영한 기계 번역에서 미가공 텍스트 데이터를 이용한 대역어 선택 중의성 해소

김 유 섭[†] · 장 정 호[††]

요 약

본 논문에서는 미가공 말뭉치 데이터를 활용하여 영한 기계번역 시스템의 대역어 선택시 발생하는 중의성을 해소하는 방법을 제안한다. 이를 위하여 은닉 의미 분석(Latent Semantic Analysis : LSA)과 확률적 은닉 의미 분석(Probabilistic LSA : PLSA)을 적용한다. 이 두 기법은 텍스트 문단과 같은 문맥 정보가 주어졌을 때, 이 문맥이 내포하고 있는 복잡한 의미 구조를 표현할 수 있다. 본 논문에서는 이들을 사용하여 언어적인 의미 지식(Semantic Knowledge)을 구축하였으며 이 지식은 결국 영한 기계번역에서의 대역어 선택시 발생하는 중의성을 해소하기 위하여 단어간 의미 유사도를 추정하는데 사용된다. 또한 대역어 선택을 위해서는 미리 사전에 저장된 문법 관계를 활용하여야 한다. 본 논문에서는 이러한 대역어 선택시 발생하는 데이터 희소성 문제를 해소하기 위하여 k-최근점 학습 알고리즘을 사용한다. 그리고 위의 두 모델을 활용하여 k-최근점 학습에서 필요한 예제 간 거리를 추정하였다. 실험에서는, 두 기법에서의 은닉 의미 공간을 구성하기 위하여 TREC 데이터(AP news)를 활용하였고, 대역어 선택의 정확도를 평가하기 위하여 Wall Street Journal 말뭉치를 사용하였다. 그리고 은닉 의미 분석을 통하여 대역어 선택의 정확성이 디폴트 의미 선택과 비교하여 약 10% 향상되었으며 PLSA가 LSA보다 근소하게 더 좋은 성능을 보였다. 또한 은닉 공간에서의 축소된 벡터의 차원수와 k-최근점 학습에서의 k값이 대역어 선택의 정확도에 미치는 영향을 대역어 선택 정확도와의 상관관계를 계산함으로써 검증하였다.

Target Word Selection Disambiguation using Untagged Text Data in English-Korean Machine Translation

Yu-Seop Kim[†] · Jeong-Ho Chang[††]

ABSTRACT

In this paper, we propose a new method utilizing only raw corpus without additional human effort for disambiguation of target word selection in English-Korean machine translation. We use two data-driven techniques; one is the Latent Semantic Analysis(LSA) and the other the Probabilistic Latent Semantic Analysis(PLSA). These two techniques can represent complex semantic structures in given contexts like text passages. We construct linguistic semantic knowledge by using the two techniques and use the knowledge for target word selection in English-Korean machine translation. For target word selection, we utilize a grammatical relationship stored in a dictionary. We use k-nearest neighbor learning algorithm for the resolution of data sparseness problem in target word selection and estimate the distance between instances based on these models. In experiments, we use TREC data of AP news for construction of latent semantic space and Wall Street Journal corpus for evaluation of target word selection. Through the Latent Semantic Analysis mothods, the accuracy of target word selection has improved over 10% and PLSA has showed better accuracy than LSA method. Finally we have showed the relatedness between the accuracy and two important factors ; one is dimensionality of latent space and k value of k-NN learning by using correlation calculation.

키워드 : 대역어 선택(Target Word Selecton), LSA(Latent Semantic Analysis), PLSA(Probabilistic Latent Semantic Analysis), k-최근점 학습(k-nearest Neighbor Learning), 상관관계(Correlation)

1. 서 론

단어의 의미를 선택하는데 있어 주어-술어, 목적어-술어 등과 같이 구문적으로 공기(co-occur)하는 단어들을 상호 의미 선정에 있어 문맥 정보로 활용할 수 있다. 기계 번역에서는 목표언어에서의 대역어를 선택하기 위한 방법으로서, 연어 정보가 많이 이용되어 왔다. [1]에서는 히브리어-영어 기계 번역 시스템에서 연어 정보를 이용한 통계적 의미모호성 해소 기법을 제안하였으며, [2]에서는 한영 기계번역 시스템에서의 영어 단어 선택을 위한 연어사전 기반의 방법을

※ 이 논문은 2003년도 한림대학교 교비연구비(HRF-2003-44)에 의하여 연구되었음.
† 종신회원 : 한림대학교 정보통신공학부 조교수
†† 준 회 원 : 서울대학교 컴퓨터공학부 박사과정
논문접수 : 2004년 4월 24일, 심사완료 : 2004년 8월 4일

제안하였다. 하지만, [1]에서 제안한 방법은 실제 말뭉치에서 나타나지 않은 경우에 대해서는 처리할 수 없는 단점을 가지고 있었고, [2]에서는 연어 사전을 구축할 때 수동적 구축 방식을 적용함으로써 비용과 일관성 문제가 제시되었다. 또한 이들 방법을 구현할 때 아무리 대규모의 말뭉치를 이용하더라도 데이터 희소성 문제는 여전히 존재한다는 문제가 있었다[3].

데이터 희소성 문제을 해결하기 위하여 [3]에서는 어휘 유사도를 말뭉치에 나타난 단어들의 실례를 가지고 통계적으로 계산하였는데, 이 때 문맥정보가 거의 배제된 상태에서 유사도를 계산하였기 때문에 다의어 및 동형이의어 처리에 문제가 있었다. 또한 [4]에서는 워드넷(WordNet)을 활용하여 유사도를 계산하여 대역어 선택을 하였는데, 워드넷은 특정 도메인에 특화하여 구축할 수 없고, 다의어의 처리에 추가적인 비용이 필요하여 비용 및 효율에 문제를 가지고 있었다.

그리고 위에서 설명한 방법론을 구현하기 위해서는, 언어처리와 관련하여 유용하게 활용되고 있는 시소러스(thesaurus), 태그된 말뭉치(tagged corpora), 전자 사전(machine-readable dictionary) 등이 필요하다. 그러나 이들 자원을 구축하는 일은 많은 비용이 드는 작업일 뿐더러 작업 자체에 인간의 직관이 크게 관여하기 때문에 그 일관성 및 무결성을 보장하는 것은 매우 어렵다. 따라서 다양한 도메인(domain)에 적합한 자원을 각각 구축하는 것은 거의 불가능하다. 이에 반하여 본 논문에서 활용할 데이터 기반 자율(unsupervised) 학습 방법은 인간의 지식, 추가적인 지식 베이스, 시소러스 또는 구문 파서와 같은 다양한 형태 및 내용의 추가 자원이 필요없이 미가공된 텍스트 데이터만을 가지고 원하고자 하는 자원을 구축할 수 있다. 본 논문에서는 이를 위하여 두 가지 기법을 활용하는데, 하나는 은닉 의미 분석(Latent Semantic Analysis : LSA)[5]이고 다른 하나는 이 모델의 프로토타입을 확률적인 모델로 재구성한 확률적 은닉 의미 분석(Probabilistic Latent Semantic Analysis : PLSA)[10]이다.

LSA는 주어진 문맥에서 나타난 단어들의 의미를 추출하고 표현하는 데 활용되는 모델이다. LSA는 단어 또는 텍스트 문서들끼리 대략의 의미 유사도를 추정할 수 있는 편리한 방법 정도로 생각할 수 있다. 그리고 이 모델은 인덱싱, 문서 일관성 측정, 또는 여러 자연언어처리 응용에서 활용되고 있다[6-8]. LSA는 여러 실험을 통하여 인간의 인지 현상들과 많은 관련을 가지고 있다고 알려져 있다. 또한 LSA는 기본 벡터공간을 은닉 공간으로 변환한 결과 벡터의 차원의 정도가 여러 응용의 성능에 있어서 매우 중요하다고 알려져 있다[9].

PLSA는 LSA에 비하여 보다 더 적절한 통계학적인 근간 위에서 설명될 수 있으며, 텍스트 문서에 대해 더 적당한 데이터 생성 모델을 정의한다[10]. PLSA 기법은 양상 모델(aspect model)에 기반하고 있으며 양상 모델은 공기 데이터에 대한 은닉 변수 모델로서, 이는 관측된 변수와 미관측된 부류(class) 변수를 서로 연관시켜 주는 모델이다[11]. 이 모델은 LSA와 마찬가지로 정보 검색의 색인 분야에서 활용되었다[12]. 또한 이는 다항 분포 공간에서 차원 축소를 수행하며 연결된 부분 단순구조(sub-simplex)는 확률 은닉 의미 공간을 통하여 식별될 수 있다[10].

결과적으로 이들 두 모델은 어휘 간 또는 문서 간의 유사도를 추정하는 데 활용될 수 있는데, 이들을 가지고 문맥의 의미를 가장 잘 표현하는 주제(topic)를 은닉 의미 구조라고 정의하고 이들 주제와 관련하여 어휘 및 문서 간 유사도를 추정한다.

본 논문에서는 영한 기계 번역에서의 대역어 선택에서 앞의 두 모델을 적용하여 보다 정확한 선택을 가능하게 하였다. 또한 두 모델의 성능을 비교하여 각각의 특성을 파악하였고 은닉 공간 상에서의 축소된 어휘 벡터의 차원수와 선택 정확도 간의 상관관계와 k-최근점 학습에서의 k값과 선택 정확도 간의 상관관계를 파악하고자 하였다. 이 과정을 보다 상세히 설명하면 다음과 같다.

첫째, 두 단어 간의 문법적인 관계를 표현하는 투플(tuple)을 저장하는 사전을 구축하였다. 여기서 문법적인 관계는 주어-자동사, 타동사-목적어, 그리고 형용사-명사로 그 범위를 제한하였다. 둘째, 대역어 선택을 위한 입력은 역시 투플로 구성되는데 이 중 한 단어는 번역되어야 하는 단어이고 나머지 한 단어는 인자로 활용되는 단어이다. 이 때 하나의 투플이 입력되면 인자로 사용되는 단어로 사전을 검색한다. 셋째, 만일 인자어가 사전에 등록되어 있지 않으면 k-최근점 학습 알고리즘을 사용하여 입력된 단어의 대역어로 어떤 대역어 부류가 가장 적절한지 결정한다. 여기서 최근점을 결정하기 위해서는 단어 간의 거리를 측정할 수 있어야 하는데, 본 논문에서는 위의 은닉 의미 분석 모델을 활용하여 그 거리를 측정하였다.

실험에서는 1988년도 AP 뉴스 말뭉치를 TREC-7 데이터[13]에서 얻어서 은닉 의미 구조를 구축하고자 하였다. 그리고 Wall Street Journal 말뭉치를 활용하여 사전과 테스트 세트를 구축하였다. 우리는 8만 여 개의 문서에서 2만 여 종의 단어를 추출하여 은닉 의미 구조를 구축하였다. 이 실험에서 디폴트 대역어를 선택한 경우에 비하여 약 10%이상의 선택 정확도의 향상을 보여 주었고 대체적으로는 PLSA가 LSA보다 근소하게 더 좋은 성능을 보여주었다. 또한 의미 공간의 차원수와 $k-$최근점 학습의 $k-$값을 대역어 선택 정확도와의 상관관계를 계산하여 각 요소들이 대역어 선택에 미치는 영향을 분석하였다.

다음 장에서는 대역어 선택의 주요 과정을 상세히 설명하며 또한 문법적 관계의 정의 및 투플의 구성과 실제 사전의

예를 제시한다. 그리고 $k-$최근접 학습 알고리즘도 설명한다. 3장에서는 두개의 은닉 의미 모델에 대하여 이론적으로 설명하며 4장에서는 본 연구를 위하여 이루어진 다양한 형태의 실험과 결과를 제시한다. 마지막으로 5장에서는 결론 및 향후 과제에 대하여 언급한다.

2. 대역어 선택 과정

2.1 문법적 관계 정의

본 논문에서는 적절한 대역어를 선택하기 위하여 사전의 형태로 저장되어 있는 문법적 관계를 활용하였다. [1]에서는 단어 간의 문법적인 관계를 표현하기 위하여 구문 투플(syntactic tuple)을 사용하였는데 이들 투플은 대역어 선택에 필요한 매우 주요한 문맥 정보를 가지고 있었다. 이들 문법적 관계는 다음과 같은 방식으로 사전에 기술되어 있다.

$$T(S_i)=\begin{cases} T_1 & \textbf{if} \quad Cooc(S_i, S_1) \\ T_2 & \textbf{if} \quad Cooc(S_i, S_2) \\ \quad \cdots & \\ T_n & \textbf{otherwise} \end{cases}$$

여기서 $Cooc(S_i, S_j)$는 원시 단어 S_i와 S_j가 문법적으로 공기하는지 여부를 표현하는 함수인데, 두 단어 중 S_i는 번역이 될 입력단어(input word)가 되고 또 S_j는 대역어 선택에 있어서 사용될 인자단어(argument word)가 된다. 그리고 T_j는 원시 단어 S_i를 번역한 결과, 즉 한국어 대역어가 된다. 그리고 $T(\,)$는 대역어 선택 과정을 가리킨다.

<표 1>은 영어의 동사 단어 S_i='*build*'와 그 목적어 명사들에 대한 문법적인 관계를 보여주고 있다. 여기서는 동사 단어 '*build*'가 입력단어가 되고 동사의 목적어 단어들이 인자단어가 된다. 이러한 문법 관계는 사전의 형식으로 저장되는데 <표 1>의 사전은 단어 '*build*'가 문법 관계에 있는 목적어 단어에 따라서 서로 다른 5개의 한국어 대역어를 취할 수 있음을 보여주고 있다. 예를 들어, '*build*'는 목적어로 '*plant*'(*공장*)을 취할 때 '*건설하다*'로 번역될 수 있으며 '*car*'(*자동차*)를 목적어로 취할 때는 '*제작하다*'로, 그리고 '*company*'(*회사*)를 목적어로 취할 때는 '*설립하다*'로 번역될 수 있다.

〈표 1〉 동사 '*build*'와 공기하는 단어들

'*build*'의 대역어(T_j)	공기하는 목적어 단어(S_j)
건설하다	plant, facility, network, …
건축하다	house, center, housing, …
제작하다	car, ship, model, …
설립하다	company, market, empire, …
구축하다	system, stake, relationship, …

그러나 이러한 방식으로 대역어를 선택할 경우에는 데이터 희소성 문제를 해결해야 한다. 즉, 이미 구축되어있는 사전에 등록되지 않은 단어가 인자 단어로 입력될 경우에는 위의 사전만을 가지고는 대역어를 선택할 수 없다. 물론 사전에 공기하는 단어 리스트를 무제한 열거할 수도 있겠으나 이는 사실상 불가능하다. 이처럼 사전에 열거되어 있지 않은 단어가 인자단어로 입력될 경우에 입력단어의 대역어를 선택하기 위해서는 이미 구축되어 있는 사전을 학습예제 공간으로 가정하는 $k-$최근점 학습 방법을 사용하여 가장 적절한 대역어를 선택할 수 있다. 다음절에서는 $k-$최근점 학습에 대하여 자세히 설명하고 있다.

2.2 $k-$최근점 학습

<표 1>을 다른 방식으로 보면, 5개의 '*build*'의 대역어는 각기 하나의 부류(class)가 되고 공기하는 단어들은 해당 부류로 분류되는 예제(example)가 된다. 입력된 인자단어가 사전에 나열되어 있지 않은 경우에는 이미 나열되어 있는 단어(예제)들을 분석하여 인자단어의 부류(class)를 유추해서 대역어를 선택할 수 있다. 이 때 유추를 위하여 본 논문에서는 $k-$최근점 학습 방법을 사용하였다. $k-$최근점 학습 알고리즘[14-15]에서는 먼저 모든 예제들은 $n-$차원의 공간 R^n상의 한 점에 대응된다고 가정한다. 이 때, $n-$차원 공간 상의 좌표는 각 예제로서의 단어를 표현하는 $n-$차원 벡터로 매핑시킬 수 있다. 그리고 한 예제의 최근점은 표준 유클리드 거리(Euclidean distance)를 사용하여 결정된다. 두 예제 x_i와 x_j간의 거리 $D(x_i, x_j)$는 다음과 같이 정의된다.

$$D(x_i, x_j)=\sqrt{\sum_{k=0}^{l}(a_k(x_i)-a_k(x_j))^2}$$

여기서 $a_k(x_i)$는 x_i의 k번째 속성값을 나타내는데 $0\le k\le l$이고 l은 전체 속성의 개수이다. 이 계산은 두 벡터 간의 코사인(cosine) 연산과 유사하다. $k-$최근점 학습은 목표함수의 결과로써 연속적인 값을 취하는 경우와 이산값을 취하는 경우가 있는데 본 논문에서는 이산값을 그 대상으로 하고, $f: R^n \rightarrow V$와 같이 표현될 수 있고 V는 유한 집합 $\{v_1, \cdots, v_s\}$으로써 n차원 공간의 한 점을 함수 f에 적용한 출력값을 의미한다. $k-$최근점 학습 알고리즘은 이산값 목표 함수를 근사화(approximating)하는 것으로써 다음과 같이 정의된다.

- **학습**
  : 모든 훈련 예제 $\langle x, f(x)\rangle$에 대하여, 각각의 예제를 **학습_예제** 리스트에 추가한다.
- **분류**
  : 질의 예제 x_q를 분류시키기 위해서는

- **학습_예제**에서 x_q와 가장 가까운 k개의 예제 $x_1, \cdots, x_k$를 찾아낸다.
- 다음을 리턴한다.

$$\hat{f}(x_q) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(x_i))$$

여기서 $a = b$이면 $\delta(a, b) = 1$이고 그렇지 않으면 $\delta(a, b) = 0$이다.

2.3 대역어 선택 예

이 절에서는 실제 사전에 등록되어 있지 않은 단어의 대역어 선택의 실례를 설명한다. 대역어 선택이 요구되는 입력단어 'build'가 입력되면 구문 분석을 통하여 인자 단어들을 추출한다. 여기서 'build'는 타동사이므로 동사의 목적어 단어가 인자 단어가 된다. 이 때, 인자 단어가 'automobile'이라고 가정하자. 단어 'automobile'은 스테밍 과정을 통하여 'automobil'로 변환되고, 현재 사전에 기술되어 있는 모든 단어들과의 의미 유사도를 추정한다. 다음 <표 2>는 단어 'automobil'과 가장 유사한 5 단어와의 의미 유사도와 각 단어들의 부류를 보여준다. 여기서 5는 k-NN의 k값에 해당하며, 의미 유사도는 LSA를 통하여 계산되었다.

〈표 2〉 단어 'automobil'과 가장 유사한 top-5 단어들

단 어	의미 유사도	부 류
car	0.025999	'제작하다'
industry	0.006838	'설립하다'
vehicle	0.004944	'제작하다'
model	0.004482	'제작하다'
engine	0.003415	'제작하다'

위 결과를 보면 단어 'automobil'과 가장 유사한 5개의 단어 중에서 4개의 단어의 부류가 '제작하다'이기 때문에 'automobil'의 부류는 '제작하다'로 결정된다. 따라서 입력 단어 'build'의 대역어는 '제작하다'가 된다.

3. 지식 추출 모델

k-최근점 학습에 있어서 가장 중요한 과정은 주어진 질의 예제와 가장 가까운 거리에 놓여진 예제들을 찾는 것이다. 본 논문에서는 예제 간 거리를 추정하고자 할 때 각 예제(단어)들의 의미 유사도를 계산하였다. 의미 유사도를 계산하기 위해서는 지식이 필요한데, 여기서는 은닉 의미 분석(Latent Semantic Analysis : LSA)과 확률적 은닉 의미 분석(Probabilistic Latent Semantic Analysis : PLSA)을 사용하여 지식을 구축하였다. 이들 방법은 기존의 미가공의 텍스트 데이터 외에는 어떠한 추가적인 인간의 지식도 불필요하다는 장점을 가지고 있다. 다음 두 절에서 LSA와 PLSA가 어떻게 두 단어 간의 의미상 유사도를 계산하는지 설명한다.

3.1 은닉 의미 분석 모델

LSA는 기본적으로 특정 문맥에서 특정 단어들이 서로 공기하는 정도가 두 단어 간의 유사도를 추정하는 데 가장 큰 정보를 제공한다는 점을 가정한다[9, 16]. LSA는 또한 특정 담화의 내용 중에서 단어들의 기대되는 문맥적 활용의 관계를 추출하고 추론한다. 또한 이것은 인간에 의해 제작된 사전, 지식 베이스, 의미 시소러스, 구문 파서와 같은 추가적인 언어 자원 및 도구는 사용하지 않고 단지 미가공된 텍스트를 입력 데이터 형식에 맞추는 도구만을 사용할 뿐이다.

첫 번째 단계에서는 주어진 텍스트 데이터를 하나의 행렬로 표현하는데, 이 때 각각의 행은 하나의 단어를 그리고 각각의 열은 하나의 독립된 문맥을 표현한다. 여기서 독립된 문맥은 하나의 문서, 문단, 문장 등 필요에 따라서 다양한 형태를 취할 수 있다. 이 행렬에서 각각의 항목은 각 단어가 각 문맥에 나타나는 빈도수를 가진다. 다음으로 LSA는 Singular Value Decomposition(SVD) 방법을 행렬에 적용시켜서 다음과 같은 세 개의 작은 행렬을 생성한다.

$$A = U \sum V^T$$

여기서 $\sum$는 AA^T 또는 A^TA의 $r(= rank(A))$개의 0이 아닌 특성값(eigenvalue)으로 구성된 대각 행렬이고 U와 V는 각각 r개의 0이 아닌 AA^T 또는 A^TA의 특성값과 연관되어 있는 직교 특성 벡터(eigenvector)이다. 부분 행렬 U는 원 행렬에서 행 항목들을 유도된 직교 벡터값으로 구성한 벡터들로서 볼 수 있으며 다른 부분 행렬 V는 마찬가지로 원래 행렬의 열 항목을 나타낸다. 그리고 세 번째 부분 행렬 $\sum$은 대각 행렬로서, 이것은 스케일 값을 가지고 있는데 이 세 행렬을 행렬곱 연산을 하면 원래의 행렬이 다시 만들어진다. 수학적으로 모든 행렬은 이와 같이 완벽하게 분해될 수 있다는 점이 증명되어 있다.

위와 같은 방법을 사용하면 원래 주어진 벡터의 차원은 단지 대각 행렬의 계수를 지움으로써 축소될 수 있다. 이를 구현하기 위해서, $k(\leq r)$개의 가장 큰 특성값들과 대응되는 단일 벡터가 새로이 k-차원 문서 공간을 정의하는데 사용된다. 이들 벡터를 이용하면 원래 $m \times r$ 행렬인 U는 $m \times k$ 행렬인 U_k로, $n \times r$ 행렬인 V는 $n \times k$ 행렬인 V_k로, 그리고 $r \times r$ 행렬인 $\sum$은 $k \times k$ 특성값 행렬 $\sum_k$로 재정의된다. 결과적으로 $A_k = U_k \sum_k V_k^T$는 k차원을 가진

〈표 3〉 은닉 의미 분석을 통한 개별 단어들의 변환 결과

단 어	초기 벡터	변환 벡터
human	(1, 0, 0, 1, 0, 0, 0, 0, 0)	(0.16, 0.40, 0.38, 0.47, 0.18, -0.05, -0.12, -0.16, -0.09)
interface	(1, 0, 1, 0, 0, 0, 0, 0, 0)	(0.14, 0.37, 0.33, 0.40, 0.16 ,-0.03,-0.07, -0.10, -0.04)
computer	(1, 1, 0, 0, 0, 0, 0, 0, 0)	(0.15, 0.51, 0.36, 0.41, 0.24, 0.02, 0.06, 0.09, 0.12)
user	(0, 1, 1, 0, 1, 0, 0, 0, 0)	(0.26, 0.84, 0.61, 0.70, 0.39, 0.03, 0.08, 0.12, 0.19)
system	(0, 1, 1, 2, 0, 0, 0, 0, 0)	(0.45, 1.23, 1.05, 1.27, 0.56, -0.07, -0.15, -0.21, -0.05)
response	(0, 1, 0, 0, 1, 0, 0, 0, 0)	(0.16, 0.58, 0.38, 0.42, 0.28, 0.06, 0.13, 0.19, 0.22)
time	(0, 1, 0, 0, 1, 0, 0, 0, 0)	(0.16, 0.58, 0.38, 0.42, 0.28, 0.06, 0.13, 0.19, 0.22)
EPS	(0, 0, 1, 1, 0, 0, 0, 0, 0)	(0.22, 0.55, 0.51, 0.63, 0.24, -0.07, -0.14, -0.20, -0.11)
survey	(0, 1, 0, 0, 0, 0, 0, 0, 1)	(0.10, 0.53, 0.23, 0.21, 0.27, 0.14, 0.31, 0.44, 0.42)
trees	(0, 0, 0, 0, 0, 1, 1, 1, 0)	(-0.06, 0.23, -0.14, -0.27, 0.14, 0.24, 0.55, 0.77, 0.66)
graph	(0, 0, 0, 0, 0, 0, 1, 1, 1)	(-0.06, 0.34, -0.15, -0.30, 0.20, 0.31, 0.69, 0.98, 0.85)
minors	(0, 0, 0, 0, 0, 0, 0, 1, 1)	(-0.04, 0.25, -0.10, -0.21, 0.15, 0.22, 0.50, 0.71, 0.62)

행렬 중 원래의 행렬과 가장 가까운 행렬이 된다. 이와 같은 방법으로 차원을 축소하면 벡터를 구성하고 있는 자질들의 의미 연관성을 반영하여 축소된 차원의 벡터로 변환된다. 이러한 특성 때문에 벡터의 차원이 축소되더라도 번역 성능에는 큰 영향을 미치지 않을 수 있다.

LSA는 이러한 방식으로 단어의 의미를 축소된 벡터로써 표현하고 있으며 이들을 사용하여 두 단어간의 의미 유사도를 계산한다. 단어간 유사도(문서 또는 문맥간 유사도도 마찬가지로)는 행렬 U_k의 행벡터들간의 내적을 계산함으로써 추정할 수 있다.(문서 또는 문맥간 유사도를 계산하려면 V_k의 행벡터들간의 내적을 계산한다) 축소된 행렬의 행벡터는 결국 축소된 은닉 공간에서의 좌표를 의미하는 것이다. 두 좌표 V_1과 V_2의 유사도를 계산하기 위한 코사인 계산 방법은 다음과 같다.

$$\cos\phi = \frac{\mathbf{V}_1 \cdot \mathbf{V}_2}{\|\mathbf{V}_1\| \cdot \|\mathbf{V}_2\|}$$

이러한 과정을 통하여 계산되는 단어간 의미 유사도는 대역어 선택 과정에서 주어진 인자단어와 가장 유사한 단어들을 사전으로부터 검색하는 데 활용된다.

<표 3>은 각 문서에서의 빈도정보를 가지는 개별 단어 벡터가 LSA를 통하여 변환되는 예를 보여주고 있다. 여기서 두 번째 열은 각 단어가 개별 문서에 나타난 빈도수로 구성한 벡터를 보여주고 있고, 세 번째 열은 초기 벡터를 LSA로 변환한 결과를 보여주고 있다. 단어간 유사도는 3번째 열의 벡터간의 코사인 계산을 통하여 계산된다. 이 과정에 대한 보다 자세한 설명은 [9]에서 찾아볼 수 있다.

3.2 확률적 은닉 의미 분석

PLSA는 두개의 모드로 구성되고 또한 공기하는 데이터를 분석하기 위한 통계적인 기법으로 언어 모델링[17]과 정보검색에서의 문서 색인[11]과 같은 응용에서 좋은 결과를 보여주고 있다. PLSA는 양상 모델(aspect model)을 기반으로 하고 있는데[18], 여기서는 실제 공기하는 데이터를 관찰함으로써 이것과 은닉 부류 변수(latent class variable) $Z=\{z_1, z_2 \cdots, z_k\}$의 연관성을 유추한다[10]. 텍스트 문서의 경우, 실제 관찰이라는 것은 단어 $w \in W$가 문서 $d \in D$에 출현하는 정도를 의미한다. 그리고 은닉 부류의 모든 가능한 상태인 $z_k (1 \le k \le K)$는 하나의 의미 토픽을 표현한다.

단어-문서의 공기 사건 (d, w)는 확률적인 방법으로 모델링되는데 여기서는 다음과 같이 나타낼 수 있다.

$$P(d,w) = \sum_z P(z)\,P(d,w \mid z)$$
$$= \sum_z P(z)\,P(w \mid z)P(d \mid z)$$

여기서 w와 d는 주어진 z에 대하여 조건부 독립이라 가정한다. $P(w|z)$와 $P(d|z)$는 각각 주어진 토픽에 대한 단어와 문서의 확률분포를 말한다. 여기서 공기하는 데이터의 분포 $P(d, w)$를 3개의 독립된 항으로 분리하여 표현하였기 때문에 LSA에서의 SVD와 유사하다고 할 수 있다. 그러나 PLSA의 목표 함수는 LSA와는 달리 다항 샘플링에서의 우도(likelihood) 함수이다.

세 개의 파라미터 $P(z), P(w|z)$ 그리고 $P(d|z)$는 모두 다음의 로그-우도(log-likelihood) 함수를 최대화함으로써 추

정될 수 있다.

$$L = \sum_{d \in D} \sum_{w \in W} n(d, w) \log P(d, w)$$

그리고 이 함수는 대부분의 은닉 변수 모델에서 그러하듯이 EM 알고리즘을 사용하여 최대화한다. 파라미터 추정과 관련하여 더 자세한 사항은 [10]을 참조하기 바란다. 이 모델에서 단어 w_1과 w_2의 의미 유사도 $sim(w_1, w_2)$는 다음과 같은 방식으로 추정되고,

$$sim = (w_1, w_2) = P(z|w_1) \times P(z|w_2)$$

$P(z|w)$는 다음의 공식에서 유도된다.

$$P(z|w) = \frac{P(z)P(w|z)}{\sum_z P(z)P(w|z)}$$

4. 실험 및 평가

4.1 공간 구성과 사전 구축을 위한 데이터

실험에서는 두 종류의 데이터가 사용되었는데, 첫 번째 데이터는 LSA와 PLSA에 의한 은닉 공간을 구성하기 위한 데이터이고, 다른 데이터는 문법적 관계를 포함하고 있는 사전을 구축하고 또한 대역어 선택 결과를 테스트하기 위한 데이터이다. TREC-7 데이터에 있는 1988년 AP 뉴스 말뭉치로부터 약 4천만 단어로 이루어진 79,919 문서를 추출하였다. 여기서 먼저 4개 이하의 문자로 이루어진 단어와 문서 형식을 나타내는 태그들을 제거하여 21,945,292 단어 텍스트 데이터를 구축하였다. 그리고 텍스트에 포함되어 있는 단어들의 어근을 추출하고 문서에서 20회 이상 나타나는 19,286 단어들을 선택하였다. 그럼으로써 텍스트 데이터의 크기는 17,071,211 단어가 되었다.

본 논문에서는 공간의 차원수와 대역어 선택 정확도와의 관련성을 분석하기 위하여 50 차원부터 300 차원까지 다양한 벡터의 차원수(PLSA의 경우에는 z의 갯수)를 조절하여 대역어 선택 결과를 분석하였다. LSA 공간을 구성하기 위하여 SVDPACK[19] 으로부터 파생된 단일 벡터 Lanczos 알고리즘[20]을 사용하였으며, PLSA는 [10, 12]에서 제시된 EM 알고리즘을 적용하였다. 지식 구축을 위하여 필요한 시간 및 공간 복잡도는 LSA에 비하여 PLSA가 월등히 높은 모습을 보여주었는데 축소된 차원의 크기에 따라 약 12배에서 60배 정도의 차이가 나타났다.

<표 4>는 3,443개의 예제 단어들로부터 임의로 추출된 5개의 단어들과 가장 의미상으로 유사한 서로 다른 5개의 단어들을 나열한 예이다. 실험을 위하여 타동사-목적어, 자동사-주어, 형용사-명사와 같은 문법적인 관계를 가지고 있는 3,443개의 문장들을 220,047개의 월 스트리트 말뭉치와 41,750개의 그 밖의 신문 말뭉치 등 총 261,797 단어 말뭉치로부터 추출하였다. <표 4>의 단어들은 여기서 인자단어로 사용된 단어들이다. <표 4>의 첫 번째 열에는 서로 다른 5개의 단어들이 나열되어 있다. 각 단어는 두 개의 행을 가지고 있는데 위의 행은 LSA 공간에서 가장 유사한 단어를 나열하고 있으며 아래 행에는 PLSA 공간에서 가장 유사한 단어들이 나열되어 있다.

전체 3,443개의 예제 중에서 2,437개는 타동사-목적어 문법 관계를 가지고 있으며, 188개는 자동사-주어 관계를, 그리고 818개는 형용사-명사 관계를 가지고 있다. 이들 관계들은 [21]의 영한 기계 번역 시스템에서 사용하고 있는 사전에서 추출하여 본 실험을 위한 사전으로 변환하여 구축하였다.

〈표 4〉 LSA와 PLSA 공간상에서 가장 유사한 단어들

선택된 단어들	가장 유사한 단어들
plant	westinghous, isocyan, shutdown, zinc, manur
	radioact, hanford, irradi, tritium, biodegrad
car	buick, oldsmobil, chevrolet, sedan, corolla
	highwai, volkswagen, sedan, vehicular, vehicle
home	parapleg, broccoli, coconut, liverpool, jamal
	memori, baxter, hanlei, corwin, headston
business	entrepreneur, corpor, custom, ventur, firm
	digit, compat, softwar, blackston, zayr
ship	vessel, sail, seamen, sank, sailor
	destroy, frogmen, maritim, skipper, vessel

그리고 대역어 선택 실험을 할 때는 5-폴드 교차 검증(5-fold cross validation) 방법을 각각의 문법적 관계에 적용시켰다. 즉 각 문법적 관계를 나타내는 샘플 문장들을 5개의 서로 교차하지 않는 샘플 집합으로 나누고, 실험에서 하나의 샘플은 테스트 샘플이 되고 나머지 4개의 샘플은 문법적 관계를 저장하고 있는 사전을 구성하는데 사용되도록 서로 결합되는 것이다. 그리고 5개의 모든 샘플들이 각각 테스트 샘플이 되어 총 5회의 실험을 반복한 후 그 결과의 평균을 구했다.

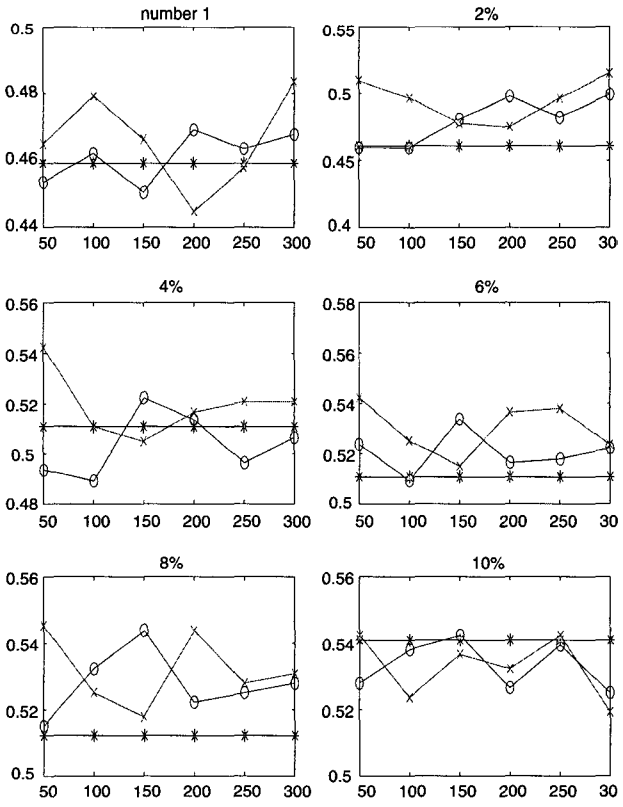
4.2 대역어 선택 결과

본 실험은 크게 두 개의 실험으로 이루어져 있다. 첫 번째 실험은 은닉 의미 분석이 대역어 선택 중의성 해소에 기여하는 정도를 측정한 실험이다. 이를 위해서, 가장 먼저 디폴트 의미(default meaning)를 사용하여 적절한 대역어를 선택하는 방법을 평가하였다. 여기서 디폴트 의미란 여러 대역어 중에서 가장 높은 빈도로 선택되어지는 대역어를 말하며 사전에 인자단어가 나열되어 있지 않은 경우에 디폴트 대역어를 무조건 선택하는 방식이다. 이러한 방식을 사용하여 약

76.81%의 정확한 대역어가 선택되었다. 두 번째 방법은 어휘를 표현하는 벡터를 전혀 가공하지 않고 원래 그대로의 원형으로 유지하고 의미 유사도를 계산하는 방법이다. 즉 모든 단어는 실험에 사용된 문서수와 동일한 차원수의 벡터를 가지고 벡터의 각 원소는 해당 단어가 해당 문서에 나타난 빈도수 그 자체를 의미한다. 이러한 방법을 이용하여 약 87.09%까지 정확한 대역어가 선택되었는데, 본 논문에서의 대역어 선택은 $k-$ 최근점 학습을 이용하여 이루어지기 때문에 k값에 따라서 그 정확도는 조금씩 달랐다. 이에 대해서는 아래에서 보다 자세하게 설명한다. 마지막으로 LSA 와 PLSA를 사용하여 어휘 유사도를 계산하여 대역어를 선택하였는데 최대 87.16%의 정확한 대역어가 선택되었다. 그런데 대역어 선택에 있어서 소요된 시간을 비교해 보면, 가공되지 않은 벡터를 사용한 경우는 LSA나 PLSA를 사용하여 기존 행렬을 가공한 경우에 비하여 약 6배에서 9배까지 더 많은 시간을 소비하는 결과를 보여주었다. 이는 인자단어가 주어졌을 때 사전에 나열되어 있는 단어들과의 유사도를 계산하는 과정에서 크고 복잡한 벡터간의 코사인 연산일수록 더 많은 시간이 소요되었기 때문이다. LSA와 PLSA의 경우에서도 결과 차원의 수와 토픽의 수가 작을 수록 대역어 선택에 필요한 시간은 점점 줄어드는 모습을 보여주었다.

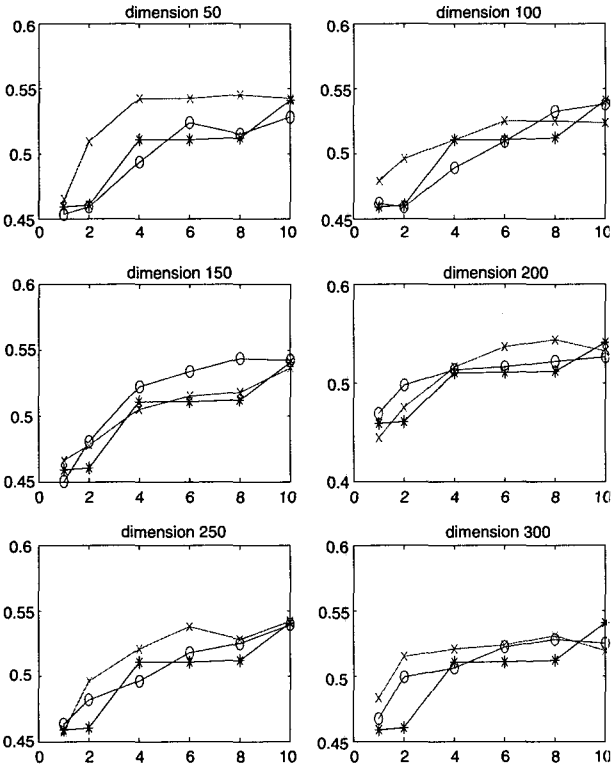
대역어 선택시 잘못 선택하는 경우를 분석해보면 다음과 같이 몇가지 원인을 찾아 볼 수 있다. 첫째, 유사도 추정에 포함된 어휘의 수가 제한되었기 때문이다. 본 논문에서는 약 20,000 어휘만을 위한 유사도 추정 행렬을 구축하였다. 때문에 이 행렬에 포함되지 못한 어휘가 인자어로 입력될 경우에는 의미상 가장 유사한 단어를 사전에서 찾을 수 없기 때문에 디폴트 의미를 적용하는데 여기서 오류가 발생할 수 있다. 예를 들어 'pantaloon', 'minimill'과 같은 단어는 유사 단어 추정이 불가능한데, 전체 오류 중 약 16%가 이 경우에 해당한다. 둘째, 어근을 찾는 알고리즘에서 문제가 발생한 경우이다. 대량의 문서 데이터에서 출현한 어휘를 대표어휘로 변환하기 위해서 본 논문에서는 단순한 형태의 스테밍 툴(stemming tool)을 활용하였다. 예를 들어 본 실험에서는 'house'와 'housing' 모두 동일한 어근인 'hous'로 대표어휘가 결정되는데 두 단어는 각각 'build'의 대역어를 '건축하다'와 '건설하다'로 다르게 선택하게 한다. 마지막으로 LSA, PLSA는 정보 검색에서 주로 활용된 방법론이라는 점이다. 기계 번역과 정보 검색은 '유사'하다는 의미가 서로 다르다. 정보 검색은 주어진 질의어와 동일 문맥에서 자주 공기한 단어들을 위주로 검색을 하기 때문에 '유사'하다는 의미는 '공기'라는 의미를 많이 물려받는다. 반면에 기계 번역에서는 의미 유형 부류가 서로 근접할 경우 '유사'하다고 한다. 예를 들어 기계 번역에서는 '의사'의 유사어로 '간호사' 또는 '변호사'처럼 서로의 의미 유형이 비슷한 단어들이 선택되어야 하지만 정보 검색에서는 '변호사' 보다는 '병원', '주사기' 등이 선택되어야 올바른 검색이 가능하다.

그리고 두 번째 실험은 차원의 복잡도와 생성된 공간의 분포 건전성(distributional soundness) 간의 상관관계가 특정한 성격을 가지고 있는지 확인하는 실험을 하였다. 이 실험을 위하여 본 논문에서는 차원의 복잡도와 대역어 선택의 정확성간의 상관관계를 계산하였으며 동시에 $k-$ 최근점 학습의 k값과 대역어 선택의 정확도간의 상관관계를 계산하였다. 이 실험에서는 50, 100, 150, 200, 250, 300의 6가지 차원수를 사용하였고 동시에 1개, 2%, 4%, 6%, 8%, 그리고 10%의 6가지의 k값을 사용하였다. 여기서 %값은 전체 샘플 중에서의 비율을 의미한다. 첫 번째 실험과는 달리, 정확도를 계산할 때 사전에는 포함되어 있지 않기 때문에 유사도 계산 및 $k-$ 최근점 학습을 적용해야 하는 샘플들만을 그 대상으로 하였다. 따라서 결과의 y축의 값이 실제 대역어 선택 정확도에 비하여 낮은 수치를 보여주고 있다. (그림 1)은 축소된 벡터의 차원수와 선택 정확도가 각각의 k값에 따라서 어떠한 관계를 가지고 있는지를 보여주고 있다. 6개의 그림이 각각의 k값에 따라서 그 경향을 보여주고 있다. -×-로 표시되는 라인은 LSA 은닉 공간을 통하여 얻어진 결과이고 -○-로 표시되는 라인은 PLSA 공간을 통하여 얻어진 결과이다. 그리고 마지막으로 -*- 라인은 차원을 축소하지 않은 경우의 결과이다. X축은 50에서 300



(그림 1) 다양한 k값에 대한 차원수와 대역어 선택 정확도간의 상관관계

까지의 차원수를, Y축은 선택 정확도를 나타내고 있다. 대개의 경우 차원을 축소하는 경우에 그렇지 않은 경우에 비해서 정확도나 속도 면에서 더 좋은 성능을 보여주고 있는 것을 알 수 있다.



(그림 2) 다양한 차원수에 따른 k값과 선택 정확도와의 상관관계

(그림 2)에서는 $k-$ 최근점 학습에서의 k값과 대역어 선택 정확도 간의 관계를 다양한 차원수에 대하여 보여주고 있다. (그림 2)의 6개의 그림은 차원수가 50에서 300까지 다양한 경우의 결과를 보여주고 있으며 각 그림에서의 3개의 라인은 (그림 1)에서의 경우와 동일하다. 그리고 그림에서의 X축은 k값을 Y축은 선택 정확도를 보여준다. 실험 결과 LSA와 PLSA 모두 선택 정확도가 k값이 증가할수록 동시에 같이 증가하는 모습을 보여주었다. 실험의 결과 벡터의 차원의 크기가 대역어 선택의 시간 복잡도에 큰 영향을 미쳤듯이 $k-$ 최근점 학습에서의 k값도 그 정도는 아니지만 역시 시간 복잡도에 영향을 미쳤다.

본 실험에서는 몇 가지 요소들의 상관계수 ($Corr(X, Y)$)를 [22]에서 제시하고 있는 다음의 방법을 통하여 계산할 수 있었다.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

여기서 X는 차원수 또는 k값을 의미하며 Y는 선택 정확도를 의미하고, $\overline{X}$는 X값의 평균을 의미한다.

<표 5>는 차원수와 선택 정확도 간의 상관계수 및 k값과 정확도 간의 상관계수를 보여준다. 표에서 보듯이 k값과 정확도의 상관계수가 차원수와 정확도의 상관계수보다 월등히 높음을 알 수 있다. 다시 말해서, k값의 선택이 벡터의 차원수보다 훨씬 더 선택 정확도에 많은 영향을 미친다고 볼 수 있다. 또한 PLSA에서는 차원수가 150일 경우에 상관계수가 정점을 이루고 있는 것을 보여주는데, 이는 이때가 PLSA 공간이 가장 은닉 의미를 정확하게 표현하고 있는 것을 의미한다. 반대로 LSA는 벡터가 200 차원으로 구축되었을 때 가장 바람직한 결과가 나타났다. 평균적으로 PLSA는 LSA보다 약간 더 높은 상관계수를 보여주었다. 이것은 PLSA를 통하여 생성된 은닉공간의 샘플 분포가 LSA를 통하여 생성된 분포보다 더 최적에 가깝다고 해석할 수 있다. 실제로 150 차원의 PLSA 공간이 200 차원의 LSA 공간보다 더 높은 선택 정확도를 보여주었다.

그리고 k값을 기준으로 보았을 때는 PLSA의 경우 k가 2(전체 예제의 2%)일 때 가장 높은 상관관계를 보여주었다. 이는 k값이 2%일 때 PLSA가 가장 의미 공간을 잘 표현한다고 볼 수 있다. 그러나 LSA의 경우에는 전체적으로 상관관계가 너무 낮아서 의미 있는 검증이 불가능하였다.

〈표 5〉 차원수 및 k값에 따른 선택 정확도와 각 모델들의 상관계수

차 원 수	50	100	150	200	250	300	평 균
PLSA	52.89	58.35	59.83	32.45	49.02	34.84	47.90
LSA	43.54	29.20	44.52	60.09	47.90	19.75	40.83
k 값	1	2	4	6	8	10	평 균
PLSA	7.89	23.91	6.46	0.12	1.79	-2.15	6.34
LSA	0.60	2.15	-5.50	-2.75	-3.11	-5.26	-2.31

5. 결 론

본 논문은 영한 기계번역의 대역어 선택시 발생하는 중의성을 해소하기 위하여 $k-$ 최근접 학습 알고리즘과 두개의 데이터 기반 모델을 사용하는 방법을 제시하였다. 본 논문에서 제시된 모델은 별도의 인간의 지식과 노력이 필요없고 단지 가공되지 않은 텍스트 데이터만을 필요로 한다. LSA와 PLSA는 모두 은닉 의미 공간을 구성하는 데 사용되는데 이 공간에서 단어간 유사도를 추정하게 된다.

본 논문에서 제시된 방법을 통하여 대역어 선택에 있어서 디폴트 의미 선택시보다 약 10%의 성능향상이 가능하게 되었다. 또한 본 논문에서는 은닉 공간의 차원수 및 $k-$ 최근접 학습의 k값과 대역어 선택 정확성간의 상관관계를 찾아서 각 모델의 은닉 공간 표현력을 분석하였다. 결과적으로 PLSA가 LSA보다 선택 정확도 및 은닉 의미의 표현력에 있어서 더 좋은 성능을 보여주었다.

향후 연구로서 이들 모델들이 가지고 있는 성능의 제한을 해결하기 위한 다양한 시도가 필요할 것이다. 예를 들어, 본 논문에서 제시된 방법론들의 결과를 서로 결합하여 앙상블 모델을 구축하거나 이들 모델을 선형적으로 결합하여 제 3의 모델을 구축하는 것이 필요하다. 또한 워드넷과 같이 본 논문에서 사용된 방법론과 이형적(heterogeneous)인 모델들을 효과적으로 결합하는 방법도 연구 대상이 될 것이다. 또한 다양한 의미 커널 알고리즘을 통하여 문제를 해결하고자 하는 것도 좋은 연구 대상이 될 것이다.

참 고 문 헌

[1] I. Dagan and A. Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus," *Computational Linguistics*, 20, pp.563-595, 1994.

[2] N. Kim and Y. Kim, "Determining Target Expression Using Parameterized Collocations from Corpus in Korean-English Machine Translation," *Proceedings of Pacific Rim International Conference on Artificial Intelligence*, 1994.

[3] I. Dagan, L. Lee and F. Fereira, "Similarity-based Models of Word Cooccurrence Probabilities," *Machine Learning*, 34, pp.43-69, 1999.

[4] Y. Kim, B. Zhang and Y. Kim, "Collocation Dictionary Optimization using WordNet and $k-$ nearset Neighbor Learning," *Machine Translation*, 16, pp.89-108, 2001.

[5] T. K. Landauer and S. T. Dumais, "A Solution to Plato's Problem : The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge," *Psychological Review*, 104, 1988.

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 41, pp. 391-407, 1990.

[7] P. Foltz, W. Kintsch and T. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," *Discourse Processes*, 25, pp.285-307, 1998.

[8] Y. Kim, J. Chang and B. Zhang, "Target Word Selection using WordNet and Data-driven Model in Machine Translation," *Lecture Notes in Artificial Intelligence*, 2417, p.607, 2002.

[9] T. K. Landauer, P. W. Foltz and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, 25, pp.259-284, 1998.

[10] T. Hoffmann, "Probabilistic Latent Semantic Analysis," *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence(UAI 1999)*, 1999.

[11] T. Hoffmann, J. Puzicha and M. Jordan, "Unsupervised Learning from Dyadic Data," *Advances in Neural Information Processing Systems*, 11, 1999.

[12] T. Hoffmann, "Probabilistic Latent Semantic Indexing," *Proceedings of the 22th Annual International ACM SIGIR conference on Research and Development in Information Retrieval(SIGIR99)*, pp.50-57, 1999.

[13] E. Voorhees and D. Harman, "Overview of the Seventh Text Retrieval Conference(TREC-7)," *Proceedings. of the Seventh Text REtrieval Conference(TREC-7)*, pp.1-24, 1998.

[14] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE trans. on Information Theory*, 13, pp. 21-27, 1967.

[15] D. Aha, D. Kibler and M. Albert, "Instance-based Learning Algorithms," *Machine Learning*, 6, pp.37-66, 1991.

[16] Y. Gotoh and S. Renals, "Document Space Models using Latent Semantic Analysis," *Proceedings of Eurospeech-97*, pp.1443-1446, 1997.

[17] D. Gildea and T. Hofmann, "Topic Based Language Models using EM," *Proceedings of the 6th European Conference on Speech Communication and Technology*, 1999.

[18] T. Hofmann, J. Puzicha and M. Jordan, "Unsupervised Learning from Dyadic Data," *Advances in Neural Information Processing Systems*, 11, 1999.

[19] M. Berry, T. Do, G. O'Brien, V. Krishna and S. Varadhan, "SVDPACKC : Version 1.0 User's Guide," *University of Tennessee Technical Report*, CS-93-194, 1993.

[20] F. R. K. Chung, "Spectral Graph Theory," *Conference Board of the Mathematical Sciences*, 92, American Mathematical Society, 1997.

[21] *http://www.smartran.co.kr/*.

[22] L. Bain and M. Engelhardt, "Introduction to Probability and Mathematical Statistics," *Thomson Learning*, pp.179-190, 1987.

김 유 섭

e-mail : yskim01@hallym.ac.krx

1992년 서강대학교 전자계산학과(학사)

1994년 서울대학교 컴퓨터공학과(석사)

2000년 서울대학교 컴퓨터공학과(박사)

2000년~2001년 서울대학교 컴퓨터신기술 공동연구소 전문연구요원

2001년 (주)아이시티 연구소장

2001년~2002년 이화여자대학교 과학기술대학원 연구전임강사

2002년~현재 한림대학교 정보통신공학부 조교수

관심분야 : 전산금융, 자연언어처리, 기계번역, 데이터마이닝, 기계학습

장 정 호

e-mail : jhchang@bi.snu.ac.kr

1995년 서울대학교 컴퓨터공학과 학사

1997년 서울대학교 컴퓨터공학과 석사

1997년~현재 서울대학교 컴퓨터공학부 박사과정

관심분야 : 기계학습, 은닉변수모델, 텍스트 마이닝, 생물정보학