식은 개념간의 거리를 측정하므로 의미적 유사성을 평가하는데 더욱 직접적인 방식이라고 할 수 있다. 에지 기반 측정은 더 직관적인 방법으로 미리 정의된 계층적 단어 네트워크의 구조에 많이 의존한다.

다음은 본 논문에서 제시하는 방법으로 실제 신문기사의 감정을 추출하고 이를 분석하였다. 신문기사는 CNN에서 제공하는 문서들을 모아 데이터셋을 만들었다. [그림 5]에서는 실험의 일부를 단계별로 제시하고 신문기사의 감정추출과 감정의 정도를 측정하는 결과를 볼 수 있다.

49 dead as Japanese train derails
Monday, April 25, 2005
Posted: 0714 GMT (1514 HKT)
TOKYO, Japan (CNN) -- In Japan's deadliest rail accident in more than 40 years, a commuter train went off the tracks during Monday morning rush hour outside Osaka in central Japan, killing 49 people and injuring more than 200 others, many seriously, authorities said.
<이하본문생략>----------

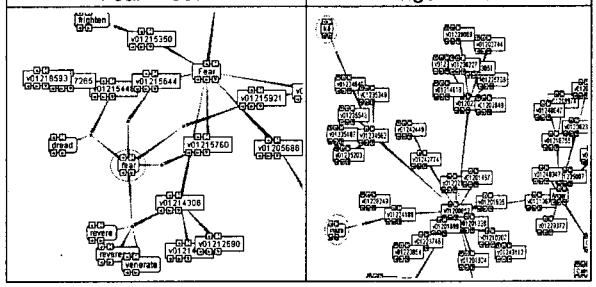
단계 1 : 뉴스기사의 구문 분석과 품사 태깅

단계 2 : 동사 추출 (동사 :33개, 감정동사 :3개)

's 's VBZ	<중간생략>----------
killing kill VBG	is be VBZ
injuring injure VBG	had have VBD
said say VBD	was be VBD
said say VBD	accompanied accompany VBN
feared fear VBD	was be VBD
go go VBP	said say VBD
operated operate VBN	is be VBZ

단계 3 : 온톨로지를 참조하여 감정 추출

Fear : 30%	Anger : 70%



[그림 5] 신문기사의 감정추출

4. 결론 및 향후 연구

본 논문은 감정기반 신문기사 검색을 목적으로 기사의 감정추출 방법을 제안하였다. 감정동사들의 관계를 온톨로지로 구축하여 신문기사내 어휘정보 중 감정동사들간의 거리를 측정하였다. 이를 이용하여 최상위 클래스의 대표감정들을 추출하고 감정의 정도를 측정하였다. 향후 이를 기반으로 감정어휘에 따른 신문기사 검색 시스템을 구축하고 감정을 이용한 컴퓨팅 분야에서 적용될 수 있는 방법을 연구할 것이다.

참고문헌

[1] 박효진, 황상웅, 안연하, 심재원 "인터랙티브 감성 공간 - 유비쿼터스 라이프", 한국정보과학회, 2003
[2] George A. Miller "Introduction to WordNet:An On-line Lexical Database" 1993
[3] Colby, M., "Modeling a paranoid mind," The Behavioral and Brain Sciences, pp. 515-560, 1981
[4] Dyer, M. G.,"In depth understanding." MIT Press,1983
[5] Wright, I. P., "Emotional Agents," Ph. D. thesis, Univ. of Birmingham, 1997
[6] 김영관, 박영택, "감정요소를 사용한 정보검색에 관한 연구," 정보처리학회논문지 B, 제10-B권, 제6호, 2003
[7] Carreras, X., I., Chao, L. Padró., M. Padró, "FreeLing: An Open-Source Suite of Language Analyzers," LREC'04, 2004