Unknown Word Lexical Dictionary의 자동 생성 방법

황명권*, 윤병수*, 정일용*, 김판구*
*조선대학교 컴퓨터공학부
e-mail: mghwang@chosun.ac.kr, neoybs@live.co.kr, iyc@chosun.ac.kr, pkkim@chosun.ac.kr

Automatic Construction Method of Unknown Word Lexical Dictionary

Myung-gwon Hwang*, Byung-su Youn*, Pan-koo Kim*
*Dept of Computer Engineering, Cho-sun University

요 약

본 연구는 의미적 정보 검색을 위한 연구 중의 하나로, 현재까지의 의미적 문서 검색에서 큰 걸림돌이었던 사전에 정의되지 않은 단어(Unknown Word)들의 어휘 사전(Lexical Dictionary)을 자동으로 생성하기 위한 것이다. 이를 위해 UW를 기존의 영어 어휘 사전인 워드넷(WordNet)에 정의되지 않은 단어로 간주하고, 웹 문서의 입력을 통하여 UW와 관련된 단어들을 추출하여 의미적 관련 정도를 확률적, 의미적 방법으로 측정한다. 본 논문에서는 UW Lexical Dictionary를 자동으로 구축하기 위한 방법에 대해서만 기술하였고, 정량적이고 객관적인 평가는 포함하지 않고 있다. 하지만 본 연구의 효용성을 확인하기 위한 몇 가지 문서로부터 추출된 결과는 본 연구가 상당히 의미적이며 가치가 높을 것으로 기대되고 있다.

1. 서론

의미적 정보 검색을 위해 많은 연구들이 수행되고 있다. 이러한 의미적 검색을 위해 문서의 semantic indexing에 관한 연구[1,3], semantic metadata creation에 관한 연구[3,4,5], 문서의 topic을 detection하는 연구[6,7] 등이 수행되었다. 그리고 많은 연구에서 어휘 사전을 기반으로 의미적 검색을 위해 노력하고 있다. 가장 대표적인 영어 어휘 사전은 워드넷[2]이라 할 수 있다. 어휘 사전을 기반으로 하는 의미적 검색의 가장 취약점은 어휘 사전에 정의되지 않은 개념들을 다루지 못한다는 점이다. 워드넷 또한 마찬가지이다. 워드넷은 10년이 넘게 확장되어 오고 있지만, 새로운 사회 현상, 기술 개발, 트렌드, 상품 개발, 유명인사(운동선수, 엔터테이너, 정치인 등)들을 모두 커버하지는 못한다. 최근에 이러한 사전에 정의되지 않는 단어들을 unknown word(UW)라 칭하고, UW들을 웹 document, data warehouse 등을 이용하여 정의하기 위한 연구가 진행되었지만[8, 9, 10, 15], 여전히 개념들 사이의 관계된 정도를 파악하지 못하는 한계점이 존재한다. 이에 본 연구는 본 연구실에서 연구된 [11, 15]의 확장으로 어휘 사전을 이용한 진정한 의미적 검색을 위해 일반 웹 문서에서 UW를 찾고, UW와 관계된 단어들을 판단한 후, 그 의미적으로 관계된 정도를 파악하여 최종적으로 UW Lexical Dictionary를 구축하기 위함이다.

본 논문의 구성은 2장에서 본 연구의 지식 베이스이며, UW Lexical Dictionary의 기반이 되는 워드넷에 대해 상세히 기술하고, 3장에서 사전의 자동 생성을 위한 방법을 기술하고 간단한 예시를 통해 본 연구의 가치와 가능성을 보인다. 그리고 4장에서 현재까지의 연구에 대해 정리하고 향후 지속될 연구에 대해 기술한다.

2. 관련연구

워드넷은 대부분의 의미적 정보검색에 적용할 수 있는 영어 어휘 사전이다. 워드넷은 인간의 어휘적 사용 구조의 심리적 이론에서 영감을 얻어 설계된 것으로, 1985년 부터 구축되고 있다. 워드넷은 어휘를 정의하기 위해 명사, 동사, 형용사 및 부사로 나누어 정의하고 동의어, 반의어, 하의어, 상의어, 부분어, 수반어 등을 개념들간의 의미적 관계로 구성하여 사용한다. 각각의 개념간의 관계는 여러가지 기호 {@, ~, #p, #m, ;c, -c, ;r, -r, …}들을 이용하고 표현하고 있다. 그리고 단어가 표현하는 개념의 의미는 신셋(synset) 번호(하나의 신셋은 하나의 의미)로 표현된다. 하나의 신셋이 여러 단어들를 가지게 되면, 즉 여러가지 단어가 동일한 신셋으로 표현되면 이러한 단어들은 동의어이다. 만약 어떤 단어가 여러가지 의미를 갖는 다의어라면 그 단어는 표현하는 의미 수 만큼의 신셋을 가지게 된다. 예를들어 신셋 "02929975"는 "a motor vehicle with four wheels"을 의미하고 5개의 단어{car, auto, automobile, machine, motorcar}를 포함한다. 그러나 5개의 단어 중에서 "car" 만을 살펴보면, 5가지 의미를 가지며 각각의 의미는 또 다른 신셋들{“02929975”, “02931574”,

"02906118", "02932115", "02931966"}로 표현된다. 이러한 워드넷의 개념들 사이의 관계정의를 이용하여 문서 분류, 자연어 처리, 시맨틱 웹(semantic web) 그리고 개념 의미 파악을 통한 텍스트 마이닝(text mining), 개념들 사이의 관계분석과 같은 정보의 의미적인 처리를 필요로 하는 다양한 분야에 적용된다. 본 연구에서는 UW 추출과 개념들 사이의 관계를 파악하기위해 워드넷을 이용한다.

3. 본론

앞에서 기술했듯이, 워드넷은 거의 모든 영어 개념을 정의하고 있다. 본 연구에서는 UW를 워드넷에 정의되지 않은 단어로 정의하고, UW Lexical Dictionary를 웹 문서의 처리를 통해 생성한다. 그리고 UW와 관계된 개념들을 co-occurrence를 통해 파악하고, 확률적 가중치와 의미적 가중치를 각각 측정하여 최종적으로 관계된 정도를 파악한다.

3.1. 사전 처리 및 Unknown Word 추출

웹 문서에서 명사 단어들을 추출하기 위해, 품사 태깅이 필요하다. 본 연구에서는 스탠포드 대학 자연어 처리 연구 그룹 (Stanford NLP Group)에서 개발하여 제공하는 pos-tagger 버전 2006-05-21을 이용한다[13]. 문서를 구성하고 있는 문장이 입력으로 주어지면, 각 단어의 품사가 태깅되어 출력된다. 품사 태깅은 각 품사의 의미를 의미하는 태그들로 출력되는데, NN(단수 명사), NNP(고유 명사), NNS(복수 명사), NNPS(복수 고유 명사)가 명사의 종류이다. <표 1>은 [12]에 기술된 유명한 축구선수 'Zidane'에 대한 문서의 한 문장을 입력으로 출력된 결과와 추출된 명사들을 보이고 있다.

<표 1> 품사 태깅 및 명사 추출

입력	Zinedine Yazid Zidane, popularly nicknamed Zizou, is a French football player of Algerian descent, famous for leading France to winning the 1998 World Cup.
출력	Zinedine/NNP Yazid/NNP Zidane/NNP popularly/RB nicknamed/VBN Zizou,/NNP is/VBZ a/DT French/JJ football/NN player/NN of/IN Algerian/NNP descent,/NNP famous/JJ for/IN leading/VBG France/NNP to/TO winning/VBG the/DT 1998/CD World/NNP Cup/NNP
명사	**Zinedine, Yazid, Zidane, Zinedine_Yazid, Yazid_Zidane**, Zizou, **football, player, football_player**, **Algerian, descent, Algerian_descent**, France, **World, Cup, World_Cup**
UW	Zinedine, Yazid, Zidane, Zinedine_Yazid, Yazid_Zidane, Zizou

<표 1>에서 명사의 연속인 경우는 복합 명사일 수 있으므로, 독립된 단어와 연결된 단어 모두를 생성한다(볼드체 부분). 추출된 명사들을 워드넷의 명사 부분과 매칭하여 일치하는 단어가 존재하지 않으면 UW로 판단한다.

3.2. Unknown Word와 관계된 후보 단어 추출

UW와 관계된 후보 단어들은 UW와 동일한 문장에서 함께 출현한 단어일 확률이 높다[11, 15]. UW와 관계가 있을 후보 단어들을 동일한 문장에서 출현한 co-occurrence 명사들로 결정한다. UW Lexical Dictionary는 워드넷과 동일하게 단어들의 신셋을 이용한다. 이에 후보 단어들의 신셋을 파악해야 하는데, 본 연구에서는 [14]에 의해 개발된 SSI 기법을 이용하여 WSD를 수행한다. <표 2>는 [12]에서 추출된 UW 중 'Zidane'에 대한 후보 개념들을 보이고 있다. [14]는 상당히 높은 정확도를 보이지만, 의미가 판단이 되지 않는 경우가 발생할 수 있다. 그런 경우는 (-)로 표기한다.

<표 2>UW와 후보 단어

UW	후보 단어
Zidane	Algerian#1, Cup#6, descent#5, football#1, France#1, June#1, player#1, World#1, World_Cup#1, attention#2, Brazil#1, country#5, Europe#1, fame(-), goal#3, play_maker#1, Year#4, April#1, football_player#1

3.3. 후보 단어들의 가중치 측정

앞의 과정에서 UW와 동일한 문장에 함께 출현한 단어들을 추출하여 후보 단어 집합을 생성하고, 정확한 의미 파악을 위해 WSD를 수행하였다. 이번 과정부터는 UW와 얼마나 관계성이 높은지를 측정하기 위해 확률적 가중치를 측정하고, 다음 과정에서 후보 단어들 사이의 관계성을 이용해 의미적 가중치를 계산한다.

후보 단어들 중에 UW와 함께 출현한 횟수가 많다면 그 단어는 UW와 관계할 확률이 높다. 이를 위해 후보 단어들의 베이지안 확률을 측정한다. 이 때, 후보 집합에서 2단어 이상 연속된 명사들은 복합명사일 가능성이 있다. 3.1에서 단일명사를 이용해 복합명사를 생성하였기 때문에, 만약 워드넷 매칭을 통해 복합명사로 판단된다면 복합명사를 구성하는 단일명사의 베이지안 확률에서 복합명사의 베이지안 확률을 제외시켜야 중복이 일어나지 않는다. 예를 들어, 문서에서 'World' 1회, 'World Cup'이 2회 출현하였다면, 'World'는 3회, 'Cup'은 2회, 'World Cup'이 2회 발생한 것으로 계산된다. 이는 'Cup'의 2회, 'World'의 2회가 독립적으로 발생하지 않은 값이 더해진 결과이다. 이와 같은 중복을 피하기 위해 수식 (1)과 같이 각각의 베이지안 확률을 측정하고 <표 3>의 복수명사의 베이지안 확률을 제외하는 과정을 추가하였다.

$$Bayesian = \frac{P(oc(uw_i)|oc(k_{ij})) \times P(oc(k_{ij}))}{P(oc(uw_i))} \quad (1)$$

oc는 단어의 출현횟수를 나타내고, uw_i는 i 번째 UW, k_{ij}는 uw_i의 j 번째 단어를 의미한다. <표 4>는 이러한 과정으로 측정된 후보 단어들의 가중치를 보이고 있다.

<표 3> 복합명사의 경우

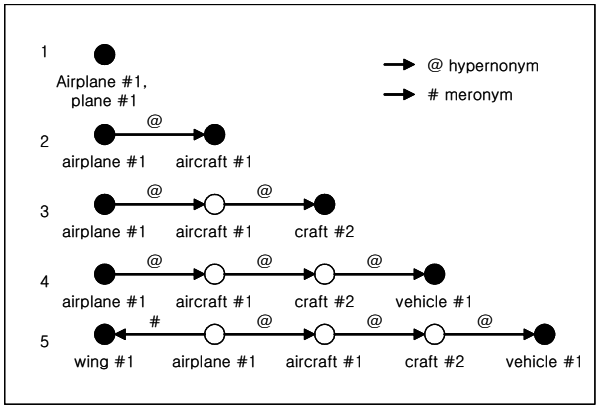
ck_{ab}를 uw_i의 후보 단어들 중 a와 b단어로 이루어진 복합명사로 정의 P'은 단일명사의 고유 베이지안 확률 $P'(oc(k_a)\|oc(uw_i))$ $= P(oc(k_a)\|oc(uw_i)) - P(oc(ck_{ab})\|oc(uw_i))$

<표 4> 후보 단어들의 가중치

uw_i	단어 (k_{ij})	베이지안 확률	단어 (k_{ij})	베이지안 확률
Zidane	Algerian	0.2	April	0.2
	football	0.2	football player	0.2
	June	0.2	player	0.2
	World Cup	0.6	attention	0.2
	country	0.2	Europe	0.2
	goal	0.2	playmaker	0.2
	descent	0.2	Brazil	0.2
	France	0.2	fame	0.2
	World	0.2	Year	0.2

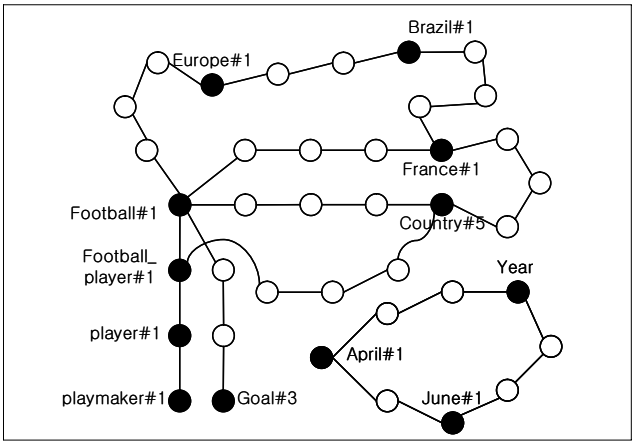
3.4. 후보 단어들의 관계성 측정 및 사전 구축

만약 UW와 함께 출현한 후보 단어들이 UW와 관계가 깊다면, 후보 단어들끼리 서로 의미적 관계를 형성할 수 있다[6, 15]. 이것은 문서에서 하나의 주제를 기술하기 위해 주제와 관련된 여러 가지 단어를 이용하는 것과 동일하다. 후보 단어들의 신셋들을 이미 파악했고, 후보 단어들 사이의 관계파악은 워드넷 매칭을 통해 가능하다. 신셋들을 이용하여 관계를 파악할 때, 5노드 사이에 관계가 형성된다면 두 신셋은 관계가 있는 것으로 판단하였다. 이는 문서에 출현한 두 단어 사이의 직접적인 관계보다는 3단계 이상 떨어져 있는 경우가 더 많기 때문이다[14]. [그림 1]은 워드넷에서 두 개념사이의 관계 형성도에 대한 예를 보이고 있다. 그림에서 숫자는 노드의 수를 나타낸다.



[그림 1] 관계 형성도

각 단어의 신셋을 워드넷에 매칭하여 신셋들 사이의 관계를 얻을 수 있다. [그림 2]는 <표 4>의 후보 단어들의 관계도를 보이고 있다.



[그림 2] 후보 단어들의 관계도

단어들 사이의 관계는 멀수록 작아지기 때문에 반비례관계로 볼 수 있다. 그리고 각 단어의 관계 가중치(R)는 5노드 거리 이내에 관계한 모든 단어들과의 관계된 정도의 합으로 계산할 수 있기 때문에 수식 (2)를 이용하여 측정한다.

$$R(k_{ij}) = \sum_{j=1}^{n} \frac{1}{D_{node}(k_{ij}, k_{il})}, j \neq l \qquad (2)$$

수식에서, n은 uw_i의 후보 단어들의 개수이며, D_{node}는 k_{ij}와 k_{il} 사이의 관계 거리를 의미한다. 수식에서 만약 두 단어 사이의 노드의 수(두 단어를 의미하는 노드 포함)가 5를 초과할 경우 $\frac{1}{D_{node}}$는 0이 된다. 그리고 3.3에서 측정한 베이지안 확률과 수식 (2)의 관계 가중치는 서로 값에 영향을 미치지 않으므로 독립이라 할 수 있다. 하지만, 만약 후보 단어들 중 어떤 단어와도 관계를 갖지 않는 단어가 존재한다면, 그 단어의 R 값은 0이 된다. 이는 <표 4>의 결과와 상관없이 0을 만들기 때문에 최종적으로 R 값에 1을 더한다. 그래서 uw_i와 의미적으로 관계된 정도는 수식 (3)에 의해 측정할 수 있다. <표 5>는 수식 (3)에 의한 결과를 보이고 있다.

$$SR(k_{ij}) = P'(oc(k_{ij})|oc(uw_i)) \times (R(k_{ij}) + 1) \qquad (3)$$

<표 5> 후보 단어들의 의미적 관련 정도

uw_i	k_{ij}	SR	k_{ij}	SR
Zidane	Algerian	0.2	April	0.32
	football	**0.59**	**football player**	**0.51**
	June	0.37	**player**	**0.46**
	World Cup	**0.6**	attention	0.2
	country	0.32	Europe	0.29
	goal	0.29	**playmaker**	**0.42**
	descent	0.2	Brazil	0.29
	France	0.32	fame	0.2
	World	0.2	Year	0.3

4. 결론 및 향후 연구

본 연구는 의미적 검색을 위해, 사전(Dictionary)에 구축되지 않은 단어(Unknown Word)들을 다양한 문서집합을 이용하여 자동으로 관계된 단어를 추출하기 위한 연구이다. 방법으로는 co-occurrence 단어들을 이용하여 후보 단어집합을 추출하고, 후보 단어들의 베이지안 확률을 이용하여 확률적 가중치 측정 및 후보 단어들 사이의 관계를 통해 의미적 가중치를 계산하였다. 현재는 본 연구의 효용성을 평가하기 위해 다양한 웹 문서들을 이용하고 있으며 결과는 상당히 연구의 가치가 높은 것으로 나타났다. 본 연구의 결과물인 UW Lexical Dictionary는 현재까지 의미적 문서 검색 등에서 걸림돌이었던 UW들을 처리하는데 아주 효율적으로 사용될 것으로 기대된다. 하지만 정량적, 객관적인 평가가 제외되어 있어 이를 위해 많은 연구에서 사용되고 있는 corpora를 사용할 필요가 있다. 이는 지속된 연구와 실험을 통해 향후에 이루어질 것이다.

감사의 글

본 연구는 문화관광부 및 한국문화콘텐츠진흥원의 문화콘텐츠기술연구소육성사업의 연구결과로 수행되었음

참고문헌

[1] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, Vol. 41, Issue 6, pp. 391-407, Jan. 1990.

[2] http://wordnet.princeton.edu/

[3] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff, "Semantic annotation, indexing, and retrieval", Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 2, Issue 1, pp. 49-79, Dec. 2004.

[4] Siegfried Handschuh, Steffen Staab, and Fabio Ciravegna, "S-CREAM - Semi-automatic CREAtion of Metadata", Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, LNCS 2473, pp. 165-184, 2002.

[5] Alani, H., Kim, S., Millard, D., Weal, M., Hall, W., Lewis, P., and Shadbolt, N. "Automatic Ontology-based Knowledge Extraction from Web Documents", IEEE Intelligent Systems, Vol. 18, Issue 1, pp. 14-21, Jan. 2003.

[6] Irina Matveeva, and Gina-Anne Levow, "Topic Segmentation with Hybrid Document Indexing", In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 351-359, June 2007.

[7] Hyunjang Kong, Myunggwon Hwang, Gwangsu Hwang, Jaehong Shim, and PanKoo Kim, "Topic Selection of Web Documents Using Specific Domain Ontology", MICAI 2006:Advances in Artificial Intelligence, LNAI 4293, pp. 1047-1056, 2006.

[8] Roberto Navigli, and Paola Velardi, "Ontology Enrichment Through Automatic Semantic Annotation of OnLine Glossaries", Managing Knowledge in a World of Networks, LNCS 4248, pp. 125-140, 2006.

[9] Paola Velardi, Alessandro Cucchiarelli, and Michael Petit, "A Taxonomy Learning Method and Its Application to Characterize a Scientific Web Community", IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 2, pp. 180-191, Feb., 2007.

[10] Michele Missikoff, Paola Velardi, and Paolo Fabriani, "Text Mining Techniques to Automatically Enrich a Domain Ontology", Applied Intelligence, ISSN 0924-669X, Vol. 18, No. 3, pp. 323-340, 2003.

[11] Myunggwon Hwang, Sunkyoung Baek, Junho Choi, Jongahn Park, and Pankoo Kim, "Grasping Related Words of Unknown Word for Automatic Extension of Lexical Dictionary", First International Workshop on Knowledge Discovery and Data Mining (WKDD 2008), pp. 31-35, Jan. 2008.

[12] http://en.wikipedia.org/wiki/Zidane

[13] http://nlp.stanford.edu/software/tagger.shtml

[14] R. Navigli and P. Velardi, "Structural Semantic Interconnections: a knowledge-based approach to word sense disambiguation", Special Issue-Syntactic and Structural Pattern Recognition, IEEE TPAMI, Vol. 27, Issue 7, 2005.

[15] Hyunjang Kong, Myunggwon Hwang, and PanKoo Kim, "The Method for the Unknown Word Classification", PKAW2006, LNCS4303, pp.207-215, August, 2006.