베이지안 네트워크를 이용한 일한 기계 번역에서의 조사 번역

Bayesian Network-based Translation of Japanese Particles in Japanese to Korean Machine Translation

指導教授　金 榮 澤

이 論文을 工學碩士學位論文으로 提出함.

1998年 10月

서울大學校 大學院
컴퓨터工學科
黃 圭 伯

黃圭伯의 工學碩士學位論文을 認准함.

1998년 12월

委 員 長	(印)
副委員長	(印)
委　　員	(印)

工學碩士學位論文

베이지안 네트워크를 이용한 일한 기계 번역에서의 조사 번역

Bayesian Network-based Translation of Japanese Particles in Japanese to Korean Machine Translation

1999年 2月

서울大學校 大學院
컴퓨터工學科

黃　圭　伯

차례

1. 서론

1.1 연구 배경

과거에 외국어로 되어 있는 문서를 접할 수 있는 가장 손쉬운 방법은 외국어로 된 인쇄물을 통하는 것이었으며 당시의 기계 번역 프로그램은 일반 텍스트 파일로 되어 있는 데이터를 대상으로 했다. 따라서 기계 번역 프로그램을 이용하여 일반 인쇄물을 번역하려면 먼저 그 인쇄물을 텍스트 파일의 형식으로 바꾸어야 했으며 이 과정의 복잡함과 부정확함으로 인하여 특수한 환경을 제외하고는 기계 번역 프로그램의 활용도가 그리 높지 못했다. 그러나 인터넷의 사용자 층이 광범위해짐에 따라 다른 언어로 되어 있는 정보에 접근할 수 있는 기회가 예전보다 크게 늘어나고 있으며 인터넷 정보는 컴퓨터 파일의 형태로 되어 있기 때문에 이를 바로 기계 번역 프로그램으로 번역할 수가 있다. 따라서 기계 번역 프로그램의 수요가 점차 늘고 있으며 이미 다수의 한국어, 영어, 일본어간의 기계 번역 프로그램이 나와 있다.

이 중 일본어와 한국어는 서로 상당한 유사점을 가지고 있기 때문에 기계 번역 프로그램을 만들기가 다른 언어들보다는 쉬운 편이다. 따라서 깊이 있는 연구가 많이 진행되지는 않았지만 실제로 출시되어 있는 일한, 한일 기계 번역 프로그램은 상당히 많다. 그러나 깊이 있는 연구가 부족한 관계로 인하여 아직은 사용자가 만족할 만한 번역 품질을 가지고 있지 않다. 번역 품질이 떨어지는 이유로는 여러 가지가 있겠지만 그 중 대표적인 것으로 형태소 분석(morphological analysis)의 부정확함과 단어의 다의성 해소(word sense disambiguation) 문제의 해결 부족 등이 있다.

형태소 분석의 부정확함은 앞으로 많은 연구가 필요한 실정이다. 일본어는 한국어나 영어와 달리 띄어 쓰기가 없으므로, 하나의 형태소(morpheme)를 구분

해 내는 일은 한국어나 영어의 그 것과는 상당히 다른 복잡한 작업이며 아직 일반 사용자들이 만족할 만한 수준은 못 되고 있다.

다른 하나의 문제점은 단어의 다의성 해소(word sense disambiguation) 문제이다. 비록 일본어와 한국어가 서로 유사한 점을 많이 가지고 있기는 하여도 모든 단어가 일대일 대응 관계를 가지고 있는 것은 아니다. 따라서 하나의 일본어 단어가 경우에 따라서 각각 다른 한국어 단어로 번역되어야 하는 경우가 생기며, 이런 문제점의 경우 일본어의 조사와 용언의 번역시 특히 많이 나타난다. 이러한 단어의 다의성 해소 문제를 해결하지 않을 경우 일반적인 사용자들이 만족할 만한 번역 품질을 결코 얻을 수 없다.

본 논문에서는 일한 기계 번역에서 단어의 다의성 해소(word sense disambiguation) 문제 중 조사 번역시의 다의성 문제 해결에 초점을 맞추고 있다.

1.2 관련 연구

일한 기계 번역(Japanese to Korean machine translation)에 관한 연구에서 단어의 다의성 해소(word sense disambiguation) 문제는 큰 부분을 차지하고 있다. 지금까지 이루어진 연구는 주로 인접해서 나타나는 단어간의 연어 관계(collocation)를 이용한 것으로 [이종우 96]과 [문경희 98]이 대표적인 예이다. 이 방법은 다의성을 가진 단어마다 수작업으로 미리 연어 패턴(collocation pattern) 정보를 구축해 놓았다가, 입력 문장에 해당 단어가 출현할 때에 패턴 매칭을 시도하여, 매칭된 패턴에 해당하는 대역어를 선정함으로써 일한 기계 번역시의 다의성을 해결하는 것이다. 이 방법은 일일이 수작업으로 패턴들을 구축해야 하므로 시간이 많이 걸리며 비효율적이다. 또한 해당하는 모든 패턴을 가지고 있다고 할 수 없으므로 완전하지 못하고 새로운 패턴이 발견될 때마다 사람이 이를 입력해 주어야 하므로 유연하지 못하다.

다른 해결 방법으로는 인공 지능 기법 중 기계 학습(machine learning) 기법을 이용하는 것이 있다. 기계 학습은 자연 언어 처리(natural language processing)의 많은 분야에 이용되어 오고 있는데, 이는 주로 대량의 말뭉치(corpus)를 이용하여 자동으로 어떤 규칙을 추출함으로써 이루어진다. 형태소 분석(morphological analysis), 복합 명사 처리, 단어의 다의성 해소(word sense disambiguation) 등이 기계 학습을 통하여 시도되고 있다. 이 중 기계 학습을 이용하여 단어의 다의성 해소를 연구한 예로는 [Gale 92], [Brown 91] 등이 있다. 이들은 2 개의 언어간의 대량의 교차 말뭉치(parallel corpus)를 이용해서 단어의 다의성 문제를 해결하였다. 우선 다의성을 가진 단어의 전후 50 단어씩을 살펴서 그 단어가 특정한 의미를 가질 때 나타나는 문맥(context) 정보를 학습했다. 그리고 실제 적용시에는, 그 단어가 문장에 나타났을 때, 앞에서 학습한 문맥 정보들과 실제 나타난 문맥을 비교한다. 그리고 실제 나타난 문맥과 가장 비슷한 문맥 정보에서 그 단어가 취했던 의미를 선정함으로써 다의성을 가지는 단어의 의미를 결정했다.

1.3 연구 내용

이 논문은 일한 기계 번역 과정 중 조사의 번역에 초점을 맞추고 있다. 일본어의 조사는 한국어의 조사에 대응하는 품사이며 비슷한 특성을 아주 많이 가지고 있다. 그러나 일한 양어(兩語)의 조사 의미 대응 관계는 일대일(1 : 1) 대응이 아닌 다대일(m : n) 대응이다. 즉, 하나의 일본어 조사는 한국어로의 번역시 경우에 따라서, 각각 다른 여러 가지의 한국어 조사나 또는 다른 품사의 단어로 번역된다. 일본어 조사의 종류에 따라서 조금씩 다르기는 하지만, 일본어의 조사에 대해 하나의 한국어 조사만을 대응시켜 번역하면 최악의 경우 40% 정도의 번역 성공률밖에 보일 수 없다.

이 문제를 해결하기 위해서 통계적 기계 학습(machine learning) 기법의 하나인 베이지안 네트워크(Bayesian network)를 도입한다. 대량의 말뭉치(corpus)를 이용해서 각각의 일본어 조사마다 해당하는 베이지안 네트워크를 학습시키며 이를 이용해서 실제의 일본어 문장에 나타나는 일본어 조사의 정확한 한국어 대역어를 추론한다.

본 논문의 구성은 다음과 같다. 2 장에서는 실험에 사용된 일한 기계 번역 시스템을 설명하고 일한 기계 번역에서의 조사 번역 문제를 정형화한다. 3 장에서는 베이지안 네트워크(Bayesian network)의 일반적인 특성과 베이지안 통계적 기법에 대해서 기술하고 4 장에서 베이지안 네트워크를 이용한 조사 번역 모델을 설명한다. 5 장에서는 실험 및 결과를 기술하고 6 장에서 결론을 맺는다.

2. 일한 기계 번역에서의 조사 번역

2.1 일한 기계 번역 시스템

기계 번역 시스템의 유형들로는 일반적으로 직접(direct) 방식, 변환(transfer) 방식, 피봇(pivot) 방식의 세 가지가 있다[김영택94]. 이 논문에서 실험에 사용된 일한 기계 번역 시스템은 직접 방식과 변환 방식의 중간 쯤에 위치하고 있다. 우선 입력으로 주어진 일본어 문장에 대하여 형태소 분석을 수행한다. 그리고 구문 분석(parsing)은 행하지 않으며 바로 변환 단계를 거쳐 한국어 문장 생성 단계로 이어진다. 형태소 분석 단계에서는 형태소 사전과 여러 가지 규칙들을 이용한다. 변환 단계에서는 변환 규칙들과 각 품사별 변환 사전, 숙어(idiom) 사전을 이용하며 생성 단계에서는 생성 사전과 생성 규칙들을 이용한다.

2.1.1 형태소 분석 단계

형태소 분석 단계에서는 주어진 일본어 문장을 형태소 별로 구분하고 품사 정보와 기타 정보를 달아서 리스트 형태의 구조를 출력한다. 이 리스트의 각 노드는 일본어의 형태소 하나에 해당한다. 간단히 표현한 형태소 분석의 예는 다음과 같다.

입력 문장) 3時から始まる.

형태소 분석 결과) 3(noun) → 時(noun) → から(particle) → 始まる(verb) → .(punctuation)

2.1.2 변환 단계

변환 단계에서는 형태소 분석 결과를 입력으로 받는다. 이 단계에서 하는 일은 해당 일본어 노드에 적절한 한국어 대역어를 달아주는 것이다. 이 때 가장 큰 참고가 되는 것은 일본어 형식 형태소에 대해 미리 구성되어 있는 변환 규칙들과 일본어 의미 형태소에 해당하는 어휘 변환 사전이다. 이들은 다의어에 대한 변환 정보를 어느 정도는 가지고 있으므로 패턴 매칭을 통해서 적절한 대역어를 선정하게 된다. 한국어가 붙여진 리스트의 모양을 간단히 표현하면 다음과 같다.

변환 결과) 3(noun)[3] → 時(noun)[시] → から(particle)[부터] → 始まる(verb)[시작되다] → .(punctuation)[.]

2.1.3 생성 단계

생성 단계에서는 한국어의 음운 규칙 등 미리 가지고 있는 생성 규칙들과 생성 사전에 의해서 리스트의 노드를 따라가며 한국어 문장을 생성하게 된다. 예제 문장에 대한 생성의 결과는 다음과 같다.

생성 결과) 3 시부터 시작된다.

2.2 번역시의 다의성 문제

2 장의 1 절에서 설명한 일한 기계 번역 시스템의 구성 가운데 변환 단계가 가장 중요하고 어렵다고 할 수 있다. 그 근본적인 이유는 하나의 일본어 단어가 여러 개의 한국어 대역어 후보를 가지고 있는 경우가 꽤 있기 때문이다. 이

는 모든 품사에 적용되지만 특히 조사와 용언의 경우에 심하다. 이 문제를 해결하지 못할 경우에는 번역의 매끄러움이 문제가 아니라 아예 원문의 뜻을 파악하지 못하는 경우까지 생기게 된다. 예를 들면 다음과 같은 경우가 있다.

일본어 조사 'から(kara)' 의 경우에 해당하는 한국어 대역어 후보들

(1) 부터(조사)
(2) (으)로(조사) + 부터(조사)
(3) 에게서(조사)
(4) 에서(조사)
(5) (으)로(조사)

위 후보들이 실제 문장에 등장하는 예는 다음과 같다.

3 時から始める. -> 3 시부터 시작한다.
日ごろの言動から考えれば, -> 평소의 언동으로부터 생각하면,
父から手紙が來た. -> 아버지에게서 편지가 왔다.
女の立場から言えば, -> 여자의 입장에서 말하면,
窓からごみを捨ててはいけない. -> 창으로 휴지를 버리면 안 된다.

또한 한국어의 조사로 번역되지 않는, 더욱 복잡한 예로는 다음과 같은 것들이 있다.

100 圓から 150 圓ほどの値段 -> 100 엔 내지 150 엔 정도의 가격
百人からの人 -> 백 명이나 되는 사람

あの男は目つき<u>からして</u>抜け目なさそうだ. -> 저 남자는 눈초리<u>부터가</u> 빈틈없어 보인다.

이들 중 복잡한 경우의 예는 그리 많지 않으며 이는 숙어 사전을 통해서 해결된다. 그러나 단순히 한국어의 조사로만(또는 한국어 조사들의 조합으로) 번역되는 경우에 일본어 조사 'から(kara)'에 대응하는 한국어 대역어는 5종류 이상이며 이를 잘못 번역할 경우 번역문이 어색해지는 경우가 많이 있다. 이 경우에 제일 많이 쓰이는 대역어 하나를 언제나 선정한다고 하면 특정 조사의 경우 번역률이 50%에도 미치지 못하게 된다.

직관적으로 생각하면 이 문제는 다의성을 해결해야 하는 단어가 나타나는 문장의 문맥(context)에 의해서 결정되는 것이다. 즉, 어느 특정한 문맥에서 대역어 후보 중 특정한 어느 하나로 결정되는 것이다. 따라서 완벽한 번역을 하려면 모든 문맥 정보를 사전에 가지고 있다가 해당 단어가 나타난 문맥을 파악하여 적절한 대역어를 선정해 주는 수밖에 없다. 현재 사용하고 있는 방법은 이런 것이다. 물론 모든 문맥 정보를 가지고 있지는 않다. 따라서 번역률이 떨어질 수 밖에 없으며 완벽한 패턴 정보를 가진다는 것은 시간적, 공간적으로 거의 불가능하다. 이 논문에서 실험에 사용한 일한 기계 번역 시스템의 원래의 번역 정확 도는 'から(kara)'의 경우에 60% 정도에 불과하다. 따라서 이 문제의 해결 방법으로 유연하면서도 효율적이며 정확도가 높은 방법이 필요하며 이에 제안하는 방법이 베이지안 네트워크(Bayesian network)를 이용한 조사의 적절한 번역이다. 다음 절에서는 조사 번역 문제의 정형화(formalization)에 대해 설명한다.

2.3 조사 번역 문제의 정형화

베이지안 네트워크(Bayesian network)를 이용하여 조사 번역시의 다의성 문제를 해결하려면 우선 조사 번역 문제의 정형화(formalization)가 이루어져야 한다. 이 절에서는 조사 번역시에 어떤 문맥 정보를 보아야 하는지를 설명하고 그 정보를 각각의 변수로 나타낸다. 우선 조사를 적절한 한국어로 번역하기 위해서는 그 조사가 나타난 부근에 어떠한 단어들이 나타났는지를 살펴보아야 한다. 우선 다음의 예제 문장을 보자.

日ごろの言動から考えれば, -> 평소의 언동으로부터 생각하면,
父から手紙が來た. -> 아버지에게서 편지가 왔다.

위의 두 예제 문장에서 일본어 조사 'から(kara)'는 각각 '(으)로 + 부터', '에게서'로 번역되었다. 이는 조사 'から(kara)'의 바로 앞에 나오는 명사 '言動(gendou)', '父(chichi)'와 동사 '考える(kangaeru)', '來る(kuru)', 그리고 두 번째 문장의 경우에는 동사 '來る(kuru)'에 걸리는 또 하나의 보어(complement)[1] '手紙(tegami)' + 'が(ga)'의 차이 때문에 생긴 것이다. 따라서 일본어 조사의 의미 선정에 영향을 주는 요소로는 그 조사가 지배를 받는 용언과 그 조사가 격을 지배하는 명사, 그리고 용언에 걸리는 또 하나의 보어(명사 + 조사)가 있다고 할 수 있다. 이 네 가지의 요소를 각각 변수 V, $N1$, $N2$, $J2$ 라고 정의한다.

또한 번역을 하려는 조사는 $J1$ 이라는 변수로 나타낸다. 이 변수들은 모두 이산적인 값(discrete value)을 가지는 이산 변수이다. 그러면 조사 번역의 문제는

[1] 여기서 보어(complement)라는 용어는 일반적인 뜻과는 달리 일본어 용언을 보조해주는 '체언(명사) + 조사'를 가리킨다.

가능한 조건(V, $N1$, $N2$, $J2$ **의 값**)**이 주어진 상태에서 가장 큰 확률을 가지는** $J1$ **의 값을 구하는 문제라고 할 수 있다**.

3. 베이지안 네트워크(Bayesian network)

베이지안 네트워크(Bayesian network)는 여러 가지 변수들 간의 확률적 관계를 표현하는 그래픽 모델이다. 베이지안 네트워크의 장점은 다음과 같다 [Heckerman 95].

첫째, 모든 변수들간의 의존 관계(dependency)들을 표현하기 때문에 일부분의 데이터만을 가지는 상황에 적절히 대처할 수 있다.
둘째, 변수들간의 인과 관계(causal relationship)를 학습하는데 사용될 수 있기 때문에 응용 분야에 대한 이해를 도울 수 있다.
셋째, 모델 자체가 원인(causality)과 확률적 의미(probabilistic semantics)를 표현하고 있기 때문에 사전 지식(prior knowledge)과 학습 데이터를 결합하는데 적합하다.
넷째, 베이지안 네트워크에 베이지안 통계 기법을 적용함으로써 데이터 과적합(data overfitting)을 막을 수 있다.

3.1 베이지안 확률 기법(Bayesian probabilistic method)을 이용한 데이터 학습

베이지안 확률 기법(Bayesian probabilistic method)을 이용한 데이터 학습을 전통적인 통계적 기법과 비교하기 위하여 다음의 예를 든다. 동전을 N 번 던진 결과 데이터를 가지고 $N+1$ 번째의 결과를 예측하는 경우이다. 동전을 던진 결과는 앞(*heads*), 뒤(*tails*)가 있다. 전통적인 통계 기법을 사용하는 경우는 동전을 던졌을 때 앞이 나올 물리적 확률(physical probability)이 알려지지는 않았지만 존재한다고 가정한다. 그리고 이 물리적 확률을 N 번의 관찰을 토대로

하여 추정하고 이를 이용하여 N + 1 번째의 결과를 예측한다. 베이지안 기법에서도 역시 앞면이 나올 어떤 물리적 확률이 존재한다고 가정한다. 그리고 이 확률에 대한 불확실성(uncertainty)을 베이지안 확률을 사용해 표현하고 확률의 법칙들을 이용하여 N + 1 번째의 결과를 예측한다.

베이지안 기법의 경우를 좀 더 자세히 설명한다. 앞면이 나올 물리적 확률을 변수 Θ로 표현한다. 이 변수가 가지는 값 q가 동전을 던졌을 때 앞면이 나올 물리적 확률이다. 이 q를 모수(parameter)라고 한다. 이 변수 Θ가 가지는 불확실성(uncertainty)을 확률 밀도 함수(probability density function) $p(q|x)$를 이용하여 표현한다(x는 지금까지 주어진 주변 정보를 나타낸다). 그리고 X_l은 l 번째 관찰의 결과를 나타내는 변수이다. l = 1, .., N + 1 이며 $D = \{X_1 = x_1$, .., $X_N = x_N\}$은 실험의 결과를 나타내는 집합이다. 베이지안 기법에서는 결국 사전 확률 분포(prior probability distribution) $p(q|x)$에서 $p(x_{N+1}|D,x)$를 계산하는 것이 문제의 해결이 된다.

이를 계산하기 위해 데이터 D가 주어졌을 때 Θ의 확률 분포를 나타내는 식을 다음과 같이 표현할 수 있다.

$$p(q \mid D,x) = \frac{p(q \mid x)\, p(D \mid q,x)}{p(D \mid x)} \qquad \text{(수식 1)}$$

여기에서

$$p(D \mid x) = \int p(D \mid q,x)\, p(q \mid x)\, dq \qquad \text{(수식 2)}$$

이제 (수식 1)의 분자의 $p(D|q,x)$를 확장하면 다음과 같다. 주어진 Θ의 값

에 대해 D에서의 관찰들은 각각 서로 독립이며 각 관찰에서 앞면(뒷면)이 나올 확률은 $q(1 - q)$이다. 따라서 (수식 1)은 다음과 같이 된다.

$$p(q \mid D,x) = \frac{p(q \mid x) q^{h} (1-q)^{t}}{p(D \mid x)} \qquad \text{(수식 3)}$$

여기에서 h와 t는 각각 D에서 앞면과 뒷면이 관찰된 횟수이다. $p(q|x)$, $p(q|D,x)$는 각각 사전 확률 분포(prior probability distribution), 사후 확률 분포(posterior probability distribution)이다. h와 t는 이항 샘플 추출(binomial sampling)에 대한 충분한 통계적 지표(sufficient statistics)이다. 최종적으로 위의 (수식 3)을 가능한 모든 Θ의 값에 대해 종합함으로써 $N + 1$ 번째 관찰에서 앞면이 나올 확률을 계산할 수 있다.

$$\begin{aligned} p(X_{N+1} = heads \mid D,x) &= \int p(X_{N+1} = heads \mid q,x)\, p(q \mid D,x)\, dq \\ &= \int q\, p(q \mid D,x)\, dq \equiv E_{p(q|D,x)}(q) \end{aligned} \qquad \text{(수식 4)}$$

위의 식을 계산하기 위해서는 Θ의 사전 확률 분포(prior probability distribution)를 알아야 한다. 널리 사용되는 방법 중 하나는 사전 확률 분포가 베타 분포(beta distribution)라고 가정하는 것이다.

$$p(q \mid x) = Beta(q \mid a_h, a_t) \equiv \frac{\Gamma(a)}{\Gamma(a_h)\Gamma(a_t)} q^{a_h - 1} (1-q)^{a_t - 1} \qquad \text{(수식 5)}$$

$a_h > 0$, $a_t > 0$는 베타 분포의 모수(hyperparameter)이며, $a = a_h + a_t$이다. $\Gamma(\bullet)$는 감마 함수이며 $\Gamma(x + 1) = x\Gamma(x)$, $\Gamma(1) = 1$이다.

(수식 3)에 의하여 사후 확률 분포(posterior probability distribution) 역시 베타 분포를 따른다[Heckerman 95].

$$p(q \mid D,x) = \frac{\Gamma(a+N)}{\Gamma(a_h+h)\Gamma(a_t+t)} q^{a_h+h-1}(1-q)^{a_t+t-1} = Beta(q \mid a_h+h, a_t+t) \quad \text{(수식 6)}$$

베타 분포에 대한 q의 기대값은 다음과 같다.

$$\int qBeta(q \mid a_h, a_t)dq = \frac{a_h}{a} \quad \text{(수식 7)}$$

따라서 N + 1 번째에 앞면이 나올 확률도 다음과 같이 간단히 계산된다.

$$p(X_{N+1} = heads \mid D,x) = \frac{a_h + h}{a + N} \quad \text{(수식 8)}$$

이제, 문제는 사전 확률 분포(prior probability distribution) $p(q|x)$를 구하는 것이다. 이는 여러 가지 방법으로 구할 수 있는데, 그 방법들로는 예측된 미래의 데이터(imagined future data), 등가 샘플(equivalent samples) 등이 있다[Heckerman 95]. 베타 사전 확률(beta prior probability)을 이용하는 것이 편리하기는 하지만 항상 정확한 결과를 주는 것은 아니다. 그런 경우에는 다수의 베타 분포의 합을 이용하는 것이 더 적절할 수 있다. 예를 들면,

$$p(q \mid x) = 0.4Beta(20,1) + 0.4Beta(1,20) + 0.2Beta(2,2) \quad \text{(수식 9)}$$

와 같은 분포를 생각할 수 있다. 이 경우, 0.4는 동전이 앞면 또는 뒷면으로 치

우치는 물리적 확률을 가질 확률을 나타내고 0.2는 치우치지 않는 물리적 확률을 가질 확률을 나타낸다. 이는 사전 확률 분포에 보이지 않는 숨겨진 변수(hidden variable)가 존재하는 경우로 볼 수도 있다. 이 숨겨진 변수가 동전이 앞이나 뒤로 치우치거나 치우치지 않을 확률을 나타낸다고 보면 된다. 즉, 베타 사전 분포가 적절하지 않은 경우에는 우리가 모르는 숨겨진 변수가 있다고 가정하고 그 변수를 존재하게 하면 되는 것이다. 지금까지 예로 든 것은 이항 분포(binomial distribution)에 관한 것이지만 이는 얼마든지 다른 분포에도 적용시킬 수가 있다. 예를 들면,

$$p(x \mid \mathbf{s}, x) = f(x, \mathbf{s}) \qquad \text{(수식 10)}$$

위와 같은 분포에서 $f(x,\mathbf{s})$는 벡터 인자 $\mathbf{s}$를 가지는 추정 함수(likelihood function)이다. 이 벡터 인자의 차원은 유한하다고 가정한다. 이러한 분포의 예로 다음과 같은 경우를 들 수 있다. 연속된 값을 가지는 변수 X가 가우시안 확률 분포(Gaussian physical probability distribution)를 가지며 평균을 m, 분산을 v라고 할 때 다음과 같은 식으로 표현된다.

$$p(x \mid \mathbf{s}, x) = (2pv)^{-1/2} e^{-(x-m)^2/2v} \qquad \text{(수식 11)}$$

이 때 $\mathbf{s} = \{m, v\}$이다.

임의의 형식의 변수들에 대해서도 위의 동전을 던지는 경우와 마찬 가지로 사전 확률(prior)을 할당하고 베이즈의 규칙(Bayes' theorem)을 이용해서 주어진 데이터에 대한 사후 확률(posterior)을 구할 수 있다. 그 식은 다음과 같다.

$$p(\mathbf{s} \mid D, x) = \frac{p(D \mid \mathbf{s}, x)\, p(\mathbf{s} \mid x)}{p(D \mid x)} \qquad \text{(수식 12)}$$

위의 식에서 s 는 모수(parameter)들의 벡터이다. 그리고 가능한 S 의 값들에 대해서 예측을 할 수 있다. 예를 들면,

$$p(x_{N+1} \mid D, x) = \int p(x_{N+1} \mid \mathbf{s}, x)\, p(\mathbf{s} \mid D, x)\, d\mathbf{s} \qquad \text{(수식 13)}$$

이다.

지수 분포에 해당하는 변수들의 경우에는 위의 식의 계산을 닫힌 형식(closed form)으로 효율적으로 할 수 있다는 것이 알려져 있다. 이러한 분포들로는 이항(binomial), 다항(multinomial), 정규(normal), 감마(Gamma), 포아송(Poisson) 분포 등이 있다. 다음의 예는 다항 샘플 추출(multinomial sampling)에 관한 것이다. 다항 샘플 추출의 경우 관찰 변수 X 는 이산적(discrete)이며 r 개의 가능한 값 x_1, .., x_r을 가진다. 이 때 추정 함수(likelihood function)는 다음과 같다.

$$p(X = x^k \mid \mathbf{s}, x) = s_k, \qquad k = 1, \ldots, r \qquad \text{(수식 14)}$$

여기에서 s = {s_2, .., s_r}은 모수(parameter)들이다. s_1 의 값은 확률의 합의 법칙에 의해 나머지 s_i 값들의 합을 1 에서 뺀 값이 되므로 굳이 벡터에 포함시킬 필요가 없다. 이 경우도 이항 샘플 추출(binomial sampling)의 경우와 마찬가지로 모수들은 우리가 모르는 어떤 물리적 확률(physical probability)에 해당한다. 이 경우에 데이터 집합 D = {X_1 = x_1, .., X_N = x_N}에 대한 충분한 통계적 지

표(sufficient statistics)는 $\{N_1, \ldots, N_r\}$이며 N_i는 D에서 x_i가 나타난 횟수이다. 다항 샘플 추출(multinomial sampling)과 함께 쓰이는 간단한 집합적 사전 확률(conjugate prior probability)은 디리힐렛 분포(Dirichlet distribution)이며 다음과 같은 식으로 표현된다.

$$p(\mathbf{s} \mid x) = Dir(\mathbf{s} \mid a_1, \ldots, a_r) \equiv \frac{\Gamma(a)}{\prod_{k=1}^{r} \Gamma(a_k)} \prod_{k=1}^{r} s_k^{a_k - 1} \qquad \text{(수식 15)}$$

여기에서 $a = a_1 + \ldots + a_r$, $a_k > 0$ ($k = 1, \ldots, r$)이다. 이 때 데이터가 주어지고 난 후의 사후 확률 분포(posterior probability distribution)는 다음과 같다.

$$p(\mathbf{s} \mid D, x) = Dir(\mathbf{s} \mid a_1 + N_1, \ldots, a_r + N_r) \qquad \text{(수식 16)}$$

다항 분포에서도 예측된 미래의 데이터(imagined future data), 등가 샘플(equivalent samples) 등의 방법을 이용하여 디리힐렛 분포(Dirichlet distribution)를 계산할 수 있다. 이러한 집합적 사전 확률 분포(conjugate prior probability distribution)와 데이터 집합 D를 가지고 다음의 관찰 결과에 대한 예측 확률 분포는 다음과 같이 구할 수 있다.

$$p(X_{N+1} = x^k \mid D, x) = \int s_k Dir(\mathbf{s} \mid a_1 + N_1, \ldots, a_r + N_r) d\mathbf{s} = \frac{a_k + N_k}{a + N} \qquad \text{(수식 17)}$$

베이지안 분석에서 중요한 또 하나의 양(quantity)은 한계 가능성(marginal likelihood) 또는 증거(evidence) $p(D|x)$이다. 이 경우는 다음과 같이 주어진다.

$$p(D|x) = \frac{\Gamma(a)}{\Gamma(a+N)} \cdot \prod_{k=1}^{r} \frac{\Gamma(a_k + N_k)}{\Gamma(a_k)} \qquad \text{(수식 18)}$$

현재 배경 지식의 상태를 나타내는 변수인 x는 확률이 어딘가에 종속적이라는 사실을 나타내는데, 이 사실을 항상 옳다고 가정하면 굳이 변수로써 표시할 필요가 없다. 따라서 다음의 수식부터는 x를 항상 생략한다.

3.2 베이지안 네트워크(Bayesian network)

베이지안 네트워크는 변수들간의 결합 확률 분포(joint probability distribution)를 효율적으로 표현할 수 있는 그래픽 모델이다. 변수 집합 **X** = $\{X_1, \ldots, X_n\}$에 대한 베이지안 네트워크는 다음의 2가지 부분으로 구성된다.

(1) 집합 **X**의 변수들간의 조건부 독립 가정(conditional independence assertion)을 표현하는 네트워크의 구조 S

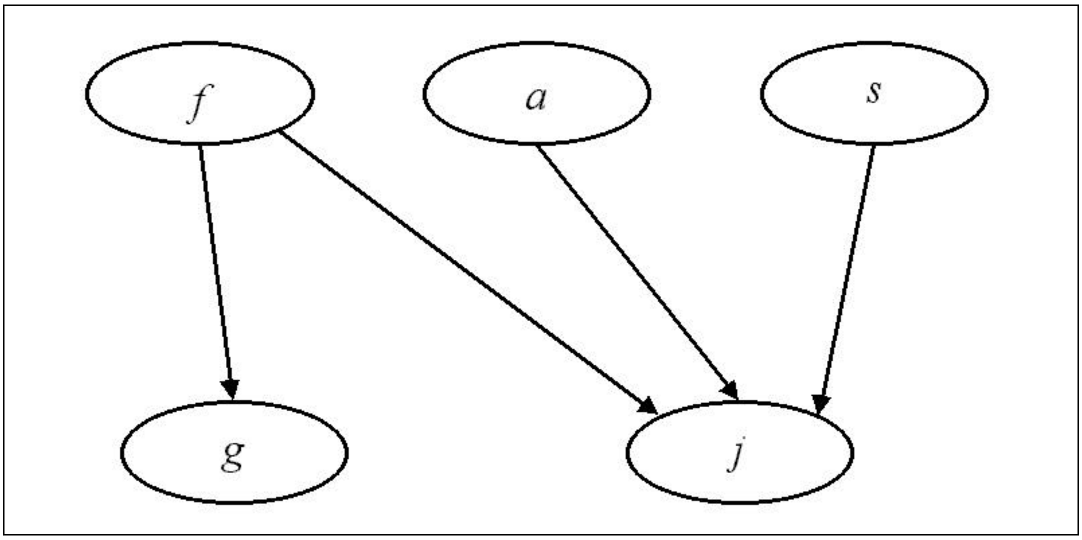
(2) 각 변수들과 연관되어 있는 지역 확률 분포(local probability distribution) 집합 P

네트워크의 구조 S는 사이클이 없는 방향성 그래프(directed acyclic graph)이다. S의 각 노드는 **X**의 각 변수들과 일대일 대응이 된다. X_i는 변수와 그 변수에 해당하는 노드를 동시에 나타낸다. $\mathbf{Pa}_i$는 그래프 S에서 X_i의 부모 노드(변수) 집합을 가리킨다. 부모 노드는 0개 이상이다. S에서 서로 연결되지 않은 노드는 서로 조건부 독립 관계에 있음을 나타낸다. 주어진 구조 S에 대해서 **X**의 결합 확률 분포(joint probability distribution)는 다음과 같이 주어진다.

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{Pa}_i) \qquad \text{(수식 19)}$$

지역 확률 분포 P는 위의 식의 Π 안의 각 항에 대응된다. 결과적으로 확률 분포 $p(\mathbf{x})$를 나타내기 위해서는 (S, P)가 필요하다. 베이지안 네트워크에 의해 표현되는 확률들은 물리적 확률(physical probability)이거나 베이지안 확률(Bayesian probability)일 수 있다. 만일 사전 지식(prior knowledge)만을 가지고 베이지안 네트워크를 구성한다면 베이지안 확률이 표현되는 것이고 데이터로부터 베이지안 네트워크를 학습하는 경우에는 물리적 확률이 표현되는 것이다 [Heckerman 95].

서로 독립인 5 개의 변수 f, a, s, g, j에 대한 베이지안 네트워크의 예는 (그림 1)과 같을 수 있다.



(그림 1)

3.3 베이지안 네트워크(Bayesian network)를 이용한 추론

베이지안 네트워크를 구성하고 난 뒤에는 주어진 조건에 대한 특정 변수의 확률 분포를 알 수 있다. 이 확률은 모델에 직접 저장되어 있는 것이 아니기 때문에 베이지안 네트워크를 이용하여 계산(추론)되어야 한다.

변수 집합 X 에 대한 결합 확률 분포(joint probability distribution)는 베이지안 네트워크에 의해서 정해지기 때문에 원칙적으로는 베이지안 네트워크를 이용해서 원하는 모든 확률을 계산해 낼 수 있다. 예를 들어 (그림 1)의 베이지안 네트워크에서 다른 변수들의 값을 관찰한 뒤 변수 f의 확률을 계산하려면 다음과 같이 계산하면 된다.

$$p(f \mid a,s,g,j) = \frac{p(f,a,s,g,j)}{p(a,s,g,j)} = \frac{p(f,a,s,g,j)}{\sum_{f'} p(f',a,s,g,j)} \qquad \text{(수식 20)}$$

하지만 변수들의 개수가 많은 경우에는 이와 같은 직접적 방법을 적용하기가 어렵다. 다행히도 모든 변수들이 이산 변수(discrete variable)인 경우에는 네트워크의 구조에 나타나 있는 조건부 독립 가정을 이용하여 이 계산을 더욱 효율적으로 할 수 있다. 위의 식은 다음과 같이 표현될 수 있다.

$$\begin{aligned} p(f \mid a,s,g,j) &= \frac{p(f)p(a)p(s)p(g \mid f)p(j \mid f,a,s)}{\sum_{f'} p(f')p(a)p(s)p(g \mid f')p(j \mid f',a,s)} \\ &= \frac{p(f)p(g \mid f)p(j \mid f,a,s)}{\sum_{f'} p(f')p(g \mid f')p(j \mid f',a,s)} \end{aligned} \qquad \text{(수식 21)}$$

(수식 21)에서 확률 $p(a)$, $p(s)$의 항이 지워진 까닭은 변수 a, s가 변수 f

와 독립이기 때문이다. 이산 변수들로만 구성되어 있는 베이지안 네트워크에서 확률을 추론해내기 위해 조건부 독립 가정(conditional independence assertion)을 이용하는 몇몇 알고리즘이 개발되었다. 예를 들면[Howard 81], [Shachter 88] 등이 있다. 가장 널리 이용되는 알고리즘은 [Lauritzen 88], [Jensen 90] 등이다.

확률의 추론 과정을 간단히 하기 위하여 조건부 독립 가정(conditional independence assertion)을 사용할 수 있기는 하지만 이산 변수들로 구성된 임의의 베이지안 네트워크에서 일반적인 경우의 정확한 추론은 NP-hard 문제임이 알려져 있다[Cooper 90]. 또한 근사적인 추론(approximate inference)도 NP-hard 문제이다. 이 어려움은 그래프에 존재할 수 있는 무방향 사이클(undirected cycle) 때문이다. 하지만 많은 경우에 이러한 사이클은 존재하지 않는다. 이 논문에서 다루는 네트워크의 위상(topology)도 마찬가지이다. 또한 무방향 사이클이 존재하는 베이지안 네트워크의 경우에도 문제의 분야(problem domain)에 따라 경험적인 여러 가지 방법들을 이용하여 해결할 수 있는 경우가 많이 있다.

3.4 베이지안 네트워크(Bayesian network)에서의 확률의 학습

이 절에서는 주어진 네트워크 구조에서 데이터를 이용해 각 노드의 지역 확률 분포(local probability distribution)를 갱신하는 방법을 논한다.

네트워크 구조가 확정되면 주어진 네트워크 구조가 변수 집합 **X** 의 물리적(실제) 결합 확률 분포(physical joint probability distribution)를 표현할 수 있다고 가정한다. 그러면 다음과 같은 식을 쓸 수 있다.

$$p(\mathbf{x} \mid \mathbf{z}_s, S^h) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i, \mathbf{z}_i, S^h) \qquad \text{(수식 22)}$$

여기에서 $\mathbf{z}_i$ 는 확률 분포 $p(x_i|\mathbf{pa}_i, \mathbf{z}_i, S^h)$에 대한 모수(parameter)들의 벡터이다. $\mathbf{z}_s$ 는 $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$이며 S^h 는 실제의 결합 확률 분포(physical joint probability distribution)가 네트워크 구조 S 를 통해서 표현될 수 있다는 가설(hypothesis)이다. 그리고 학습할 데이터로는 무작위로 추출한 샘플(random sample) $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$이 있다고 가정한다. D 의 원소 $\mathbf{x}_l$ 은 l 번째의 경우(case)라고 한다. 3장 1절에서와 비슷하게 $\mathbf{z}_s$ 에 대한 불확실성(uncertainty)을 벡터 형태의 변수 $\mathbf{Z}_s$ 를 정의함으로써 나타내고 사전 확률 밀도 함수(prior probability density function) $p(\mathbf{z}_s| S^h)$를 정한다. 이제 베이지안 네트워크에서 확률을 학습하는 문제는 다음과 같이 기술할 수 있다. 주어진 무작위의 샘플 D 에서 사후 확률 분포 $p(\mathbf{z}_s|D, S^h)$를 계산한다.

$\mathbf{z}_i$ 에 대한 함수라고 볼 수 있는 확률 분포 $p(x_i|\mathbf{pa}_i, \mathbf{z}_i, S^h)$를 지역 확률 분포 함수(local distribution function)라고 부른다. 이 함수는 확률적 분류(probabilistic classification)나 회귀(regression) 함수와 같은 꼴을 가지고 있다. 결국 베이지안 네트워크는 확률적 분류/회귀 모델들을 조건부 독립 가정(conditional independence assertion)에 의해 조직해서 모아 놓은 것으로 볼 수 있다. 확률적 결과를 내는 분류/회귀 모델들은 여러 가지가 연구되었으나 이 중 널리 연구된 것은 무제한 다항 분포(unrestricted multinomial distribution), 가우시안 노이즈를 가지는 선형 회귀(linear regression with Gaussian noise), 일반화된 선형 회귀(generalized linear regression) 등이 있다. 이들 중 무제한 다항 분포(unrestricted multinomial distribution)의 예를 들어 본다.

이 경우 각 변수 $X_i \in \mathbf{X}$ 는 이산적이다. 즉, r^i 개의 이산 값(discrete value)들을 가진다. 그리고 각 지역 확률 분포 함수는 $\mathbf{Pa}_i$ 의 각 상태(configuration)에 대한 다항 분포(multinomial distribution)들의 집합이다. 이를 식으로 표현하면 다음과 같다.

$$p(x_i^k \mid \mathbf{Pa}_i^j, \mathbf{z}_i, S^h) = z_{ijk} > 0$$
$$\text{단}, \mathbf{pa}_i^1, \ldots, \mathbf{pa}_i^{q_i} \left(q_i = \prod_{X_i \in \mathbf{Pa}_i} r_i\right)\text{는 } \mathbf{Pa}_i \text{가 취할 수 있는 값들} \qquad \text{(수식 23)}$$
$$\mathbf{z}_i = ((z_{ijk})_{k=2}^{r_i})_{j=1}^{q_i}$$

표현의 편리함을 위해서 모수(parameter)들의 벡터를 다음과 같이 정의한다.

$$\mathbf{z}_{ij} = (z_{ij2}, \ldots, z_{ijr_i}) \qquad \text{(수식 24)}$$

무제한(unrestricted)이라는 말은 이 확률 분포를 낮은 차원의 $\mathbf{Pa}_i$에 대한 함수와 구분하기 위하여 사용되었다. 이러한 형식의 지역 확률 분포 함수와 다음의 두 가지 가정이 주어지면, 사후 확률 분포(posterior probability distribution) $p(\mathbf{z}_s|D, S^h)$는 효율적으로 닫힌 형식(closed form)으로 계산될 수 있다. 첫 번째 가정은 무작위 샘플 D에 빠진 데이터(missing data)가 없다는 것이다. 이러한 샘플 D를 완전하다(complete)고 한다. 두 번째 가정은 모수 벡터 $\mathbf{z}_{ij}$가 서로 독립이라는 것이다. 이는 다음의 식으로 표현된다.

$$p(\mathbf{z}_s \mid S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\mathbf{z}_{ij} \mid S^h) \qquad \text{(수식 25)}$$

이러한 가정은 [Spiegelhalter and Lauritzen 90]에 의하여 소개되었으며 이를 모수 독립(parameter independence)이라고 부른다. 완전 데이터와 모수 독립의 가정 하에서는 무작위 샘플들이 주어진 후에도 모수들은 서로 독립이다.

$$p(\mathbf{z}_s \mid D, S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\mathbf{z}_{ij} \mid D, S^h) \qquad \text{(수식 26)}$$

이렇게 해서 각각의 모수 벡터 $\mathbf{z}_{ij}$를 독립적으로 학습할 수 있다. 각각의 벡터 $\mathbf{z}_{ij}$가 사전 확률 분포(prior probability distribution) $Dir(\mathbf{z}_{ij}|\alpha_{ij1}, .,. \alpha_{ijri})$를 가진다고 하면, 사후 분포(posterior distribution)는 다음과 같이 된다.

$$p(\mathbf{z}_{ij} \mid D, S^h) = Dir(\mathbf{z}_{ij} \mid a_{ij1} + N_{ij1}, \ldots, a_{ijr_i} + N_{ijr_i}) \qquad \text{(수식 27)}$$

앞의 동전 던지기의 예에서와 같이 다음의 경우(case)를 예측하기 위하여 $\mathbf{z}_s$의 모든 상태(configuration)에 대해 기대값을 구할 수 있다. 예를 들어 $p(\mathbf{x}_{N+1}|D, S^h)$는 다음의 식으로 예측이 가능하다.

$$p(\mathbf{x}_{N+1} \mid D, S^h) = E_{p(\mathbf{z}_s|D,S^h)}\left(\prod_{i=1}^{n} z_{ijk}\right) \qquad \text{(수식 28)}$$

이 기대값을 구하기 위해서 D가 주어졌을 때에도 각 모수들은 서로 독립이라는 사실을 이용한다.

$$p(\mathbf{x}_{N+1} \mid D, S^h) = \int \prod_{i=1}^{n} z_{ijk}\, p(\mathbf{z}_s \mid D, S^h) d\mathbf{z}_s = \prod_{i=1}^{n} \int z_{ijk}\, p(\mathbf{z}_{ij} \mid D, S^h) d\mathbf{z}_{ij} \qquad \text{(수식 29)}$$

따라서 앞의 식 (수식 17)에 의해 다음의 결과를 얻을 수 있다.

$$p(\mathbf{x}_{N+1} \mid D, S^h) = \prod_{i=1}^{N} \frac{a_{ijk} + N_{ijk}}{a_{ij} + N_{ij}} \qquad \text{(수식 30)}$$

3.5 베이지안 네트워크가 조사 번역에 적합한 이유

베이지안 네트워크는 각 변수들간의 확률적 관계를 표현하는데 적합하다. 2 장의 마지막 절에서 일본어 조사 번역 문제를 주어진 변수 4 개(V, $N1$, $N2$, $J2$)의 값이 주어졌을 때 변수 $J1$ 이 어떤 값을 취할 확률을 구하는 문제로 정형화 하였다. 그 확률이 가장 높은 값을 $J1$ 이 취하게 하면 번역이 이루어지는 것이다. 따라서 5 개의 변수들간의 확률적 관계를 알아야 번역을 할 수 있다. 하지만 그 확률적 관계를 사람이 직접 계산할 수는 없다. 다만 사람이 알 수 있는 것은 5 개의 변수간의 인과 관계(causal relationship) 정도이다. 이 인과 관계를 사전 지식(prior knowledge)이라고 하면 이는 네트워크의 구조(위상)를 구성하는데 사용될 수 있다. 또한 주어진 일본어 말뭉치(corpus) 데이터는 무작위의 샘플(random sample)이라고 할 수 있다. 이 데이터로는 주어진 네트워크 구조의 각 노드들의 지역 확률 분포(local probability distribution)를 계속 갱신(update)할 수 있다. 즉, 사람이 가지고 있는 사전 지식(문장의 각 성분들간의 인과 관계)과 실험 데이터를 쉽게 결합할 수 있는 방법이 바로 베이지안 네트워크이다.

또한 위의 변수들은 언제나 항상 값을 취하는 것은 아니다. 문장 성분이 생략될 수 있기 때문에 조사 $J1$ 의 격지배를 받는 명사 $N1$ 을 제외하고는 모두 생략될 수 있다. 따라서 불완전한 샘플이 데이터 내에 많이 존재할 수 있다. 이런 경우의 학습에도 베이지안 네트워크는 적합하다. 이런 경우의 학습 알고리즘에 대해서는 뒤에서 설명한다. 그리고 위의 변수들은 모두 이산 변수(discrete variable)이므로 데이터를 이용한 학습이 쉽게 이루어질 수 있다.

3. 베이지안 네트워크(Bayesian network)

베이지안 네트워크가 구성되고 말뭉치 데이터를 이용해서 네트워크를 학습시킨 후에는 비교적 간단한 방법으로 필요한 확률을 추론(inference)할 수 있다. 이 것은 네트워크의 구조가 비교적 간단하기 때문이다.

4. 베이지안 네트워크를 이용한 조사 번역 모델

4.1 네트워크 위상(network topology)의 구성

베이지안 네트워크의 위상(구조)은 학습 성과를 결정하는데 있어 중요한 요소이다. 결국 이 그래프는 각 변수들과 그 변수들간의 관계를 표현하는 것이다. 네트워크의 구조를 결정하는데 있어서 가장 중요한 것은 그 네트워크가 적용될 분야에 대한 지식이라고 할 수 있다.

문제 해결에 이용될 베이지안 네트워크의 위상을 구성하는 방법에는 크게 두 가지가 있다. 첫째는 사람이 직접 각 변수들간의 인과 관계(causal relationship)를 이용하여 네트워크의 위상을 결정하는 방법이다. 베이지안 네트워크에서 각 노드는 변수에 해당한다. 사람은 우선 문제 해결에 중요한 변수들을 결정한다. 그 뒤에 그 변수들 중 인과 관계가 있다고 생각되는 노드들을 화살표로 연결하는 것이다. 두 번째 방법은 대량의 데이터를 이용해서 네트워크의 구조를 결정하는 것이다. 이 경우에는 우선 필요한 변수들을 설정한다. 그리고 이 변수들간의 인과 관계를 대량의 데이터를 이용해서 찾아내는 것이다. 여기에는 각 모델들의 적합도(fitness)를 결정할 수 있는 기준(criterion)이 필요하며 이 기준을 이용해서 필요한 모델들을 찾아낼 수 있는 탐색 기법(searching method)이 필요하다.

이 논문에서는 사람이 직접 네트워크의 구조를 결정하는 방법을 채택했다. 그 이유는 관련된 변수들의 개수가 5 개로 그리 많지 않고, 문제를 해결하는데 필요한 인과 관계도 비교적 명확하기 때문이다. 네트워크를 구성하기 위해 2 장에서 설명 했던 조사 번역에 영향을 주는 요소들을 각 변수들로 이용한다. 이들은 다음과 같다.

$N1$: 번역하려는 조사에 의해 격지배를 받는 명사

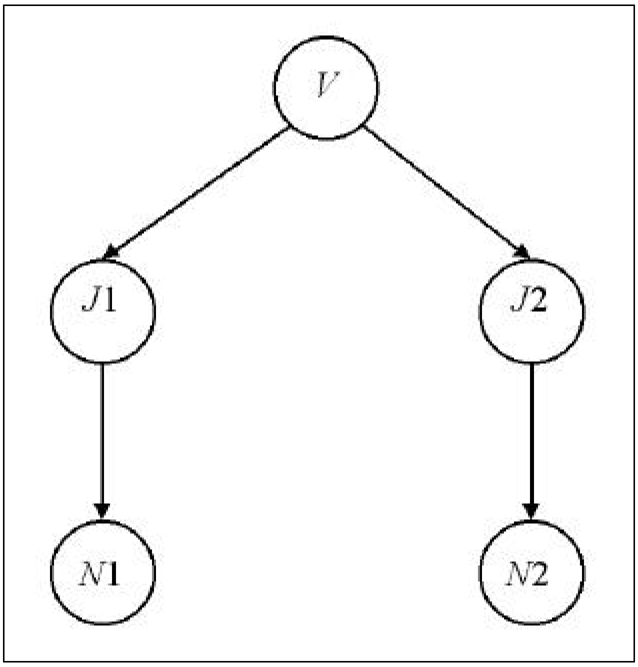
$J1$: 번역하려는 조사

$N2$: 번역하려는 조사를 지배하는 용언의 또 다른 보어가 되는 명사

$J2$: $N2$ 를 격지배하는 조사

V: 번역하려는 조사를 지배하는 용언

위의 다섯 개의 변수로 이루어진 네트워크를 번역하려는 일본어 조사에 대해서 각각 구성한다. 위의 변수들의 인과 관계는 다음과 같다. 동사 V 는 자신이 지배하는 조사 $J1$, $J2$ 에 대해서 인과 관계를 가진다. 각 조사 $J1$, $J2$ 는 자신이 격지배하는 명사 $N1$, $N2$ 에 대해서 각각 인과 관계를 가진다. 이를 그림으로 표현하면 (그림 2)와 같다.



(그림 2)

다음에 필요한 것은 각 변수들이 취할 수 있는 값의 범위이다. 변수들 중 조사를 나타내는 변수 *J*2는 번역에서 고려하는 일본어 격조사 6 개(が(ga), から(kara), で(de), に(ni), まで(made), を(wo))를 값으로 취하는 이산 변수(discrete variable)이다. 그리고 번역하려는 조사 *J*1은 해당 일본어 조사가 번역될 수 있는 한국어 대역어 후보들을 값으로 가진다(2 – 5 개). 그러나 명사와 용언을 나타내는 변수는 취하는 값을 한정해 주어야 한다. 명사와 용언이 취할 수 있는 값의 범위는 모델의 성공도를 결정하는데 있어서 매우 중요하다고 할 수 있다. 우선 명사나 용언이 취할 수 있는 모든 값들을 변수가 취할 수 있다고 하면 그 경우의 수가 너무 많아진다. 예를 들어 현재 이 논문에서 실험에 사용하고 있는 일한 기계 번역 시스템의 명사 사전의 엔트리는 약 4 만 개 정도이고 동사(용언) 사전의 경우는 이보다는 적은 2 만 개 정도이다. 그러나 이

를 모두 사용해서 실험을 하려고 하면 말뭉치 데이터가 적어도 100 만 문장 이상이어야 한다. 하지만 실험에 사용한 말뭉치 데이터는 약 7 천 문장이고 이 중 다의성 해소를 하려고 하는 조사가 나타나는 문장의 경우 각 일본어 조사에 따라서 차이는 있지만 대략 900 – 4000 문장에 불과하다. 따라서 명사와 용언에 의미 클래스의 개념의 도입이 필요하다.

의미 클래스(semantic class)

의미 클래스는 자연 언어 처리(natural language processing)의 거의 모든 영역에서 필요한 개념이다. 즉 여러 가지의 단어들을 품사가 아닌 의미(semantic)에 따라서 그룹을 짓는 것이다. 이에 대한 연구로는 미국의 Wordnet, 한국의 [문유진 96] 등이 있다. 일본어에 해당하는 의미 클래스 연구는 [大野 82]와 같은 것들이 있다. 하지만 이들은 의미 클래스의 개수가 너무 많으므로 조사 번역 문제의 해결에 있어서는 조사 번역에만 필요한 정도로 의미 클래스의 개수를 축소시킬 필요가 있다. 본 논문에서는 [황호성 98]에 이용되었던 한국어 의미 클래스를 본 연구에 맞도록 수정하여 사용한다. [황호성 98]의 연구에서는 한영 기계 번역에서 한국어 부사격 조사의 번역을 위하여 명사와 용언에 각각 다음의 의미 클래스를 부여했다. 이는 일본어의 명사와 용언에도 무리 없이 그대로 적용이 된다. 명사의 의미 클래스는 (표 1)과 같이 22 개로, 용언의 의미 클래스는 (표 2)와 같이 13 개로 정했다. 이는 일어일문학 전공자들의 도움을 받아서 이루어졌으므로 사람이 직접 각 명사와 용언에 의미 클래스를 부여해 주었다고 할 수 있다.

의미 클래스	예	의미 표지
장소	學校(gakkou), 會社(kaisha)	P

시간	夜(yoru), 時(toki)	T
감정적 행위	微笑み(hohoemi), 愛(ai)	AC
감각적 행위	聲(koe), 感じ(kanji)	AF
이성적 행위	考え(kangae), 理解(rigai)	AR
단위	圓(en), メ-トル(metoru)	AU
형태	十字(juuji), 四角(sikaku)	AT
지적 추상물	法(hou), 規則(kisoku)	AI
도구	刀(katana), 鉛筆(enpitsu)	CI
물리적 행위	移動(idou), 接觸(setsyoku)	AP
생리 현상	いたみ(itami), 侵り(nemuri)	AB
자연 현상	雨(ame), 風(kaze)	AN
사회 현상	革命(gakumei), 戰爭(densai)	ASP
재료	紙(kami), 木材(mokuzai)	CMT
탈 것	車(kuruma), 地下鐵(jikatetsu)	CT
기계	テレビ(terebi), ラジオ(razio)	CMA
통신 수단	手紙(tegami), 電話(denwa)	ACM
지위	社長(satyou), 總理(souri)	CS
물품	お金(okane), タイヤ(taiya)	CC
인간 부속물	目(me), 耳(mimi), 口(kutsi)	CH
언어	英語(eigou), 韓國語(kankokugou)	AL
상태	始まり(hazimari), 結末 (ketsumatsu)	AS

(표 1)

의미 클래스	예	의미 표지
수여	くれる(kureru)	G
발화전달	言う(iu), 話す(hanasu)	S
이동	行く(iku), 來る(kuru)	D
수익	聞く(kiku), 受ける(ukeru)	R
기여	親切だ(sinsetsuda)	B
관계	大切だ(daisetsuda)	REL
한정	禁ずる(kinzuru)	L
형성	飾る(kazaru)	C
인식	知る(siru), 考える(kangaeru)	K
변형	變える(kaeru), 切る(kiru)	TR
생리현상	死ぬ(sinu), 侵る(neru)	HC

피동	泣かされる(nakasareru)	PA
감정	喜ぶ(yorokobu), 哀しい (kanasii)	F

(표 2)

명사와 용언을 나타내는 변수들이 각각 의미 클래스의 값을 취한다고 하면 변수 V는 13 개, $N1$, $N2$는 22 개의 값을 취하는 이산 변수이다. 이렇게 구성된 베이지안 네트워크는 내부에 사이클(cycle)을 포함하지 않는다. 따라서 변수들이 무제한 다항 분포(unrestricted multinomial distribution)를 따른다는 가정을 하면 3 장에서 설명한 학습 기법과 추론 기법을 그대로 이용할 수 있게 된다.

4.2 네트워크의 학습

베이지안 네트워크의 구조가 확정되었으면 가지고 있는 말뭉치 데이터를 이용해서 베이지안 네트워크가 가져야 할 확률 분포를 학습해야 한다. 네트워크의 노드는 V, $J1$, $J2$, $N1$, $N2$의 다섯 가지가 있으며 이들은 각각 이산 변수들이다. 각 변수들이 서로 독립임을 이용해서 학습을 할 수 있다. 그러나 문제는 말뭉치(corpus)에서 수집한 데이터가 불완전 데이터(incomplete data)라는 점이다. 완전 데이터(complete data)는 데이터의 모든 경우에 대해서 각 변수들이 모두 값을 가지고 있는 데이터이다. 그러나 이 논문에서 사용되는 데이터는 근본적으로 불완전 데이터일 수 밖에 없다. 그 이유는 언어를 사용할 때 생략이 빈번히 일어나기 때문이다. 그 예는 다음과 같다.

3 時から始まる.

위의 문장에서 V에 해당하는 값은 동사 始まる(hazimaru)의 의미 클래스인 TR 이고, N**1** 에 해당하는 값은 명사 時(zi)의 의미 클래스에 해당하는 T 이다. 그런데 $J2$, $N2$ 에 해당하는 값은 나와 있지 않다. 따라서 불완전한 경우(incomplete case)인 것이다.

불완전한 데이터를 학습하기 위해서 이 논문에서는 몬테-카를로(Monte-Carlo) 기법 중의 하나인 깁스 샘플 추출(Gibbs sampling)을 이용한다. 이 방법의 대략적인 개요는 다음과 같다.

깁스 샘플 추출(Gibbs sampling)

주어진 변수들 $\mathbf{X} = \{X_1, .., X_n\}$의 결합 확률 분포(joint probability distribution) $p(\mathbf{x})$에 대해서, 깁스 샘플러(Gibbs sampler)를 가지고 $p(\mathbf{x})$에 관한 함수인 $f(\mathbf{x})$의 기대 값의 근사치를 구할 수 있다. 방법은 다음과 같다.

(1) $\mathbf{X}$ 의 변수들의 초기 상태를 무작위로 할당(assign)한다.
(2) 변수들 중 하나인 X_i 를 정해서 그 변수의 상태를 무효화(unassign)하고 다른 $n-1$ 개의 변수들의 상태를 이용해서 그 변수의 값의 분포를 계산한다.
(3) 그리고 그 분포를 이용해서 X_i 의 값을 추출하고 $f(\mathbf{x})$의 값을 구한다.
(4) (2), (3)을 반복하며 $f(\mathbf{x})$의 평균값을 계산한다.

두 가지 조건을 만족할 경우, 경우의 수가 무한히 커지면 $f(\mathbf{x})$의 평균값은 기대 값 $E_{p(\mathbf{x})}(f(\mathbf{x}))$와 같아진다. 두 가지 조건은 깁스 샘플러(Gibbs sampler)가 축소 불가능(irreducible)해야 한다는 것과 각 변수들이 무한히 많이 선택되어야 한다는 것이다. 깁스 샘플러가 축소 불가능하다는 것은 초기에 주어진 $\mathbf{X}$ 의 상태(configuration)에서 시작하여 다른 어떤 상태(configuration)라도 나올 수 있다는

것이다.

네트워크의 학습 알고리즘은 다음과 같다.

우선 모수들의 사전 확률 분포(prior probability distribution)를 평가한다. 여기에는 등가 샘플 크기(equivalent sample size)를 5 로 한 등가 샘플(equivalent samples) 방법[Heckerman 95]이 이용되었다.

그리고는 깁스 샘플 추출(Gibbs sampling)을 이용하여 말뭉치(corpus) 데이터로부터 네트워크의 확률 분포를 학습한다. 이 과정은 다음과 같다. 주어진 말뭉치 데이터를 불완전 데이터(incomplete data) $D = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$이라고 한다. 여기서 $\mathbf{y}_l$ 은 하나의 불완전한 경우(incomplete case)이다. 각각의 경우들에 대해서 관찰되지 않은 변수들을 임의로 초기화한다. 이 결과로 완전한 데이터 D_c 가 생긴다. 그 다음 원래의 데이터 D 에서는 관찰되지 않았던 변수 X_{il} 의 상태를 다음의 확률 분포에 따라서 재할당한다.

$$p(x'_{il} \mid D_c \setminus x_{il}, S^h) = \frac{p(x'_{il}, D_c \setminus x_{il} \mid S^h)}{\sum_{x''_{il}} p(x''_{il}, D_c \setminus x_{il} \mid S^h)} \qquad \text{(수식 31)}$$

위의 식에서 $D_c \setminus x_{il}$ 은 x_{il} 이 제거된 데이터이며 분모의 합은 변수 X_{il} 이 가질 수 있는 모든 경우에 대해서 계산된다. 위의 식은 (수식 18)에 의해서 다음과 같이 계산될 수 있다.

$$p(D \mid S^h) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(a_{ij})}{\Gamma(a_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + N_{ijk})}{\Gamma(a_{ijk})} \qquad \text{(수식 32)}$$

위의 식에서 i 는 샘플 벡터에서 나타나는 변수의 위치를 가리키고 j 는 그

변수의 부모 변수들의 벡터가 가질 수 있는 상태(configuration) 중 어느 하나를 가리키며 k 는 그 변수가 가질 수 있는 값들 중 어느 하나를 가리킨다. 위의 재할당을 관찰되지 않았던 모든 변수들에 대해서 행한다. 이 결과로 새로운 완전데이터(complete data) D'_c 를 가지게 된다. 그리고 (수식 25), (수식 27)에 의하여 사후 확률 밀도(posterior probability density) $p(\mathbf{z}_s|D'_c, S^h)$를 계산한다. 위의 과정을 1000 번 반복하여 얻은 평균값을 네트워크가 학습한 사후 확률 밀도로 사용하였다.

4.3 네트워크로부터의 추론

네트워크의 학습이 끝났으면 주어진 네트워크를 실제 문제에 적용할 수 있다. 조사의 번역이 필요한 경우에 주어질 수 있는 경우의 변수의 상태는 다음과 같은 4 가지가 있을 수 있다. 각 예는 일본어 조사 'から(kara)' 와 'で(de)'의 경우이다.

(1) V, $J2$, $N2$, $N1$ 이 주어지는 경우

예) 窓からごみを捨ててはいけない.
V = 捨てる(suteru), 의미 클래스 D
$J2$ = を(wo)
$N2$ = ごみ(gomi), 의미 클래스 CC
$N1$ = 窓(mado), 의미 클래스 CC

(2) V, $N1$ 이 주어지는 경우

예) 女の立場から言えば,
V = 言う(iu), 의미 클래스 S

$N1$ = 立場(tachiba), 의미 클래스 CS

(3) $N1$, $J2$, $N2$ 가 주어지는 경우

예) 東京から京都まで
$N1$ = 東京(tokyo), 의미 클래스 P
$J2$ = まで(made)
$N2$ = 京都(kyoto), 의미 클래스 P

(4) $N1$ 이 주어지는 경우

예) 10 月で
$N1$ = 月(gatsu), 의미 클래스 T

위의 4 가지의 경우에 대해서 변수 $J1$ 의 확률 분포를 구하는 식은 다음과 같이 각각 유도된다.

(1)의 경우는 변수 네트워크의 구조에 의해서 $J2$, $N2$ 와 $J1$ 은 서로 독립이다. 따라서 $J1$ 의 값에 영향을 주는 변수는 V 와 $N1$ 이다. 따라서 네트워크의 구조와 확률의 법칙에 의해 다음과 같이 $J1$ 의 값을 구할 수 있다.

$$
\begin{aligned}
J1 &= \mathop{\mathbf{argmax}}_{J1}\ p(J1 \mid V, J2, N1, N2) \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(J1 \mid V, N1) \\
&= \mathop{\mathbf{argmax}}_{J1} \frac{p(V, N1 \mid J1)\, p(J1)}{p(V, N1)} \\
&= \mathop{\mathbf{argmax}}_{J1} \frac{p(V, N1 \mid J1)\, p(J1)}{p(V)\, p(N1)} \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(V \mid J1)\, p(N1 \mid J1)\, p(J1) \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(V)\, p(J1 \mid V)\, p(N1 \mid J1)
\end{aligned}
\qquad \text{(수식 33)}
$$

(2)의 경우는 (1)과 같다.

(3)의 경우는 네트워크의 구조상 $J2$, $N1$ 만이 $J1$ 에 영향을 준다. 따라서 다음과 같이 $J1$ 의 값을 구할 수 있다.

$$
\begin{aligned}
J1 &= \mathop{\mathbf{argmax}}_{J1}\ p(J1 \mid N1, J2) \\
&= \mathop{\mathbf{argmax}}_{J1} \frac{p(N1, J2 \mid J1)\, p(J1)}{p(N1, J2)} \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(N1, J2 \mid J1)\, p(J1) \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(N1 \mid J1)\, p(J2 \mid J1)\, p(J1) \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(N1 \mid J1)\, p(J1, J2) \\
&= \mathop{\mathbf{argmax}}_{J1}\ p(N1 \mid J1) \sum_{V} p(J1)\, p(J2)
\end{aligned}
\qquad \text{(수식 34)}
$$

(4)의 경우는 다음과 같이 구해진다.

$$\begin{aligned} J1 &= \underset{J1}{\mathbf{argmax}}\; p(J1 \mid N1) \\ &= \underset{J1}{\mathbf{argmax}}\; p(N1 \mid J1)\, p(J1) \end{aligned} \qquad \text{(수식 35)}$$

위의 방법으로 주어진 문맥(context)에서의 조사 $J1$ 의 값을 구할 수 있다.

5. 실험

5.1 데이터의 수집

본 실험에서 이용한 데이터는 일본의 아사히 신문 및 혼다 자동차 홈페이지에서 수집한 약 7000 문장이다. 이들 중 각 조사가 출현하는 문장을 따로 모아서 각각의 실험에 사용했다. 각 조사가 출현하는 문장을 모았으면 데이터에 대한 선처리(preprocessing)가 필요하다. 우선 각 경우(case)에 대해서 실제 문장에 나타난 변수 V, $N1$, $J2$, $N2$ 의 값을 찾아야 한다. 여기에는 몇 가지의 경험적 방법이 사용되었다. 데이터의 경우는 각 조사마다 다른데 실험에 사용된 조사는 'から(kara)', 'で(de)' 의 두 가지이다. 각 조사가 나타난 문장은 각각 723 문장과 2367 문장이다.

5.2 실험 결과

4 장에서 설명한 방법대로 두 개의 조사 'から(kara)', 'で(de)' 에 대해서 실험을 했다. 각 문장 중 80%만을 네트워크의 학습에 이용했다. 번역 성공률은 전 문장에 나타난 조사를 대상으로 검토되었다. 그 결과는 아래의 (표 3)과 같다.

다의성 조사	의미	출현빈도	성공횟수	번역률(%)
から(kara)	부터	144 회	128 회	88.9%
	(으)로+부터	173 회	156 회	90.2%
	에게서	44 회	39 회	88.6%
	에서	217 회	189 회	87.1%
	(으)로	145 회	127 회	87.6%
	합계	723 회	639 회	88.4%

で(de)	에서	996 회	907 회	91.1%
	(으)로	1203 회	1089 회	90.5%
	에	168 회	134 회	79.8%
	합계	2367 회	2130 회	89.9%

(표 3)

'から(kara)' 의 경우는 문장에 나타난 723 회의 경우 중에 한국어로 번역했을 때 자연스러운 번역이 되는 경우가 '부터', '(으)로부터', '에게서', '에서', '으로' 각각 144, 173, 44, 217, 145 회였으며 이를 올바로 번역한 경우의 횟수는 각각 128, 156, 39, 189, 127 회였다. 이를 종합한 번역 성공률은 88.4%이다. 'で(de)' 의 경우는 한국어로 번역했을 때 자연스러운 번역이 되는 경우가 '에서', '(으)로', '에' 각각 996, 1203, 168 회였으며 이에 대한 번역 성공 횟수는 각각 907, 1089, 134 회이다. 이 경우의 번역 성공률은 88.9%이다.

이번 실험에서 나타난 결과를 예전의 방법과 비교하였다. 그 결과는 (표 4) 에 있다.

	から(kara)	で(de)
가장 빈도수가 높은 역어로 번역	30.0%	50.8%
기존의 연어 매칭 방식	59.8%	70.3%
베이지안 네트워크를 이용한 방식	88.4%	89.9%

(표 4)

위의 비교표에서 가장 빈도수가 높은 역어로 번역한 결과는 'から(kara)' 의 경우는 위의 실험 데이터에서 가장 많이 나타난 한국어 대역어인 '에서' 로만 번역했을 때, 'で(de)' 의 경우는 '(으)로' 로만 번역했을 때의 번역 성공률이며, 기존의 연어 매칭 방식은 기존의 일한 기계 번역 시스템에서 사용하던 방식으

로서 일어일문학 전공자가 직접 입력한 연어 패턴을 이용해서 번역한 결과이다. 연어 패턴은 동사 변환 사전, 변환 규칙에 입력되어 있다.

6. 결론

이 논문에서는 일한 기계 번역시의 조사 번역의 다의성 문제를 해소하기 위해서 베이지안 네트워크(Bayesian network)를 이용하여 대량의 말뭉치(corpus)에서 문맥(context)과 조사의 대역어간의 확률적 관계를 학습하는 방법을 취했다. 그 결과 예전의 방법보다는 더 좋은 번역 결과를 가져왔다. 기존의 연어 매칭 방식도 해당 연어 패턴의 수를 늘리면 베이지안 네트워크를 이용한 방식과 비슷한 번역률을 보일 수 있을 것이다. 하지만 베이지안 네트워크의 경우 학습 데이터를 늘림으로써 자동적으로 번역률을 올릴 수 있다.

이번 실험에서 결과가 더 좋지 못 했던 까닭은 다음의 두 가지라고 볼 수 있다. 하나는 더욱 큰 대량의 말뭉치(corpus) 데이터가 필요하다는 점이다. 베이지안 네트워크의 경우 데이터가 늘어날수록 네트워크가 표현하는 확률을 더 정확히 학습할 수 있다. 또 하나는 변수의 값이 정확하게 들어가지 않을 수 있다는 점이다. 즉 일본어 문장에 대해서 구문 분석(parsing)을 행하지 않기 때문에 해당하는 변수들의 값을 자동으로 찾는 것이 100% 정확하다고 확신할 수 없다. 이 두 가지 문제점이 해결된다면 더 좋은 결과를 가져올 것이다.

또한, 이러한 접근 방법은 일한 기계 번역에서의 조사 번역 뿐만 아니라 거의 모든 경우의 단어의 다의성 해소(word sense disambiguation)에 이용될 수 있다. 대량의 말뭉치 데이터를 가지고 언어의 각 성분간의 확률적 관계를 알아낼 수 있다고 본다. 이를 위해서는 각 언어의 대량의 말뭉치 보유와 구문 분석에 대한 더욱 더 심도 있는 연구가 필요하다고 본다.

참고문헌

[김영택 94] 김영택, "자연언어처리", 교학사, 1994.

[김정인 97] 김정인, 문경희, 이종혁, 이근배, "일한 기계번역에서 양상자질테이블 구동 방식에 의한 한국어 술부 생성", 한국정보과학회논문지, 제 24 권, 11 호, pp.1284-1296, 1997.

[문경희 98] 문경희, 이종혁, 김정인, 양기주, "일-한 기계 번역 시스템: 연어 패턴을 이용한 어휘 다의성 해소", 한국정보과학회논문지, 제 25 권, 8 호, pp.1270-1280, 1998.

[문유진 96] 문유진, 의미론적 어휘개념에 기반한 한국어의 명사 Wordnet 의 설계와 구축, 서울대학교 대학원 컴퓨터공학과 공학박사 학위논문, 1996.

[박태진 98] 박태진, 한일 기계번역에서 형태소 단위의 숙어 처리, 서울대학교 대학원 컴퓨터공학과 공학석사 학위논문, 1998.

[이종우 96] 이종우, 일한 기계 번역기에서의 변환 기법에 관한 연구, 서울대학교 대학원 컴퓨터공학과 공학석사 학위논문, 1996.

[지경래 95] 지경래, "일본어 문법 개설", 태학사, 1995.

[황호성 98] 황호성, 한영 기계번역에서 부사격 조사의 번역, 서울대학교 대학원 컴퓨터공학과 공학석사 학위논문, 1998.

[金泰錫 91] 金泰錫, 浦昭二, 日韓兩國語間の類似性を利用した機械飜譯, Technical Report, No. 91001, 慶應義塾大學管理工學科, 1991.

[大野 82] 大野普, 西正人, "類語新辭典", 角川書店:東京, 1982.

[寺村 90] 寺村秀夫, "日本語의 構文과 意味", 법문사:서울, 1990.

[Allen 95] Allen, J., *Natural Language Understanding*, Benjamin Cummings, 1995.

[Brown 91] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D. and Mercer, R. L., Word-sense disambiguation using statistical method, In *Proceedings of Association for Computational Linguistics*, ACL, pp.264-270, 1991.

[Charniak 91] Charniak, E., Bayesian Networks without Tears, *AI Magazine*, Vol. 12, No. 4, pp.50-63, 1991.

[Charniak 94] Charniak, E., *Statistical Language Learning*, Cambridge, MA:MIT Press, 1994.

[Cherkassky 98] Cherkassky, V. and Mulier, F., *LEARNING FROM DATA Concepts, Theory, and Methods*, John Wiley & Sons, 1998.

[Cooper 90] Cooper, G., Computational complexity of probabilistic inference using Bayesian belief networks (Research note), *Artificial Intelligence*, Vol. 42, pp.393-405, 1990.

[Gale 92] Gale, W. A., Church, K. W. and Yarowsky, D. A., A method for disambiguating word senses in a large corpus, *Computers and the Humanities*, 1992.

[Heckerman 95] Heckerman, D., A Tutorial on Learning With Bayesian Networks, Technical Report MSR-TR-95-06, Microsoft Research, Redmond, WA, 1995.

[Howard 81] Howard, R. and Matheson, J., Influence diagrams, In Howard, R. and Matheson, J. editors, *Readings on the Principles and Applications of Decision Analysis*, Vol. 2, pp.721-762, Strategic Decisions Groups, Menlo Park, CA, 1981.

[Jensen 90] Jensen, F., Lauritzen, S. and Olsen, K., Bayesian updating in recursive graphical models by local computations, *Computational Statistics Quarterly*, Vol. 4, pp.269-282, 1990.

[Jensen 96] Jensen, F., *An Introduction to Bayesian Networks*, Springer, 1996.

[Lauritzen 88] Lauritzen, S. and Spiegelhalter, D., Local computation with probabilities on graphical structures and their application to expert systems, *J. Royal Statistical Society B*, Vol. 50, pp.157-224, 1988.

[Mitchell 97] Mitchell, T. M., *MACHINE LEARNING*, McGraw-Hill, 1997.

[Pearl 87] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA:Morgan Kaufmann, 1987.

[Shachter 88] Shachter, R., Probabilistic inference and influence diagrams, *Operations Research*, Vol. 36, pp.589-604, 1988.

[Spiegelhalter and Lauritzen 90] Spiegelhalter, D. and Lauritzen, S., Sequential updating of conditional probabilities on directed graphical structure, *Networks*, Vol. 20, pp.579-605,990.