모듈화된 신경망을 이용한 한국어 중의성 해결 시스템

한태식, 송만석
연세대학교 전산과학과

Word sense disambiguation using modular neural networks

Tae-Sik Han, Man-suk Song
Dept. of Computersience, YonSei University

요약

문장 안에서 한 단어가 가지는 올바른 의미를 얻기 위해 모듈화된 신경망을 이용하였다. 앞부분에 놓인 신경망은 코호넨 신경망으로 사용자의 지도가 개입되지 않은 상태로 자율학습(Unsupervised learning)이 이루어지고, 뒤에 놓인 신경망은 앞에서 결과로 얻은 2차원의 자기 조직화 형상지도(Self-organizing feature map)를 바탕으로 역전파 신경망을 이용한 지도학습(Supervised learning)을 하게 하였다. 입력 자료는 구문분석된 문장의 조사 정보를 활용하여 입력 위치를 정해준 명사의 의미표지와 동사의 의미표지를 사용하였다. 중의성이 있는 단어를 가지는 문장은 중의성의 가지수 만큼 테스트 입력 자료가 되어 신경망을 통과하여 의미를 결정하도록 한다.

I. 서론

한 단어가 의미적으로 중의성을 가지고 있다는 말은 그 단어의 의미가 두 가지 이상의 의미로 대응되는 것을 말한다. 문장에서 발생한 중의성을 제거하여 올바른 의미를 얻어내기 위한 시도는 크게 선택 제약을 사용하는 방법과 통계적인 방법을 사용하는 방법으로 구분될 수 있다. 선택 제약을 사용하는 방법은 기호 주의에 바탕을 둔 방법으로 언어에 관한 지식을 규칙으로 표현하여 언어를 분석하는 방법이다. 이 방법은 분석의 범위를 넓힐수록 규칙이 늘어나게 되고, 규칙 사이의 간섭이 심하게 일어나 규칙을 확장시켜 중의성을 줄이는 것이 힘들어서 시스템 확장이 어렵다. 이런 단점을 보완하기위해 최근에 들어와서 활발히 연구되고 있는 분야가 통계를 기반으로 하는 방법들이다. 필요한 확률값을 얻어내기 위해 말뭉치에서 분석 대상이 되는 단어를 중심으로 일정한 거리에 놓인 단어들에 관한 정보를 사용하는 것이 그러한 시도들 중의 하나이며, 신경망을 자연어처리 분야에 도입한 것도 그러한 노력의 한 갈래라고 볼 수 있다.

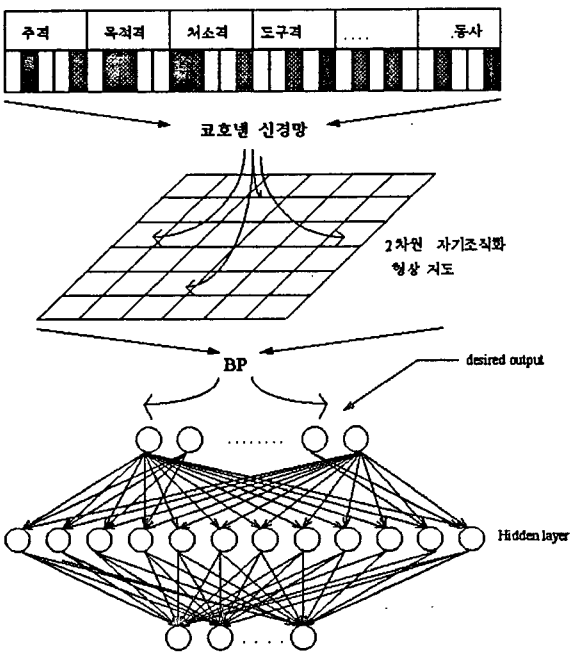
신경망을 사용함으로써 얻을 수 있는 가장 큰 장점은 기호주의 패러다임(paradigm)의 단점인 지식 습득 병목 현상(knowledge Acquisition bottleneck)을 극복할 수 있는 가능성이다.

학습에 사용되는 말뭉치가 어느 정도 언어현상을 정확히 반영해주고 있다고 볼 때, 수치로 표현된 정보를 통해 원하는 내용을 정해 학습시키는 것이 신경망 학습의 주된 작업이다.

본 논문에서는 단어의 의미 중의성을 해결하기 위해 코호넨 신경망(Kohonen-networks)과 역전파 신경망(Backpropagation networks) 두 가지로 구성된 모듈화된 신경망을 이용하였다. 말뭉치에서 중의성을 가지는 단어가 포함된 문장의 구문 분석 결과를 이용하여 조사 정보에 따라 명사의 의미 표지와 동사의 의미 표지를 붙여 학습시킨 뒤, 중의성을 지닌 단어가 들어있는 문장이 시험 문장으로 들어왔을 때 가질 수 있는 의미의 모든 경우를 고려하여 만든 시험 문장이 신경망을 거쳐 나간 결과를 토대로 의미를 결정하도록 하였다.

II. 시스템 개요

본 시스템에서 구축한 신경망의 구조는 <그림 1>과 같다.



<그림 1>

위의 <그림 1>에서와 같이 입력 자료는 처음에 코호넨 신경망을 통해 자율 학습을 하게 된다. 학습이 끝나면 각각의 문장에서 얻은 자료를 원형(proto type)으로하여 2 차원의 자기조직화 형상 지도(Self-Organizing Feature Map)를 얻게 된다. 이 2차원 지도는 단어의 의미를 원하는 결과(desired output)로하여 역전파 신경망으로 입력되어 학습을 한다.

코호넨 신경망을 사용하여 모듈화된 신경망을 사용한 이유는 코호넨 신경망이 인간의 두뇌를 가장 잘 모델링한 신경망으로 여기에서 얻을 수 있는 2 차원 지도를 중의성 구분을 위한 역전파 신경망 입력 자료로 사용하기 위해서이다. 코호넨 신경망은 자율학습과 경쟁학습 방법에 기반하는 신경망 모델로 입력 데이타 공간이 크고 복잡한 경우 가중치 벡터(weight vector)들을 통해 적절히 조절된 보다 다루기 쉬운 공간으로 대응시켜주는 성질을 가지며, 자율 학습을 하는 신경망으로 다른 신경망에 비해 빠른 속도로 학습을 할 수 있는 장점을 지닌다.

III. 학습 과정

학습에 쓰인 입력 자료는 구문 분석 결과에서 조사의 정보에 따라 위치가 결정된 명사의 의미 표지와 동사의 의미 표지로 이루어진 이진수 자료이다. 학습 자료로 사용할 문장은 말뭉치(국민학교 국어 교과서)에서 중의성을 가지는 단어를 포함하는 문장으로 한 문장 안에서 의미가 결정될 수 있으면서 구문 분석된 결과에서 조사 정보를 사용할 수 있는 것만을 대상으로 하였다. 따라서 본 시스템은 입력 데이타가 이미 구문분석이 되어 필요한 명사와 동사의 의미 표지가 분리되어 있는 상태의 자료가 입력되어 작동하도록 구성하였으며 단어의 중의성은 한 문장 내에서만 해결하도록 되어 있다.

조사 정보에 따라 결정된 위치에 할당되어야할 의미 표지는 "국어 어휘의 분류 목록에 대한 연구"[12]에서 각 명사에 준 숫자를 사용하였다. 명사 의미 표지 숫자는 모두 세자리 수로 각각의 숫자가 계층적으로 의미를 분류하여 나타내고 있다. 각 자리의 숫자에 10 bit씩 할당하여 입력 자료로 표현하였다.

동사 의미 표지는 연세 대학교 한국어 사전 편찬실의 말뭉치를 분석하여 분류한 자료[14] 를 사용하였다.

격조사 선택 때 인용격이나 호격은 동사와의 관계를 통해 중의성 해결에 도움을 주지 않는 것으로 판단되어 입력 자료 구조에서 제외시켰고 보어 성분은 동사와의 관계를 보이는데 필요해서 입력 자료 구조로 포함 시켰다.

결과적으로 입력 구조는 다음 표와 같이 14가지 항목을 가진다.

<table>
<tr><td rowspan="14">조사정보</td><td colspan="3">주격</td></tr>
<tr><td colspan="3">목적격</td></tr>
<tr><td rowspan="12">부사격</td><td rowspan="4">처소격</td><td>낙착점</td></tr>
<tr><td>출발점</td></tr>
<tr><td>향방</td></tr>
<tr><td>한계선</td></tr>
<tr><td colspan="2">도구격</td></tr>
<tr><td colspan="2">자격격</td></tr>
<tr><td rowspan="2">비교격</td><td>모양</td></tr>
<tr><td>정도</td></tr>
<tr><td colspan="2">동반격</td></tr>
<tr><td colspan="2">변성격</td></tr>
<tr><td colspan="3">보어정보</td></tr>
<tr><td colspan="3">동사정보</td></tr>
</table>

< 표 1 >

사람의 다리(leg)와 건너는 다리(bridge)의 두가지 의미를 가지는 "다리"라는 단어를 예로 학습 과정을 살펴보자. 말뭉치에서 다음의 (1)과 같이 다리에 해당되는 예문을 뽑아내었다고 하면,

(1) ... 다리를 싸맨 젊은이는 ...

이 문장은 다음의 <표 2>과 같이 의미 표지가 할당된다.

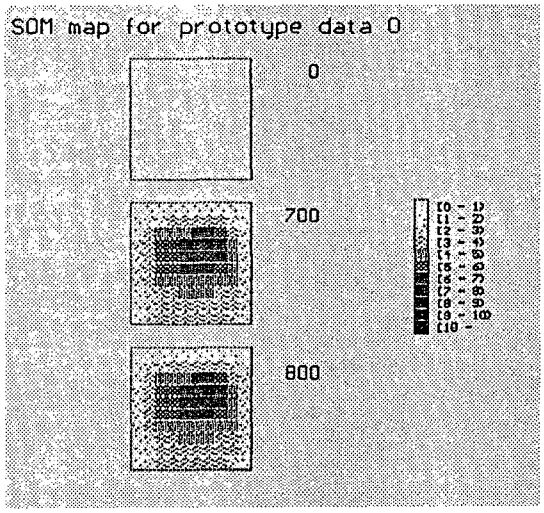
주격	목적격	...	동사
젊은이(117)	다리(132)	...	싸매다(101)

< 표 2 >

시스템은 이것을 이진수로 바꾼 숫자를 입력으로 코호넨 신경망을 통해 자율학습(Unsupervised learning)을 하며, 그 결과로 각각의 문장이 코호넨 신경망을 통해 2차원의 자기조직화 형상지도(Self-organizing Feature Maps)로 표현된다. 이 2차원 지도는 단어의 본래 의미를 나타내는 값(desired output)과 함께 역전파 신경망을 통해 지도학습(supervised learning)을 하게 된다.

국민학교 교과서로 구성된 말뭉치에서 "다리"를 포함한 문장들을 뽑아 구문 분석한 뒤 동일하게 의미 표지가 붙게 된 것은 모두 하나로 간주하여 leg의 의미를 가지는 결과와 bridge의 의미를 가지는 학습 데이타를 각각 8개씩 얻을 수 있었다.

코호넨 신경망의 2차원 지도는 10x10크기로 하였다. 말뭉치에서 얻은 "다리"에 관한 16 가지 구문 분석되어 의미 표지를 한 자료를 코호넨 네트웍을 통해 800회 학습 시킨 뒤 학습에 사용한 하나의 자료에 대한 2차원 지도를 다음 < 그림 2 >같이 관찰할 수 있었다.



< 그림 2 >

이러한 결과로 얻어진 2차원 지도 16개를 이용하여 역전파 신경망을 이용해 지도학습 시킨다. 사용한 역전파 신경망은 하나의 은닉층(hidden layer)을 가지며 입력층은 100개의 뉴런을 출력층은 5 개의 뉴런을 갖게했다. 출력층에서의 뉴런 수는 구분하고자하는 의미의 갯수 만큼을 주어야한다. "다리"의 예에서는 2 개로 충분하지만 다른 단어에서의 중의성 발생을 고려하여 5 개를 두었다. 은닉층의 뉴런의 갯수는 출력층의 뉴런의 갯수를 N이라고 했을 때 (2N+1)개로 정하도록하였다. 여기에서는 모두 11 개의 은닉층 뉴런을 두어 역전파망이 입력 자료를 2 차원지도 입력 신호에 따라 분류해내도록 하였다.

IV. 의미 결정 과정

시험 문장으로 다음 (2)와 같은 문장이 들어 왔다고 가정하자.

(2) "제비는 다리를 다쳤다."

이 경우 얻어낼 수 있는 구문 분석 정보는 "주어 + 목적어 + 서술어"이며, 목적어인 중의성을 지닌 단어 "다리"의 의미를 결정하기 위해 이미 "다리"와 관련된 문장을 학습한 신경망에 다음 <표 3>과 같이 "다리"의 모든 중의성을 고려한 데이타를 입력시킨다. 지도 학습 때 원하는 결과(desired output)로 사용하기위해 다리(leg)에 10000를 다리(bridge)에 01000을 할당하였다.

주격	목적격	...	동사	Desired output
제비(262)	다리(132)	...	다치다(102)	10000
제비(262)	다리(587)	...	다치다(102)	01000

< 표 3 >

원하는 결과를 제외한 의미 표지로만 이루어진 자료는 코호넨 신경망을 거쳐 2차원의 지도를 얻게 되고 이 지도가 역전파망을 거쳐 나온 결과가 자신의 원하는 결과(desired output)와 동일하게 나오는 경우 의미를 결정지을 수 있게 하였다. 같은 것이 두 개 이상 나오는 경우는 중의성을 해결 못한 것으로 간주했다.

V. 결론

본 시스템은 중의성 해결을 위해 인간의 두뇌를 가장 잘 모델링한 신경망 중의 하나인 코호넨 신경망과 간단한 구조로 분류(clustering)를 잘 해내는 역전파 신경망을 결합하여 사용하였다. 단순히 역전파망을 사용하여 입력 자료의 유형을 구분하기 보다는 신경망이 가지는 인지적인 특성을 반영하기 위해 코호넨 신경망을 전체 시스템의 앞 부분에 두었다.

아직은 중의성을 가진 많은 단어와 여러 가지 경우에 관한 문장들에 대한 실험이 이루어지지 못 했기 때문에 이 방법으로 더 많은 문장에 대해서도 단어의 중의성이 잘 해결될 것이라고는 말하기 힘들다. 또한 입력 자료의 구조에서 보았듯이 조사와 관련된 명사의 정보와 동사 의미 정보에만 의존하였기 때문에 다른 문장 성분의 고려가 없었다. 이러한 정보까지 고려해서 중의성 분석이 이루어져야 하겠다.

VI. 참고문헌

[1] Philippe G. Schyns. A Modular Neural Network Model of Concept Acquisition. Cognitive Science 15. 461-508.1991.

[2] Allen. Natural Language Understanding. Benjamin/Cummings Publishing Company. 227-262, 295-327. 1995.

[3] 이근배. 자연어처리에있어서 연결 주의와 기호 주의의 비교.한국 정보과학회 논문지 93. 8 Vol. 20, No. 8, August. 1993.

[4] Matthew Zeidenberg. Neural Networks in Artificial Intelligence. 195-235.Ellis Horwood Press. 1990.

[5] Garrison W.Cottrell and Steven L. Small. A Connectionist Scheme for Modeling Word Sense Disambiguation. Cognition and Brain Theory, 6(1), 89-120. 1993.

[6] David L. Waltz and Jordan B. Pollack. Massively Parallel Parsing : A Strongly Interactive Model of Natural Language Interpretation. Cognitive Science 9, 51-74. 1995.

[7] Risto Mikkulainen and Michael G. Dyer Natural Language Processing With Modular PDP Networks and Distributed Lexicon. Cognitive Science 15, 343-399. 1991.

[8] 이선정, 한상영. 신경망을 이용한 한국어 단어 범주 애매성 해소. 한국 정보과학회 논문지 제 21권 제 2호.1994.

[9] Teuvo Kohonen, The Self-Organizing Map. Proceedings of The IEEE, VOL. 78, No. 9, september. 1464-1480. 1990.

[10] Simon Haykin, Neural Networks. Macmillan. 106-200, 397-427.1994.

[11] 김대수, 신경망 이론과 응용. 하이테크정보. 169-189.1992.

[12] 임홍빈, 국어 어휘의 분류 목록에 대한 연구. 국립국어연구원.1992.12.

[13] 남기심, 고영근. 표준 국어문법론. 1993.

[14] 박영자. 자연어처리를 위한 한국어 동사,명사의 개념 분류. 제 4 회 한글 및 한국어 정보처리 학술 발표 논문집. 1992.