의미 경계의 현실화를 위한 공기정보의 자동 군집화

신사임[0], 최기선
한국과학기술원 전산학과, 전문용어 언어공학연구센터, 언어자원은행[1]
{miror, kschoi}@world.kaist.ac.kr

Automatic word sense clustering using collocation for practical sense boundaries

Saim Shin[0], Key-Sun Choi
KAIST, KorTerm, BOLA

요 약

본 논문에서는 다의어의 현실적인 의미 분포의 결정에 대해 이야기 하고자 한다. 수동으로 구축한 의미체계인 사전이나 시소러스들은 그 의미구분의 경계가 모호하고 비현실적인 부분이 많아서 언어처리 시스템의 적용에 문제점으로 지적되고 있다. 그러므로, 본 연구에서는 대용량 코퍼스에서 추출한 공기정보와 자동 군집화 방법들을 사용하여 실질적인 다의어의 의미 경계를 발견하는 방법을 제안하였다.

수동 구축된 사전과 코퍼스 기반 사전의 다의어 의미 분포와 비교해 본 결과, 본 논문에서 제안한 방법의 결과가 코퍼스 기반 사전의 의미 분포와 매우 유사한 결과를 보이는 것을 확인할 수 있었다.

1. 서 론

컴퓨터의 언어 처리 연구에서 의미처리와 이의 응용에 대한 연구는 오래 전부터 활발히 연구되어 오고 있는 분야이다. 컴퓨터에서의 의미처리를 위하여 그 기준이 되는 의미 구분은 전문가의 분석과 수작업에 의해 구축된 사전과 시소러스의 구분을 적용하여 사용하여 왔다. 그러나, 기존의 수작업 언어자원들은 그 의미 사이의 경계가 너무 세분화되어 있거나 모호하여 때로는 사람이 그 의미의 차이를 구분하는 데에도 어려운 경우가 있다. 또한, 실제로 사용되지 않는 많은 의미들이 의미구분에 포함되어 있다. 이러한 수작업 의미구분의 비현실성은 언어처리 시스템에 적용할 경우 그 정확도와 실용성을 떨어뜨린다는 지적이 있다. [11]

따라서, 본 논문에서는 대용량 코퍼스에서 추출한 공기정보에 자동 군집화 방법을 적용하여 실용적인 의미분포를 추출하는 방법을 제안한다.

2장에서는 본 연구에 대한 관련연구를, 3장에서는 의미분포의 추출대상인 공기정보에 대하여 정의하고, 4장에서는 자동 군집화 방법에 대하여 소개한다. 마지막으로, 5장에서 실험결과와 분석결과를 설명한 후 6장에서 결론을 맺는다.

2. 관련 연구

의미구분의 비현실성을 극복하기 위하여 코퍼스의 의미분포를 분석하여 실제 사용되는 의미의 통계적 분석을 기반으로 코퍼스 기반의 사전들이 구축되고 있다. 그러나, 이러한 코퍼스 기반의 사전들은 여전히 수작업 중심으로 이루어지고 있어서 구축비용과 시간이 지나치게 많이 요구된다.

이러한 수작업 사전 구축 작업의 높은 구축 비용을 줄이고 코퍼스로부터 실제적인 의미를 추출하기 위하여, 대용량코퍼스에서 추출한 공기 정보를 활용하고자 하였다. Wortchartz는 '함께 등장하는 공기정보가 단어의 의미를 나타낸다'는 가정 하에 구축된 공기정보 사전이다 ([3], [4]). 이 사전은 같은 의미의 공기정보는 언어공통적인 패턴을 가지고 있다는 사실을 증명하였다. 또한, 공기정보의 유사도를 측정하여 다의어의 의미경계를 구분할 수 있을 것이라는 가능성을 보여준다.

[5]은 대용량 코퍼스에서 추출한 공기정보에 제안하는 CBC (Committee Based Clustering) 방법을 사용하여 의미경계를 추출하고자 하였고, 객관적인 실험결과를 통하여 정당성을 증명하고자 하였다. 그러나, 이 연구는 사용한 대상 공기정보를 코퍼스 빈도수에 의해 선택한 1000개의 명사로 제한하였다. 본 논문에서는 좀 더 정확한 문맥 패턴의 추출방법을 제안하였다.

이 밖에도 [1]는 정보검색의 성능향상을 위하여 베이지안 네트웍을 사용하여 공기정보 지도를 구축하여 복합명사와 고유명사 애매성해소에 적용하여 성능향상에 기여하였다. [6]은 기계번역의 번역어 선택의 의미 애매성 해소에 초점을 두고 다국어 공기정보를 사용하여 단어의 의미경계를 자동으로 추출하는 연구를 제안하였다.

3. 의미경계 추출을 위한 공기정보

3.1 사전 의미분포의 비현실성

다음은 대표적인 다의어인 '전자'라는 단어의 우리말 큰사전[2]의 의미분포를 보여준다.

1. 농부

1 이 논문은 과학기술부, 과학재단의 지원에 의하여 이루어짐

2 한글학회, 어문각, 1997

2. 전모
3. 지난번. 두 가지 사물이나 사람을 들어 말하였을 때, 먼저 든 사물이나 사람
4. 거리낌이 없이 제 마음대로 내키는 대로 함
5. 한 원자 속에 음전기를 띠고 원자핵의 둘레를 도는 소립자의 한가지
6. 한자의 글씨체의 한가지. 대전과 소전의 두 가지.
7. 어떤 나라 말을 그 적힌 소리대로 그 나라 글자로 맞추어 씀
8. 전자기의 준말

8가지 의미 중 실제로 많이 사용하는 의미는 3, 5, 8 번 의미이고, 나머지 의미들을 사용되고 있지 않은 의미이다. 실제로 많이 사용되는 의미번호가 현실과 동떨어져서 의미분포에서 3번 이후로 뒤에 등장하는 현재의 분포도 문제가 된다. 또한, 3번과 8번 의미의 경우는 '음전기를 띠는 소립자'라는 의미의 범위가 상당 부분에서 겹치기 때문에, 의미구분에 적용할 경우 사람이 결정하는 경우에도 어려움이 있을 것이다.

코퍼스의 의미분포를 분석하여 구축한 연세사전에 수록된 '전자'의 의미는 다음과 같다.

1. 앞에서 말한 두 가지의 사물이나 사람 가운데에서, 앞에 먼저 말한 사물이나 사람.
2. 한 원자 속에서 음전기를 띠고 원자핵의 둘레를 도는 기본적 소립자의 한가지. 질량이 매우 작으면서 안정되어 있으며, 자기적인 성질을 가지고 있음.
3. 획의 굵기가 일정하며, 곡선이 부드럽고, 글자가 전체적 균형을 이루며, 한자 특유의 상형적 성질을 암시하는 특성이 있는, 한자 붓글씨체.

이처럼 전문가에 의해 수작업으로 구축한 사전의 의미분포는 코퍼스에서 등장하는 의미분포와, 또한 실제로 코퍼스의 의미를 분석하여 수작업으로 구축한 사전과도 큰 차이를 보이고 있다. 이러한 사전 의미분포의 비현실성은 코퍼스를 통한 학습으로 코퍼스의 단어를 의미구분 해야 하는 의미구분시스템과 언어처리 시스템에 그대로 적용하기에 어려움이 있다.

3.2 공기정보의 다의성

다음은 ' 전자' 의 공기정보를 코퍼스에서 추출한 결과의 일부이다.

전하량, 전자공여력, 연결성, 열역학, 표, 책임성, *원소*, 종교영역, 북한, …

이탤릭 단어만이 연세사전의 2번 의미로 '전자'가 사용되었을 경우 함께 등장하는 공기정보들이다. 그러므로, 이 공기정보를 그대로 언어처리에 적용한다면 나머지 문맥정보들은 정확한 문맥정보의 반영에 방해요소로 작용하게 된다. 그러므로, 공기정보의 정확한 사용을 위하여, 식 (1)과 같이 같은 의미에서 사용한 x의 문맥정보들을 군집으로 묶어내는 작업이 필요하다.

$$g \circ f(x,w,c) = \begin{Bmatrix} \langle g(x_1^{-w}),\ldots, g(x_1^{-1}), g(x), g(x_1^{+1}),\ldots, g(x_1^{+w})\rangle \\ \ldots\ldots\ldots \\ \langle g(x_m^{-w}),\ldots, g(x_m^{-1}), g(x), g(x_m^{+1}),\ldots, g(x_m^{+w})\rangle \end{Bmatrix} \quad (1)$$

f 는 코퍼스 c에서 공기정보 추출의 대상인 대상단어 x[3]의 공기정보를 추출범위 w로 추출하여 벡터화하는 함수이다. 즉, m개의 문맥 중, x_i^{-j}는 x와 함께 공기하는 i번째 문맥에서 왼쪽으로 j번째 단어이다. 본 논문에서 공기정보 추출의 범위 w는 x과 같은 문장으로 한정한다. g는 f 에서 만들어진 문맥벡터를 적합한 의미로 연결하는 함수를 나타낸다. 이 과정에서 단어 w의 코퍼스로부터의 의미분포 또한 추출할 수 있게 된다.

본 논문에서는 의미와 문맥벡터와의 연결을 위한 함수 g를 위하여 여러 가지 자동 군집화 방법을 사용하여 구분하는 방법을 제안한다.

4. 공기정보의 자동 군집화

본 연구의 공기정보를 사용한 의미경계 추출을 위한 가정은 다음과 같다. 같은 의미의 단어와 함께 등장하는 문맥정보는 서로 비슷한 패턴을 보인다. 그러므로, 대상단어의 각 의미 별 공기정보 - 대상단어의 문맥정보 - 를 문맥정보의 문맥정보 - 본 논문에서는 '주변단어'라고 정의한다 - 의 패턴을 분석하여 유사한 문맥정보끼리 군집화 한다면, 의미 별로 정확한 공기정보를 추출할 수 있다. 또한, 이 과정에서 대상단어의 정확한 의미구분과 그 경계를 코퍼스의 공기정보를 통하여 발견할 수 있다.

4.1 문맥정보의 정규화

군집화 작업을 위하여 추출한 공기정보는 정규화 과정을 거친다. 정규화 과정이란, 추출한 주변단어들 중 의미적 특징이 불분명한 정보들을 제거하는 작업을 의미한다.

본 연구에서는 두 가지 정규화 방법을 적용하여서 공기정보의 의미적 관련성을 표면화하여서 관련성이 부족한 문맥정보를 제거하였다.

첫째, tf• idf값을 사용하여 의미적으로 중요하지 않은 공기정보들을 제거하였다. 추출한 공기정보들 중에 의미에 상관없이 자주 등장하는 고빈도 단어, 지나치게 많은 의미를 가진 다의어, 문법적 성격이 강한 의존명사 등은 의미경계 추출을 위한 군집화 작업에 방해요소로 작용한다. tf는 단어가 문서에 등장하는 빈도수이고, idf는 해당 단어가 등장하는 문서의 개수의 역수이다. 그러므로, tf• idf값이 큰 단어들은 의미에 상관없이 다양한 문서에 출현하는 단어임을 나타내므로, 문맥의 의미 별 등장패턴과는 관련이 적다. 의미적 연관성이 없는 정보들을 제거하여 정확한 군집의 추출을 기대할 수 있다.

둘째, 의미적 연관성이 적은 문맥단어 별 주변단어들을 은닉의미색인 (Latent Semantic Indexing)을 통하여 그 내포된 의미적 특징을 표면에 나타내고자 한다.

4.2 자동 군집화 방법

대상단어의 문맥정보들을 정규화 과정을 거친 주변단어들로 벡터화 한다. 이 결과를 다양한 군집화 방법을 적용하여 그 결과를 분석해 보고자 한다.

군집화 방법은 주변단어 벡터들 사이의 유사도 비교를 통하여 주변단어 벡터가 유사한 문맥단어들을 같은 군집으로 구분한다. 군집화를 통하여 추출한 군집들은 대상단어의 코퍼스에 등장하는 각 의미 별 문맥정보이다.

본 논문에서 적용한 군집화 방법은 아래와 같다.

- K: K-평균 군집화 방법. (*K-means clustering*) [2]
- B: 휴리스틱으로 초기값을 결정하는 K-평균 군집화 방법. (Buckshot) [7].
- C: 커미티 기반 군집화 (Committee based clustering) [5]
- M1, M2: 마르코프 군집화 (Markov clustering). 그래프기반 군집화 방법. 그래프 표현방법[4]에 따라 두 가지로 분류. ([8], [9])
- F1, F2: 퍼지 군집화 (Fuzzy clustering) [10]

5. 실험 및 분석

대용량 코퍼스에서 추출한 공기정보에서 대표적인 다의어들을 대상으로 본 논문에서 제안하는 군집화 방법을 적용하여 군집을 추출하였다. 추출한 군집의 개수와 기존의 사전의 의미분포를

3 본 논문에서 추출한 공기정보에 포함되는 어휘는 문장 내에 등장하는 실질어인 명사, 형용사, 동사로 한정한다.

4 M1:주변단어의 공기 여부, M2:M1+주변단어의 유사도 (0.5 이상)

비교하여 그 차이점을 분석하였다. 실험에 사용한 데이터는 다음과 같다.

- 영어 코퍼스: Penn tree bank[5]
- 영어 비교 사전: WordNet 1.5[6] (수작업 의미구분 결과인 정교한 구분 (W-F)과 코퍼스 기반 의미구분인 개략적인 의미구분 (W-C)의 두 가지 의미구분 정보를 제공한다.)
- 한국어 코퍼스: KAIST 코퍼스[7]
- 한국어 비교 사전: 우리말 큰사전 (K-D : 수작업 사전), 연세사전[8] (Y-D: 코퍼스 기반 수작업 사전)

5.1 실험 결과

위 데이터를 통한 영어와 한국어의 군집화 실험결과는 다음과 같다. 실험의 대상 다의어는 SENSEVAL2에서 의미구분 대상 단어로 제공하고 있는 영어 393개 단어와 한국어 20개 단어의 실험결과이다. [11]

	K	B	C	F1	F2	W-C	W-F
의미수	1392	1434	2222	1468	2006	1603	4571
평균 의미수 (명사)	3	3.26	6.01	3.92	4.04	3.26	8.94
평균 의미수 (형용사)	3.05	3.22	3.22	2.23	5.05	2.89	8.6

표 1 영어 데이터의 실험 결과

	K	B	C	F1	F2
의미수	35	35	66	34	31
명사	2.917	2.917	5.5	2.833	2.583

	M1	M2	K-D	Y-D
의미수	49	46	135	40
명사	4.083	3.833	11.25	3.333

표 2 한국어 데이터의 실험 결과

두 번째 실험은 군집화 결과의 군집 내의 요소들이 얼마나 정확하게 구분되었는지를 알아보았다. SENSEVAL2[9]에서 구축된 의미구분 데이터에서 의미 별 대상단어의 공기정보를 추출한다. 추출한 의미 별 공기정보와 자동 군집화 결과의 일치율은 식 (2)와 같이 구한다

$$\text{일치율} = \frac{1}{N}\sum \frac{\text{일치하는 문맥의 수}}{(\text{군집의 문맥} \cap \text{의미별 공기정보의 문맥}) \text{의 수}} \quad (2)$$

N : 실험대상 다의어의 수

식 (2)에서 분모는 군집화 결과의 문맥과 의미구분 코퍼스에서 추출한 공기정보와의 교집합의 원소의 개수를 의미한다. 의미구분 코퍼스와 공기정보 추출 코퍼스가 완벽하게 일치하지 않기 때문에, 이들 사이에 공통으로 등장하는 원소만을 대상으로 일치율을 평가하였다. 즉, 일치율은 두 군집 사이의 공통 원소들의 군집 일치율을 의미한다.

두 결과의 가능한 모든 군집 쌍 간의 일치율을 계산하여, 가장 높은 일치율을 보이는 사전의미와 군집을 매핑한 후, 문맥의 일치율을 조사한 결과는 표 3과 같다.

	K	B	C	F1	F2
일치율	98.666	98.578	90.91	97.316	88.333

표 3 군집화 결과와 의미 별 공기정보와의 문맥 일치율

5.2 결과 분석

표 1과 표 2의 실험 결과에서 보듯이, 기존의 세분화된 사전과 자동 군집화 결과는 다른 결과를 보여준다. 또한, 본 논문에서 제안한 방법에 의한 군집화 결과는 코퍼스 기반 사전의 의미구분 결과와 더 유사하다. 그러므로, 본 논문에서 제안한 방법이 코퍼스로부터 자동으로 실제적인 의미경계를 추출할 수 있다는 것을 알 수 있다.

6. 결론

본 논문에서는 자동 군집화 방법과 코퍼스에서 추출한 공기정보를 사용하여 다의어의 의미경계를 자동으로 추출하는 방법을 제안하였다. 코퍼스 기반 사전과 의미구분된 코퍼스로부터 추출한 문맥정보와의 비교를 통하여 본 논문에서 제안한 방법이 적은 비용으로 실질적인 다의어의 의미분포를 코퍼스에서 추출할 수 있음을 알 수 있다.

향후 연구로는 자동 의미추출 방법을 시소러스와 사전의 자동 구축에 적용하기 위한 연구가 필요하다.

참고 문헌

[1] Young C. Park and Key-Sun Choi, "Automatic Thesaurus Construction Using Bayesian Networks", Information Processing and Management, 1996

[2] Ray S. and Turi R. H., "Determination of Number of Clusters in K-means Clustering and Application in Colour Image Segmentation", Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999

[3] Heyer G., Quasthoff U., Wolff C., "Information Extraction from Text Corpora", IEEE Intelligent Systems and Their Applications, 2001

[4] Läuter M., Quasthoff U., Wittig T., Wolff C., Heyer G., "Learning Relations using Collocations", IJCAI, 2001

[5] Patrick Pantel and Dekang Lin., "Discovering Word Senses from Text", Proceedings of ACM Conference on Knowledge Discovery and Data Mining, 2002

[6] Hyungsuk Ji,Sabine Ploux and Eric Wehrli., "Lexical Knowledge Representation with Contexonyms", Proceedings of the 9th Machine Translation,2003

[7] Eric C. Jensen,Steven M. Beitzel,Angelo J. Pilotto,Nazli Goharian,Ophir Frieder, "Parallelizing the Buckshot Algorithm for Efficient Document Clustering", Proceedings of the 2002 ACM International Conference on Information and Knowledge Management , 2002

[8] Stijn van Dongen, "A cluster algorithm for graphs", Technical Report INS-R0010,National Research Institute for Mathematics and Computer Science in the Netherlands, 2000

[9] Stijn van Dongen, "A stochastic uncoupling process for graphs", Technical Report INS-R0011, National Research Institute for Mathematics and Computer Science in the Netherlands, 2000

[10] Song D.,Cao G. and Bruza P. D., "Fuzzy K-means Clustering in Information Retrieval.", DSTC Technical Report, 2003

[11] Philip Edmonds, "Introduction to Senseval", ELRA Newsletter, October 2002.

[12] 신사임, 이운재, 서충원, 최기선, "k-평균 군집화 방법을 사용한 공기정보의 의미기반 군집화", 한국인지과학회, 2003

[5] http://www.cis.upenn.edu/~treebank/home.html

[6] http://www.cogsci.princeton.edu/~wn/

[7] http://kibs.kaist.ac.kr/, 1999-2003

[8] http://clid.yonsei.ac.kr:8000/dic/default.htm, 2003