사건과 상태의 선호도 분류

양재군° 배재학
울산대학교 컴퓨터·정보통신공학부
{jgyang°, jhjbae}@ulsan.ac.kr

Classifying Preference Degree of Events and States

Jae-Gun Yang° Jae-Hak J. Bae
School of Computer Engineering & Information Technology, University of Ulsan

요 약

Plot Unit는 이야기를 형성하는 줄거리 또는 줄거리에 나오는 여러 사건을 하나로 구성하여 표현한다. 글을 읽고 Plot Unit를 파악한다는 것은 그 글의 내용을 이해하고 있다는 것이다. 본 논문에서는 이러한 Plot Unit의 정서상태 선호도를 결정하는 방법으로 범주 재분류를 생각하였다. Roget 범주들을 양, 음 기준에 따라서 양범주, 음범주, 중성범주로 재분류하였다. 또한, 개연규칙과 Plot Unit의 대응에 이 결과를 적용해 봄으로써, 범주 재분류를 활용하여 Plot Unit의 사건유형을 결정할 수 있음을 확인하였다.

1. 서 론

자연언어처리에서는 이야기에 등장하는 주인공들 간의 상호작용을 표현하는 한 방법으로 Plot Unit[1,2]를 활용한다. 이는 등장인물들의 심리적인 상태와 사건을 연결하여 전체적인 줄거리 구성을 이끌어 내는 방법이다. Plot Unit을 파악하는 것은 글을 이해하는 것이다.

본 논문에서는 Plot Unit을 구성하는 세 가지 정서상태(Affect states) 중에서 긍정적인 사건(Positive Event)과 부정적인 사건(Negative Event)의 구분에 사용할 목적으로 Roget 범주를 양, 음 그리고 중성 범주로 재구성하는 방법을 모색하였다.

2. Roget 시소러스

Roget 시소러스[3]는 의미 분류를 기초로 총6개의 강(Class)으로 구성되었다. 각 강은 하부에 부(Division), 과(Section) 등의 계층구조로 세분화되었다. 각 계층은 고유한 범주표제를 가지고 있으며 계층구조의 말단에는 총1044개의 범주(Category)가 존재한다. 각 범주에는 품사별로 유의어 목록이 나열되어 있다. 한편, 유의어 목록에서 특정 어휘가 다른 범주를 참조하는 경우에는 "어휘 &c.(표제어) 표제번호"의 형식으로 표현한다. 이를 참조정보라 하였다. 본 논문에서는 이러한 범주표제와 참조정보를 범주 재분류에 활용하고자 한다. 또한, Roget 시소러스의 원형을 기계가 사용할 수 있는 형태로 미리 전처리한 어휘사전 ROTIP(ROget's Thesaurus In Prolog)[4]을 이용하였다.

2.1 이원성과 대칭성

과학과 수학은 덧셈과 뺄셈에서부터 초대칭성 이론까지 이원성과 대칭성이라는 단순한 개념을 이용한다. 이원성과 대칭성은 언어 체계에서도 중요한 성질이다[5]. Roget 시소러스의 범주 시스템 역시 개념을 공유하면서 긴밀한 관계가 있는 반의어들은 서로 이웃한 범주에 배치되어있다. 이처럼 사람은 어떤 의미차원에 대해서 그 속에서 대비되는 두 개념을 생각해내고 그 두 개념은 보통 해당 차원의 양 극단에 위치한다.

서로 반대말인 두 어휘가 공유하는 개념의 정도와 그렇지 않은 두 어휘의 그것과 비교해 보면 반의어쌍 쪽이 그 정도가 크다. 예를 들어, 더위(Heat)과 추위(Cold)는 온도 차원(속성)을 서로 공유한다. 온도에 관련된 다른 어휘들을 위상에 따라 나열할 수 있는 것과 마찬가지이다. 이러한 연속체 혹은 범위상에 존재하는 Cold와 Heat 간의 어휘순열은 {cold, cool, tepid, luke-warm, warm, heat}이다. {Heat-Cold} 어휘쌍은 {Heat-Light} 등의 어휘쌍에 비해서 더 많은 공통점이 있다. 이처럼 반의어쌍은 공유하는 개념차원에서 극성이 다르지만 어휘 전체 차원에서는 가까운 위치에 있다고 할 수 있다. Roget 시소러스에서 Heat는 #382.범주이고 Cold는 #383.범주이다. 두 범주는 온도(Heat)라는 범주에 공통으로 속해있다. 한편 Light는 #402.범주이며 Heat, Cold 등의 범주와는 먼 거리에 위치한다.

본 논문에서는 서로 대칭인 범주들을 위주로 Roget 범주 시스템을 양, 음 그리고 중성 등 세 가지 그룹으로 재분류하였다.

2.2 범주 재분류

[그림 2]는 Roget 범주 시스템의 일부이다. 그림에서는 사회적인 정서에 해당하는 범주들을 보여준다. 여기에서 서로대칭인 범주쌍은 다음과 같다: {#888.-#889.}, {#890.-#891.}, {#892.-#893.}, {#894.-#895.}, {#897.-#898.}, {#903.-#904.}, 그리고 서로 대칭이 아닌 {#896., #899., #901., #901a., #902., #905.}. 이 범주쌍들을 Plot Units의 E+와 E-구분 기준인 Positive와 Negative 성질의 관점에서 세 그

룹으로 분류할 수 있다: {Positive : #888., #890., #892., #894., #897., #903.}과 {Negative : #889., #891., #893., #895., #898., #904.}, 그리고 {미분류 : #896., #899., #901., #901a., #902., #905.}. 이처럼 Roget 범주들을 Plot Units를 구성하는 엔터티로 재분류하는 것이 가능하다. 한편, Plot Units에서는 엔터티를 심상(Mental)과 사건(Event)으로 나눈 후에 사건(Event)을 E+와 E-로 분류하여 총 세 가지 정서상태로 구성한다. 여기에서 첫 과정인 심상과 사건의 분별은 기존연구[6]를 활용한다.

⊙ Class VI: Words Relating to the Sentient and Moral Powers
- SECTION III. SYMPATHETIC AFFECTIONS
  - 1. SOCIAL AFFECTIONS

#888. Friendship. (+)
#889. Enmity. (-)
#890. Friend. (+)
#891. Enemy. (-)
#892. Sociality. (+)
#893. Seclusion. Exclusion. (0)
#894. Courtesy. (+)
#895. Discourtesy. (-)
#896. Congratulation. (0)
#897. Love. (+)
#898. Hate. (-)
#899. Favorite. (0)
#900. Resentment. (0)
#901. Irascibility. (0)
#901a. Sullenness. (0)
#902. [Expression of affection or love.] Endearment. (0)
#903. Marriage. (+)
#904. Celibacy. (-)
#905. Divorce. (0)

[그림 1] 분류된 Roget 범주 시스템의 일부

※ 범주표제 끝부분의 기호는 재분류 결과를 나타낸다.

2.3 양음범주 결정

앞 절에서 예시한 재분류 방법을 모든 Roget 범주에 적용했다. 분류기준인 Positive, Negative 성질에 해당하는 범주들을 각각 양범주, 음범주로 부르기로 한다. 이때 재분류된 양, 음범주의 레벨은 0이다<표2>.

<표 2> 양, 음범주 분류 결과 (레벨 0)

Class	범주개수	양,음범주	비율
1. Abstract Relations	187	38	20.32%
2. Space	139	18	12.95%
3. Matter	142	32	22.54%
4. Intellect	157	58	36.94%
5. Volition	232	112	48.28%
6. Affections	187	104	55.61%
합계	1044	362	34.67%

이 과정에서 서로 대칭이면서 이웃하지만 양, 음 성질로는 분류하기 어려운 예도 있다(#16. Uniformity. - #16a. Nonuniformity.). 이 경우에는 두 범주의 분류를 보류한다.

레벨이 0인 양, 음 범주들을 토대로 미분류 범주들의 양, 음 성질을 결정한다. 이 과정에서는 참조정보를 이용한 범주 재분류 방법[7]을 이용하였다. [그림 3]은 미분류 범주의 성질을 결정하는 과정이다. 이렇게 결정된 범주들의 레벨은 1이다.

```
// 주요용어
Cnc      // 미분류 범주
Cr       // 참조관계에 있는 Roget 범주
Cpn      // Cnc의 양, 음 성질
Ir       // 참조정보
Np       // 양(Positive) 수치
Nn       // 음(Negative) 수치
Ncr      // Cr의 개수

repeat (모든 Cnc에 대해서)
  Cnc의 Ir를 탐색해서 Cr을 찾는다.
  repeat (모든 Cr에 대해서)
    if (Cr의 성질이 양이다.)
      Np를 1만큼 증가
    else
      Nn을 1만큼 증가
    end if
  end repeat
  // 양, 음의 비율이 Ncr의 반을 넘으면
  // Cnc의 성질을 결정한다.
  if (Np / Ncr > 0.5)
    Cpn = 양      // Cnc는 양범주이다.
  else if
    Cpn = 음      // Cnc는 음범주이다.
  else
    Cpn = 미분류  // Cnc의 성질은 미정.
  end if
end repeat
```

[그림 2] 미분류 범주의 성질을 결정하는 과정

레벨 0인 양, 음 범주에 새로 결정된 레벨 1의 양, 음 범주를 더하여 위 과정을 반복하여 레벨 2의 양, 음 범주를 얻는다. 이 과정을 모든 범주의 성질이 결정되거나 혹은 범주의 성질을 결정할 수 없을 때까지 반복한다. 실험 결과, 레벨 5까지만 발견할 수 있었다<표 3>. 나머지 범주들은 중성범주이다.

<표 3> 재분류 결과

레벨	양	음
0	181	181
1	10	7
2	20	43
3	7	6
4	1	2
5	1	0

3. Plot Unit

Plot Unit는 이야기(Narrative)를 표현하는 단위이다. 이는 이야기에 등장하는 주인공들 사이에서 일어나는 상호작용을 표현하고, 원문에서 언급된 사건이나 상황을 묘사한다.

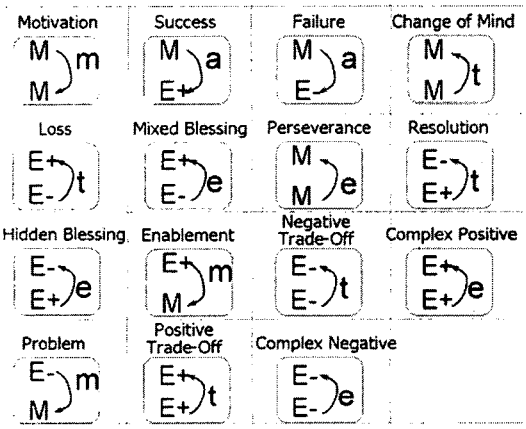
3.1 Plot Unit의 구성

Plot Unit는 정서상태(Affect States)라는 엔터티로 구성된다. 이것에는 중립적인 심상을 의미하는 "M : Mental"과 긍정적인 사건을 표현하는 "E+ : Positive"와 부정적인 사건을 묘사하는 "E- : Negative"등이 있다. 정서상태는 등장인물의 관념적인 계획, 목적, 외부 사건

에 대한 반응들을 나타낸다. Plot Unit는 2개의 심상을 연결한 인과사슬(Causal Links)로 표현된다. 이 사슬은 두 심상의 관계를 motivation(m), actualization(a), termination(t), equivalence(e) 등으로 구분한다.

3.2 Primitive Plot Unit

인과관계 사슬로 연결된 Plot Unit는 Primitive Plot Unit와 Complex Plot Unit로 나눌 수 있다. Primitive Plot Unit는 등장인물 하나 또는 둘 사이의 정서상태에 관한 모든 가능한 관계들을 표현한다. Complex Plot Unit는 Primitive Plot Unit들이 결합된 것이다. [그림 1]은 두 심상을 하나의 인과사슬로 연결한 15개의 Plot Unit이다.



[그림 3] Primitive Plot Unit의 예

3.3 개연규칙과 Plot Unit의 대응에 활용

<표 4>는 특정 이야기(Narrative)의 한 문장과 그 경우의 개연규칙[8]이다. 실험 결과를 예시한 개연규칙과 Plot Unit을 대응하는데 활용하고자 한다.

<표 4> Plot Unit 발견에 적용할 개연규칙의 예

원 문	I assumed since the wedding was *canceled*, the gifts would be *returned*.
개연규칙	사라진 것을 복구 하다.
어 휘	canceled <= returned
OfN정보	event(obliteration) <= event(restitution).

canceled는 #522.범주에 속하고 return은 #790.범주에 속하는 어휘이다. #790.범주는 레벨 1 양범주에 해당하고 #552.범주는 레벨 0 음범주에 해당한다. Plot Units의 심상은 각각 E+, E-가 된다. 이에 해당하는 정서상태 유형은 Resolution과 Hidden Blessing이다. 두 유형은 인과사슬(Causal Links)에 따라서 구분할 수 있는데, 이 경우는 canceled과 return이 equivalence보다는 termination에 가깝다. 이상을 정리하면 E- <-(t)- E+ 형태임을 알 수 있고 이것은 Plot Units 중에서 Resolution 유형[그림 3]에 해당한다.

4. 결론

본 논문에서는 개연규칙과 Plot Unit의 대응에 활용할 목적으로, Plot Unit 사건과 상태의 선호도를 분류하였다. 구체적으로는, Roget 범주들을 양, 음 기준에 따라서 양범주, 음범주, 중성범주로 재분류하는 방법을 이용하였다. 이 과정에서는 개념을 공유하면서 긴밀한 관계인 반의어들은 서로 가까운 범주에 배치한 Roget 범주 시스템의 원리를 이용하였다.

구체적인 결정 과정은, Roget 범주 시스템의 이원성과 대칭성을 활용하여 레벨 0인 양범주, 음범주를 결정한다. 이후에는 레벨 0 범주들을 토대로 참조정보를 이용한 범주 재분류 방법으로 레벨 1인 양, 음범주를 찾아낸다. 이 방법을 반복 적용해서 레벨 5 까지의 양, 음범주를 탐색할 수 있었다.

재분류 결과를 개연규칙과 Plot Unit의 대응에 이 결과를 적용해 봄으로써, 범주 재분류를 활용하여 Plot Unit의 사건 유형을 결정할 수 있음을 확인하였다.

본 논문에서 고안한 방법으로, 특정 어휘의 성질이 양인지 음인지 혹은 중성인지를 결정할 수 있다. 이를 이용하면 어휘사슬(lexical chain)을 선택하는 경우, 사슬을 구성하는 각 노드에 기호를 부여할 수 있어서 기존의 방법을 보완할 수 있을 것이다.

<Acknowledgements>
본 연구는 한국과학재단 목적기초연구 R05-2004-000-12362-0 지원으로 수행되었음. 또한 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업의 연구결과로 수행되었음.

참고문헌

[1] Lehnert, W.G., "Narrative complexity based on summarization algorithms.", Proc. of IJCAI, pp. 713-716, 1983.

[2] Lehnert, W.G., Loiselle, C.L., "Plot unit recognition for narratives." (Tech. Rep. No. 83-39) Amherst, MA: University of Massachusetts, Department of Computer and Information Science, 1983.

[3] Roget's Thesaurus, http://www.gutenberg.org/dirs/1/0/6/8/10681/10681-h-body-pos.htm.

[4] 양재군, "시소러스의 기계 가용화에 대한 연구", 울산대학교 석사학위논문, 2000.

[5] Old, L. John, (2003). The Semantic Structure of Roget's, A Whole-Language Thesaurus. (Doctorial dissertation, Indiana University, 2003). Dissertation Abstracts International (UMI No. AAT 3122723)

[6] 양재군, 배재학, "온톨로지 정보를 이용한 범주 재편성: 로젯 시소러스의 경우", 한국정보처리학회 제18회 춘계학술발표대회 논문집, 제9권, 제1호, pp. 515-518, 2002.

[7] 양재군, 배재학, 이종혁, "온톨러지 재사용을 위한 범주 재분류", 한국정보처리학회논문지 B, 제12-B권, 제 1호, pp. 69-80, 2005.

[8] 김곤, 배재학. 개연성 규칙과 문장추상화를 활용한 문서요약. 한국정보처리학회 제18회 추계학술발표대회 논문집, 제9권, 제2호, pp. 359-362, 선문대학교.