어휘 간의 관계를 고려한 중의성 해소 방법 및 수화 생성 시스템에의 응용

김상철*[1] 박광현[2] 변증남[1]
한국과학기술원 전자전산학부[1]
광운대학교 정보제어공학과[2]
sangchul@kaist.ac.kr akaii@kw.ac.kr zbien@ee.kaist.ac.kr

Word Sense Disambiguation Considering Words Relations and Its Application to Sign Language Generation System

Sangchul Kim[1] Kwang-Hyun Park[2] Zeungnam Bien[1]
Department of Electrical Engineering and Computer Science, KAIST[1]
Department of Information and Control Engineering, Kwangwoon University[2]

요 약

어휘 중의성 해소는 다양한 분야에서 중요한 위치를 차지하고 있는 자연어처리 관련 문제이다. 본 논문에서는 어휘 중의성 해소의 실제 응용과 관련하여 문헌 내에 여러 개의 중의어가 존재할 때의 중의성 해소 문제를 다룬다. 기존의 연구에서는 다루지 않았던 다수의 중의어가 존재할 때의 중의성 해소 문제에 대하여 고찰한 후, 중의어 간의 연관성을 이용한 중의성 해소 개념을 제시한다. 또한 이를 이용한 구체적인 해소 방안 제안 및 본 연구의 한국어-한국수화 번역 시스템에의 응용 예를 소개한다. 결론 및 향후 과제에서는 본 논문에서 언급된 방법의 향후 개선 방안에 관하여 언급한다.

1. 서론

어휘 중의성 해소는 전산 언어학의 핵심적인 연구 문제 중 하나로 문장 내에 존재하는 중의성을 갖는 단어가 어떤 의미로 사용되었는지 다양한 언어 정보원을 이용하여 파악하는 과정을 말한다[1]. 어휘 중의성 해소는 기계 번역, 인터넷 검색 등 전산 관련 분야 및 생물 정보학 등 다양한 응용 분야에서 활발하게 이용되고 있다[2][3].

이러한 어휘 중의성 해소에 대한 활발한 연구가 있어왔으며 많은 방법론들이 제시되었다[4][5]. 이러한 연구 결과를 실제로 응용하기 위해서는 일반적인 문서 자료에서의 중의어 존재 특성을 고려할 필요가 있다. 많은 문서에서 한 문장당 1개 이상의 중의어가 발견되며 특히 수 개의 문장을 고려하면 등장하는 중의어의 수는 대개 2개 이상이다. 이와 같이 여러 개의 중의어가 존재하는 경우에 관한 중의성 해소 연구가 존재하나 중의어 간의 관계와 관련된 문제를 다루고 있지는 않다[6][7].

본 논문에서는 이와 관련하여 다수의 중의어를 포함한 문서 단위가 중의성 해소의 대상이 되는 경우에 중의어 간의 관계를 고려한 중의성 해소 문제를 다루고 있다. 즉, 다수의 중의어가 존재하는 경우 중의어 간의 상호 관계가 중의성 해소의 성능에 영향을 미칠 수 있음을 고려하여, 이러한 문제를 체계적으로 정리하고 구체적으로 문제화하여 가능한 해결 방안을 제안한다. 또한 본 연구의 구체적인 시스템 응용 예를 소개하고, 결론 및 향후 과제에서는 본 논문에서 언급된 방법의 특성을 논의하며 향후 개선 방안에 관하여 언급한다.

2. 다수의 중의어가 존재할 때의 중의성 해소 문제

2.1. 기존의 방식 및 한계

다수의 중의어를 포함한 문서 단위가 중의성 해소의 대상이 되는 경우에 대한 기존의 연구에서는 단일 중의어의 중의성 해소 방법과 유사한 방식을 적용하며 그 중의성 해소 순서에 관한 특별한 고려를 하고 있지 않다. 또한 하나의 중의어의 의미를 결정하였을 때 그 중의성이 해소된 중의어가 다른 중의성 미해소 중의어에 미치는 영향에 대한 명시적인 고려가 존재하지 않는다[6][7].

하지만 다수의 중의어가 존재하는 경우 그 중의어 사이에는 언어학적 특성에 따른 유의미한 상호 관계가

있으며 이는 중의어들의 중의성 해소에 있어 중요한 정보가 될 수 있다. 이러한 상호 관계에 대한 정보를 고려하지 않는 경우 주어진 정보를 최대로 이용할 수 없다는 한계가 존재하며 특히 대상 문서 단위 내에 여러 개의 중의어가 나타나는 경우 문제의 중요성은 더 부각된다.

2.2. 중의어 간 상호 관계로부터 얻을 수 있는 정보

중의어 간 상호 관계로부터 얻을 수 있는 정보는 구체적으로 다음과 같다.

먼저 하나의 중의어의 의미를 확정하게 되면 그 중의어의 확정된 의미가 다른 중의어의 중의성 해소를 위한 단서가 될 수 있다. 이때 어떤 중의어의 중의성을 먼저 해소하는지에 따라 이후의 중의성 과정에서 이를 유용하게 이용할 수 있을지의 여부가 결정된다. 가령 특정 중의어의 경우 그 중의성 해소의 난이도는 높으나 다른 중의어와 큰 관련이 없을 수도 있으며 이 경우 이 중의어의 중의성을 먼저 해소하는 것은 그 자체로 부정확할 수 있으며 또한 다른 중의어의 중의성 해소에도 도움이 되지 않는다. 반대로 어떤 중의어의 경우 그 용례가 특정 의미에 대부분 한정되어 있을 뿐만 아니라 다른 중의어들과 밀접한 연관이 있을 수 있으며 이때는 이 중의어의 중의성을 먼저 해소하는 것이 유리한 선택이 된다.

이와 같이 중의어 간의 상호 관계를 이용하여 다른 중의어의 중의성 해소 과정에 도움을 받을 수 있으며 이를 극대화시키기 위해서는 적절한 중의성 해소 순서의 결정이 요구된다. 또한 해당 중의어의 중의성 해소 난이도에 따라 의미 가정 여부를 결정하는 것이 전체 중의어 관점에서 볼 때 최대의 정확도를 얻을 수 있는 전략이다.

이러한 중의어 간의 상호 관계를 이용하지 않는 경우 유용한 정보의 손실이 발생하게 되며 특히 중의어의 빈도가 매우 높은 경우에는 정보의 부족을 야기하게 된다.

3. 상호 연관성을 이용한 중의성 해소 개념

이제 2.2절의 내용을 보다 구체화시켜 실제로 중의성 해소 과정에서 중의어 간의 상호 관계가 어떤 형태로 이용될 수 있는지를 논의한다.

3.1. 중의성 해소 순서에 대한 고려

앞서 설명한대로 여러 개의 중의어가 존재하는 경우 어떤 중의어의 중의성을 먼저 해소할 것인지의 문제는 나머지 중의어의 중의성 해소 성능에 영향을 미친다.

중의성 해소의 순서를 결정하기 위하여 다음과 같은 조건을 고려할 수 있다. 선 순위로 중의성이 해소되는 중의어의 경우 그 중의성 해소 과정이 다른 중의어들에 비해 용이하여야 한다. 사전 확률의 관점에서 볼 때 이러한 중의어들은 가능한 여러 개의 의미 중 보통 하나의 특정 의미로 사용될 확률이 높은 중의어가 된다.

또한 주변의 중의어가 아닌 단서들이 해당 중의어의 중의성을 해소하는데에 결정적인 영향을 주는 경우 또한 중의성 해소 과정이 용이하게 된다. 또한 해당 중의어의 중의성을 해소했을 때 다른 중의어들의 의미가 얼마나 명확해지는지의 여부도 중요한 고려 사항이 된다. 특정 중의어의 중의성을 해소하였을 때 나머지 중의어들의 조건부 의미의 확률 분포가 중의성을 해소하기 쉬운 방향으로 변화한다면 해당 특정 중의어는 우선적으로 중의성 해소가 되어야 한다.

3.2. 중의어의 의미 가정 여부에 대한 고려

다수의 중의어에 대하여 특정 중의어의 의미를 가정하기 위해서는 다음과 같은 사항을 고려하여야 한다.

대상 문서 내에 존재하는 중의어들 중 그 의미를 가정하기 어렵거나 다른 중의어의 중의성 해소에 영향을 주지 않는 단어의 경우 의미를 가정하는 것이 전체 중의성 해소 성능에 도움이 되지 않는다. 오히려 해당 중의어의 의미를 부정확하게 가정함으로써 다른 중의어의 중의성 해소에 부정적인 영향을 끼칠 수 있게 된다. 때문에 중의어의 의미 가정 여부에 관한 고려는 신중하게 이루어져야한다.

이와 관련하여 하나의 대안으로 중의어의 의미를 확정짓지는 않되 그 의미를 확률을 고려하여 가정하게 되면 의미 가정으로 인한 부작용을 경감할 수 있게된다.

4. 순서 및 의미 가정 여부를 고려한 중의성 해소 방법

본 절에서는 앞서의 논의를 토대로 중의성 해소 순서 및 중의어의 의미 가정 여부를 고려한 중의성 해소 방안을 제시한다.

사용되는 기호의 정의는 다음과 같다.

N: 중의어의 총 수

w_i: i번째 중의어. $i=1, \cdots, N$

M_i: 중의어 w_i가 가지는 의미의 개수

w_{ij}: 중의어 w_i의 j번째 의미. $j=1, \cdots, M_i$

$P_{w_{ij}}$: 중의어 w_i가 의미 w_{ij}를 가질 확률. 즉, $P_{w_{i1}} + \cdots + P_{w_{iM_i}} = 1$.

σ_{w_i}: w_i에 대하여 각 의미별 확률 $P_{w_{i1}}, \cdots, P_{w_{iM_i}}$의 표준편차.

I: 가장 큰 표준편차를 가지는 w_i의 인덱스 i. 즉, $\sigma_{w_I} = \max(\sigma_{w_1}, \cdots, \sigma_{w_N})$

J: 중의어 w_i에 대하여 $P_{w_{ij}}$가 가장 큰 값을 가질 때의 j 값. 즉, $P_{w_{iJ}} = \max(P_{w_{i1}}, \cdots, P_{w_{iM_i}})$.

$\sigma_{w_i, w_{IJ}}$: w_I를 의미 J로 가정하고 나머지 w_i, $i \neq I$ 에 대하여 구한 σ_{w_i}.

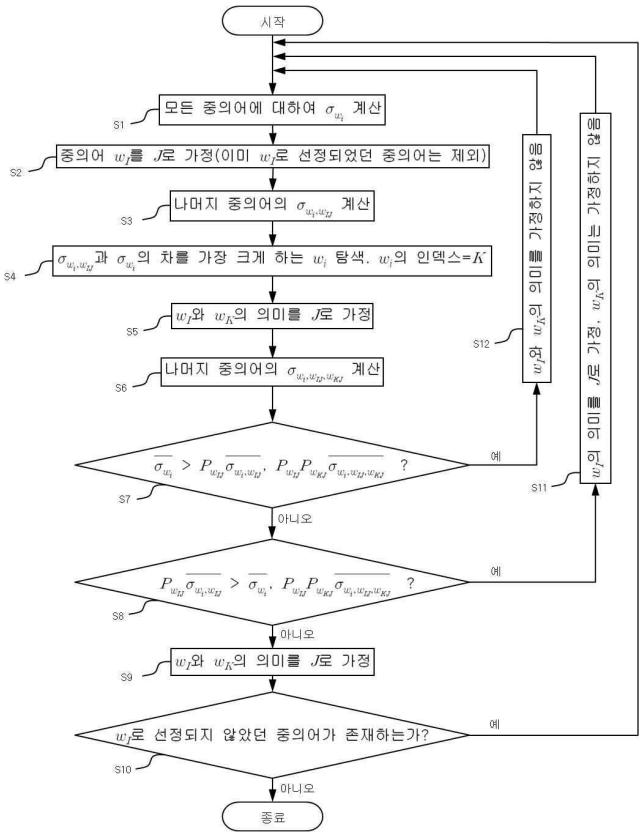

그림 1. 복수 중의어의 중의성 해소 과정

구체적인 순서는 다음과 같다. 그림 1은 이후 설명할 내용을 순서도로 나타낸 것이다. 먼저, 모든 중의어에 대하여 각각 σ_{w_i}를 계산한다(S1). 이 때 다른 중의어들은 문장 내에 존재하지 않는다고 가정한다. 표준편차가 클 수록 어떤 특정 의미일 확률이 높다는 뜻이므로 중의어의 올바른 의미를 유추하는 것이 더욱 용이하다. 표준편차의 값을 구하는 방법은 여러 가지가 있을 수 있다. 가장 간단한 방식으로 사전에 조사된 중의어들의 사용 빈도 정보표를 이용하여 사전 확률의 분포를 구할 수 있다. 그림 2는 그러한 정보표의 예이다[8]. 가령 어떤 특정 중의어의 각 의미에 대한 사용 빈도가 서로 비슷하게 나타난다면 표준 편차의 값은 작을 것이고, 하나의 의미로 치우쳐 있다면 반대의 결과가 나타나게 된다.

차례	항목	품이	품사	빈도	계수	교재	교과	교양	문학	신문	잡지	대본	구어	기타
16399	가01		명	7	3	2	0	0	0	0	4	0	0	1
10933	가01	우물 -	명	13	10	2	0	1	4	0	2	1	0	3
10934	가07	家	명	13	3	0	0	12	1	0	0	0	0	0
38207	가80	假	명	1	1	0	0	1	0	0	0	0	0	0
25486	가가03	假家	명	3	2	0	1	0	2	0	0	0	0	0
30138	가가대소	呵呵大笑	명	2	1	0	0	0	0	0	0	0	0	2
38208	가가대소하다	呵呵大笑 -	동	1	1	0	0	0	0	0	0	0	0	1
38209	가가소소하다		형	1	1	0	0	1	0	0	0	0	0	0
30139	가가호호	家家戶戶	부	2	2	0	0	0	0	2	0	0	0	0
22315	가감01	加減	명	4	3	0	0	0	1	3	0	0	0	0
38210	가객02	歌客	명	1	1	0	0	0	0	1	0	0	0	0
38211	가가거거		부	1	1	0	0	0	0	1	0	0	0	0
38212	가가날		명	1	1	0	0	0	0	1	0	0	0	0
38213	가거주	假居住	명	1	1	0	0	0	0	1	0	0	0	0
22316	가건물	假建物	명	4	3	0	0	0	1	1	2	0	0	0
30140	가검물	可檢物	명	2	1	0	0	0	0	2	0	0	0	0
1195	가게		명	192	50	34	11	9	37	12	20	51	2	16
17945	가곤방	-房	명	6	4	1	0	0	2	0	1	2	0	0
999	가격03	價格	명	250	45	8	18	70	18	109	25	0	0	2
25487	가격하다	加擊-	동	3	3	0	0	1	0	1	0	1	0	0
30141	가결하다01	可決-	동	2	1	0	0	0	0	2	0	0	0	0
14013	가계06	家系	명	9	7	0	0	6	0	1	2	0	0	0
9885	가계08	家計	명	15	8	0	0	1	1	2	10	0	0	1
30142	가계부	家計簿	명	2	2	0	1	0	0	1	0	0	0	0
38214	가계비	家計費	명	1	1	0	0	1	0	0	0	0	0	0
30143	가계성	家計性	명	2	1	0	0	0	0	2	0	0	0	0
22317	가계약	假契約	명	4	1	0	0	0	0	4	0	0	0	0

그림 2. 단어의 사용 빈도 정보표

그 다음, 가장 큰 표준편차 값을 가지는 중의어 w_I에 대하여 그 의미를 사전 확률이 가장 높은 J로 가정한 후(S2), 나머지 중의어들의 $P_{w_{i1}}, \cdots, P_{w_{iM_i}}$ 및 그 표준편차 값 $\sigma_{w_i, w_{IJ}}$를 계산한다(S3). $\sigma_{w_i, w_{IJ}}$는 w_I의 의미가 주어졌다고 가정하고 구하는 값이므로 초기의 σ_{w_i}와는 다른 값을 가진다. 이때 새로운 확률들은 조건부 확률이 된다.

이렇게 조건부 표준편차 $\sigma_{w_i, w_{IJ}}$를 구하고 나면, 각 중의어별로 $\sigma_{w_i, w_{IJ}}$와 σ_{w_i}의 차이를 구하여 그 차이가 가장 큰 중의어 w_K를 선정한다. $\sigma_{w_i, w_{IJ}}$와 σ_{w_i}의 차이가 크다는 것은, w_I의 의미를 가정하였을 때 w_K의 의미별 확률 간의 편차가 w_I의 의미를 가정하지 않았을 때와 비교하여 증가한다는 것을 의미한다. 이는 중의어 w_I와 w_K가 서로 연관성이 높다는 것을 나타내는 것이며 또한 w_I의 의미가 J일 가능성이 높다는 것을 시사하는 것이기도 하다. 중의어들 간에 서로 연관성이 없다면, 어느 한 중의어의 의미별 확률은 다른 중의어의 의미를 가정하는

지 여부에 상관없이 변하지 않는다. 이는 확률의 독립성에 기인한다. 즉, 서로 독립인 두 확률변수 X, Y 경우, X의 확률 P(X)는 Y를 임의의 값으로 가정했을 때의 x의 확률 P(X|Y)와 같다(즉, P(X)=P(X|Y)). 따라서, 중의어 w_I의 의미를 J로 가정하였을 때, w_I와 연관성이 있는 중의어는 그 의미별 확률분포가 변하고, 연관성이 없는 중의어는 그 의미별 확률분포가 변하지 않게 된다. 따라서, 이러한 방식을 통해 서로 연관성이 있는 중의어를 선별할 수 있는 것이다.

w_I와 w_K의 의미를 각각 해당 중의어에서 가장 큰 확률을 가지는 의미인 J로 가정한 후(S5) 다시 나머지 중의어들에 대한 σ_{w_i}의 합을 구한다. 이때의 표준편차를 $\sigma_{w_i,w_{IJ},w_{KJ}}$라 표기한다(S6).

다음으로, σ_{w_i}, $\sigma_{w_i,w_{IJ}}$, $\sigma_{w_i,w_{IJ},w_{KJ}}$의 평균값을 각각 구하고, 이를 각각 $\overline{\sigma_{w_i}}$, $\overline{\sigma_{w_i,w_{IJ}}}$, $\overline{\sigma_{w_i,w_{IJ},w_{KJ}}}$ 로 표기한다. 세 값을 비교함에 있어 $\overline{\sigma_{w_i,w_{IJ}}}$와 $\overline{\sigma_{w_i,w_{IJ},w_{KJ}}}$는 각각 w_I와 w_K의 의미를 가정한 상태에서 구한 것이므로, 이로 인한 불확실성을 반영하기 위하여 각각의 값에 $P_{w_{IJ}}$와 $P_{w_{IJ}}P_{w_{KJ}}$를 곱한다. 즉, $\overline{\sigma_{w_i}}$, $P_{w_{IJ}}\overline{\sigma_{w_i,w_{IJ}}}$, $P_{w_{IJ}}P_{w_{KJ}}\overline{\sigma_{w_i,w_{IJ},w_{KJ}}}$ 의 세 값을 비교한다(S7~S12).

만약, $\overline{\sigma_{w_i}}$이 가장 큰 경우 w_I와 w_K의 의미를 가정하지 않는다(S12). 만약, $P_{w_{IJ}}\overline{\sigma_{w_i,w_{IJ}}}$이 가장 큰 경우 w_I의 의미를 J로 가정하고, w_K의 의미는 가정하지 않는다(S11). 만약, $P_{w_{IJ}}P_{w_{KJ}}\overline{\sigma_{w_i,w_{IJ},w_{KJ}}}$이 가장 큰 경우 w_I와 w_K의 의미를 각각 J로 가정한다(S9).

한국어의 언어학적 특성상 세 개 이상의 중의어가 동일 문장 내에서 밀접한 관계를 가질 확률은 적으므로 한 번 선정되었던 중의어 w_I를 w_I 후보에서 제외하고 나머지 중의어에 대하여 새로이 같은 과정을 거친다.

과정이 진행됨에 따라 의미가 가정되는 중의어의 개수가 증가하게 된다.

마지막까지 의미가 가정되지 않는 중의어의 경우 출현 순서대로 중의성을 해소하여 중의성 해소과정을 완료한다.

5. 제안된 중의성 해소 방식의 적용 예

제안된 중의성 해소 방식은 구체적으로 다음과 같은 시스템에 적용될 수 있다.

그림 3와 그림 4에 나타낸 시스템은 현재 개발 보완 중인 한국어-한국수화 번역 시스템으로 그 중의성 해소 성능의 개선을 위해 본 논문에서 설명한 중의성 해소 순서 및 중의어 의미 가정 여부를 고려한 중의성 해소 방법을 적용해 볼 수 있다.

그림 3. 한국어-한국수화 번역 시스템 흐름도

그림 4. 한국어-한국수화 번역 시스템의 실행 예

청각 장애인을 대상으로 하는 본 프로그램의 특성상 중의어의 의미를 명확히 파악하는 것은 다른 응용에서보다 중요하며 본 연구에서 제안된 방법의 적용은 성능 개선을 가져올 수 있다.

6. 결론 및 향후 과제

본 논문에서는 이제까지 다수의 중의어에 대한 중의성 해소 문제를 고찰하였다. 이와 관련하여 여러 개의

중의어가 대상 문서의 구분 내에 존재하는 경우 중의어 간의 상호 관계를 고려한 중의성 해소 방법을 제안하였다. 특히 본 논문은 다수의 중의어가 존재하는 경우의 중의성 해소 순서 및 의미 가정 여부에 관한 내용을 구체적으로 문제화하여 향후 연구의 주제를 제안한 것에 의의가 있다.

본 논문에서 언급된 중의어의 상호 관련성을 이용한 중의성 해소에 대한 논의 및 제시된 해결 방안은 몇 가지 측면에서 향후 추가적인 보완 연구 및 개선이 요구된다. 먼저 해결 방안의 구체적 성능 개선 정도에 대한 정량적 분석이 필요하다. 본 연구는 문헌 내에 존재하는 모든 중의어들의 중의성 해소에 대해 논의하고 있으므로 정량적 분석을 위해서는 대량의 시험 자료 구축이 선행되어야한다. 이는 향후 관련 연구자들의 협업을 통하여 가능할 것이다.

한편 4절에서 구체적으로 제시된 방안의 경우 지금까지 언급된 기본 개념들을 간단히 구체화해 본 것이므로 차후 개선의 여지가 크다. 단순히 조건부 확률에 기반한 의미 분포 특성 변화에 의존하는 것이 아니라 학습을 통한 반복적인 전략 수정 및 개선 방식을 적용해 볼 수 있을 것이다. 또한 그래프 모델(Graphical Modelling) 방법을 사용하면 중의어 간의 상호 관계를 보다 정확하게 파악할 수 있으므로 관계에 기반한 성능 향상의 개선 정도를 증가시킬 수 있게 될 것이다.

참고 문헌

[1] 이용구, 사전 정보를 이용한 단어 중의성 해소 모형에 관한 실험적 연구, 연세대학교 대학원 박사 학위 논문, pp. 1-6, 2006.

[2] Rada Mihalcea, “Word Sense Disambiguation and Its Application to Internet Search,” Master Thesis of Southern Methodist University, 1999.

[3] Martijn J. Schuemie, Jan A. Kors, Barend Mons, "Word Sense Disambiguation in the Biomedical Domain: An Overview," *Journal of Computational Biology*, Vol. 12, No. 5, pp. 554-565, 2005.

[4] Ide, N., and Veronis, J., "Word sense disambiguation: the state of the art," *Computational Linguistics*, 24(1), pp. 1-40, 1998.

[5] Agirre, E. and Edmonds, P. (Editors), Word Sense Disambiguation : Algorithms and Applications, Springer, pp. 12-18, 2006.

[6] Marisa Ulivieri, Elisabetta Guazzini, Francesca Bertagna, Nicoletta Calzolari, "Senseval-3: The Italian All-words Task," *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 2004.

[7] Veronique Hoste, Anne Kool, and Walter Daelemans, "Classifier optimization and combination in the English all words task," *Proceedings of the SENSEVAL-2 workshop*, pp. 84-86, 2001.

[8] 21세기 세종계획 산출물 [http://www.sejong.or.kr]