잠재의미구조 기반 단어 유사도에 의한 역어 선택

장정호*○ 김유섭** 장병탁*
*서울대학교 컴퓨터공학부 **이화여자대학교 과학기술 대학원
jhchang@bi.snu.ac.kr yskim01@ewha.ac.kr btzhang@bi.snu.ac.kr

Target Word Selection using Word Similarity based on Latent Semantic Structure in English-Korean Machine Translation

Jeong-Ho Chang*○ Yu-Seop Kim** Byoung-Tak Zhang*
*School of Computer Science and Engineering, Seoul National University
**Ewha Institute of Science and Technology, Ewha Woman's University

요 약

본 논문에서는 대량의 말뭉치에서 추출된 잠재의미에 기반하여 단어간 유사도를 측정하고 이를 영한 기계번역에서의 역어 선택에 적용한다. 잠재의미 추출을 위해서는 latent semantic analysis(LSA)와 probabilistic LSA(PLSA)를 이용한다. 주어진 단어의 역어 선택시 기본적으로 연어(collocation) 사전을 검색하고, 미등록 단어의 경우 등재된 단어 중 해당 단어와 유사도가 높은 항목의 정보를 활용하며 이 때 k-최근접 이웃 방법이 이용된다. 단어들간의 유사도 계산은 잠재의미 공간상에서 이루어진다. 실험에서, 연어사전만 이용하였을 경우보다 최고 15%의 성능 향상을 보였으며, PLSA에 기반한 방법이 LSA에 의한 방법보다 역어선택 성능 면에서 약간 더 우수하였다.

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 역어를 선택하기 위한 방법으로서, 연어 정보가 많이 이용되어 왔다. [1]에서는 히브리어-영어 기계 번역 시스템에서 연어 정보를 이용한 통계적 의미 모호성 해소 기법을 제안하였으며, [6]에서는 한영 기계번역 시스템에서의 영어 단어 선택을 위한 연어사전 기반의 방법을 제안하였다. 하지만, 이러한 연어 사전의 수작업 구축에는 많은 노력과 비용이 소요되며, 대규모 말뭉치를 이용하더라도 데이터 희소성 문제가 여전히 존재한다 [2].

본 논문에서는 연어사전을 기본으로, 연어 사전에 등재되지 않은 미지의 단어에 대해서는 유사도가 높은 단어를 선택하고, 선택된 단어에 해당되는 정보를 이용하여 주어진 단어에 대한 역어를 결정하는 접근법을 제시한다. 연어사전의 각 항목은 문법관계에 의해 연결된 두 단어들의 쌍으로 구성되며, 주어-서술어, 목적어-서술어, 형용사 수식어-피수식어 관계를 고려한다. 단어간 유사도는 잠재의미 추출을 위한 모델에 기반하여 측정하며, latent semantic analysis(LSA)와 probabilistic LSA(PLSA)를 이용한다.

LSA는 단어 사용에 따른 문맥정보에 기반하여 잠재의미를 추출하는 방법으로서 정보검색 등의 분야에서 널리 활용되고 있다[8]. 모델의 학습에는 행렬의 최소제곱 근사 방법인 singular value decomposition (SVD)이 이용된다. PLSA는 확률적 mixture decomposition에 기반한 방법으로서 데이터에 대한 생성 모델을 다항분포로 정의하고 이에 대한 유사도값을 최대화함으로써 모델을 학습한다. 모델의 학습은 EM 알고리즘에 의해 수행된다[4]. 실험에서, LSA나 PLSA에 의한 단어 유사도 정보를 활용함으로써 연어사전에 의한 기본적인 역어 선택 방식에 비해 최고 15% 정도의 정확도 향상이 있었으며, 특히 PLSA에 의한 방법이 LSA 보다 약간 더 우수한 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에는 연어사전에 의한 역어선택을 설명하며, 3장에서는 잠재의미 추출에 사용된 두 모델과 분석 결과에 기반한 단어간 유사도 측정에 대해 설명한다. 4장에서는 대규모 말뭉치에 기반한 실험결과를 제시하며, 5장에서는 결론 및 향후 연구 방향을 제시한다.

2. 문법관계에 의한 역어 선택

주어진 단어의 기본적 역어 선택을 위해 이 논문에서는 단어들의 문법관계를 이용하며, 이는 사전의 형태로 저장된다. 단어 w에 대한 사전 항목은 다음과 같다.

$$T(w) = \begin{cases} T_1 \ \ if \ \ Cooc(w, w_1) \\ T_2 \ \ if \ \ Cooc(w, w_2) \\ \cdots \ \ \cdots \\ T_n \ \ \text{otherwise} \end{cases}$$

$Cooc(w, w_i)$는 단어 w와 w_i의 문법적 공기관계를 의미하며, 각 선택 항목은 w가 해당 문법관계 하에서 w_i와 같이 나타나면 w의 역어는 T_i로 선택됨을 의미한다. 표 1은 영어단어 'build'의 술어-목적어 관계에서의 예이다.

이러한 연어기반의 역어 선택 방식이 지닌 문제는 기본적으로 데이터 희소성 문제에 따른, 사전에 등록되지 않은 단어에 대해서는 적당한 역어를 추천하기 위한 과정이 없다는 것이다. 이 문제를 해결하기 위해 미지의 단어에 대해서는 가장 유사한 등록단어의 역어를 선택하는 것이 일반적인데, 단어들간의 유사도를 어떻게 측정할 것인가가 문제가 된다. 본 논문에서는 대량의 말뭉치에 기반한 잠재의미추출 모델을 적용하여 단어들간의 유사도를 측정한다.

역어 (build)	목적어		
건설하다	plant	facility	bridge
건축하다	house	center	housing
제작하다	car	ship	model
설립하다	company	market	empire
구축하다	system	stake	relationship

표 1. 동사 build에 대한 연어 사전 항목의 예

3. 잠재의미 추출 모델

3.1 Latent Semantic Analysis

LSA는 텍스트 구성 요소들간의 의미관계 파악을 자동화하기 위한 대표적인 방법으로서, 정보검색이나 텍스트 응집도 분석 등에 적용되어 왔다. LSA에서는 수작업으로 구축된 사전이나, 지식베이스, 시소러스 없이 단어들의 문서내에서의 공기관계에 기반하여 자동적으로 정보를 구축하며, 대규모 말뭉치로부터 단어나 절(passage)들간의 직/간접적인 관계에 대한 수학적 분석에 기반하고 있다[3][8]. 분석 과정에서는 보통 SVD가 이용된다.

문서의 개수가 m이고 단어 개수가 n인 문서집합을 $n \times m$ 행렬 X로 표현하면, 이는 SVD에 의해 다음과 같이 행렬들의 곱으로 분해될 수 있다.

$$X = U\Sigma V^T, \quad UU^T = V^TV = I \tag{1}$$

위 식에서 U는 $n \times r$ 행렬, V는 $m \times r$ 행렬, Σ는 $r \times r$ 대각행렬로 주어진다. 의미적으로 볼 때, U와 V는 각각 XX^T와 X^TX의 eigenvector이며, Σ의 대각원소들은 XX^T, X^TX의 eigenvalue의 제곱근 값으로서 X의 singular value들이다.

대규모 문서의 경우 보통 r보다는 적은 수의 singular vector가 이용되는데 singular value가 큰 상위 k ($< r$) 개를 선택하며, 이는 식(2)와 같이 최소제곱면에서 행렬 X에 대한 최적 근사화를 제공한다.

$$\tilde{X} = U_k\Sigma_kV_k^T \tag{2}$$

3.2 Probabilistic Latent Semantic Analysis

Probabilisitc LSA (PLSA)는 공기(cooccurence) 데이터 분석을 위한 통계적 기법으로서 언어 모델링, 정보검색, 정보여과 등의 분야에 적용되어 왔다[4]. PLSA 기법에서는 공기 데이터에 대한 aspect 모델에 기반하여 데이터 각 항목에 대해 잠재 변수 $z \in Z = \{z_1, z_2, \cdots z_K\}$를 도입한다. 텍스트 문서의 경우 각 데이터 항목은 단어 w와 문서 d의 쌍 (w, d)로 주어지며, 각 z_i는 하나의 의미자질(semantic topic)을 나타낸다.

PLSA에서는 LSA와는 달리 (w, d)에 대한 확률적 모델링을 제공하는데, 이는 식 (3)과 같이 주어진다.

$$P(d, w) = \sum_z P(z)P(d, w|z) = \sum_z P(z)P(w|z)P(d|z) \tag{3}$$

$P(w|z)$와 $P(d|z)$는 각각 특정 주제에 대한 단어와 문서의 분포를 나타내며, z값이 주어질 때 w와 d는 조건부 독립조건을 만족한다. 이러한 분해는 그 관계가 확률적으로 주어지며, 데이터에 대한 다항 분포에 기반한 로그 유사도 함수를 최대화 함으로써 각 확률값들을 추정한다는 점에서 LSA와는 차이가 있다고 할 수 있다. 데이터에 대한 로그 유사도 함수는 식 (4)와 같다.

$$L = \sum_w \sum_d n(w, d)\log P(w, d) \tag{4}$$

이 함수의 최대화에 기반한 식 (3)의 조건부 확률값 추정을 위해서는 EM 알고리즘이 이용된다[4].

3.3 단어간 유사도 계산

일반적으로 단어들간의 유사도는 행렬 X의 행 벡터들간의 내적으로 정의되는데, LSA에서는 식 (5)에 의해 계산된다.

$$XX^T = (U\Sigma V^T)(V\Sigma^TU^T) = U\Sigma^2U^T \tag{5}$$

결국, 단어 w_i는 $U\Sigma$ 행렬의 행 벡터 $u_i\Sigma$에 의해 표현되며, 단어들간의 유사도는 이 벡터들의 내적에 의해 계산됨을 알 수 있다. 본 논문에서는 실제 계산시 각 벡터를 정규화하며($|u_i\Sigma|^2 = 1$), 이는 정보검색에서 흔히 사용되는 cosine 유사도에 해당된다.

PLSA의 경우도 LSA에서와 비슷하며, 두 단어 w_i, w_j 간의 유사도는 식 (6)과 같이 계산한다.

$$Sim(w_i, w_j) = \sum_{k=1}^{K} P(z_k|w_i)P(z_k|w_j) \tag{6}$$

$$P(z_k|w) = \frac{P(z_k)P(w|z_k)}{\sum_{k'=1}^{K} P(z_{k'})P(w|z_{k'})}$$

4. 실 험

기본적인 언어 사전 구성을 위해서 Wall Street Journal 말뭉치와 기타 신문 기사 문서를 사용하였으며, 총 문장의 수는 261,797 문장이다. 이 중에서 술어-목적어, 주어-술어, 수식어-피수식어 문법 관계에 대해 각각 2,437, 188, 818개의 예제를 추출하였다. 단어들간의 유사도를 측정하기 위해서는 79,919개의 문서로 구성된 1998년 AP 뉴스 말뭉치를 사용하였으며, stemming 알고리즘을 적용하고 불용어와 출현 문서의 수가 20 이하인 단어를 제외하고 난 후 단어의 수는 19,286개이다. 표 2는 LSA와 PLSA 적용 후 식 (5)와 식 (6)에 의해 측정된 유사 단어들의 예이다.

표 3은 세 종류의 문법 관계에 대해 역어 선택의 정확도를 보인다. k-최근접 이웃 방식을 적용하여 역어를 선택하였으며 각 관계에 대해 5-fold 교차검증(cross validation)을 통해 평균 정확도를 계산하였다. 잠재의미 공간 차원은 LSA의 경우 200, PLSA의 경우 128로 하였다. 그리고 PLSA 학습 시 EM 알고리즘의 반복횟수는 최대 50회로 제한하였다. 제시된 결과는 k의 값을 1, 5, 10으로 한 결과 중에서 가장 좋은 성능에 대한 것이다. 잠재의미에 기반하여 계산된 단어 유사도를 이용함으로써 연어사전기

반의 기본 역어 선택보다 11% ~ 15%의 성능향상이 있었으며, 또한 PLSA에 의한 결과가 LSA에 의한 결과보다 약간 높은 정확도를 보였다.

단어	유사한 상위 5 단어				
plant	westinghous	isocyan	shutdown	zinc	manur
	radioact	hanford	irradi	tritium	biodegrad
car	buick	oldsmobil	chevrolet	sedan	corolla
	highwai	volkswage n	sedan	vehicular	vehicle
ship	vessel	sail	seamen	sank	sailor
	destroy	frogmen	maritim	skipper	vessel

표 2. 단어에 대한 상위 5개의 유사어의 예. 각 단어에 대해 첫 줄은 LSA, 두 번째 줄은 PLSA에 의한 결과이다.

문법관계	모델	정확도
주어-술어	LSA	84.41% (k = 1)
(71.85%)	PLSA	**86.05%** (k = 10)
목적어-술어	LSA	84.62% (k = 5)
(75.93%)	PLSA	**87.49%** (k = 5)
수식어-피수식어	LSA	80.93% (k = 10)
(70.54%)	PLSA	**82.76%** (k = 10)

표 3. 역어 선택의 정확도. 첫 번째 열의 괄호 안의 숫자는 연어사전만 이용할 때의 정확도이다.

그림 1은 주어-술어 문법관계에 대해 k의 값에 따른 정확도를 나타낸 것이다.

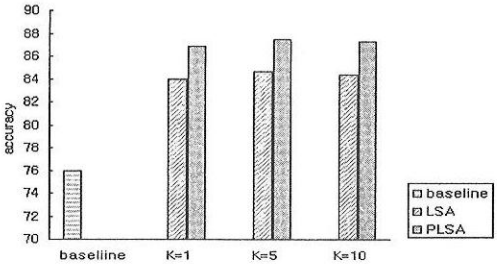

그림 1. 주어-술어 관계에 대한 역어 선택의 정확도

5. 결 론

본 논문에서는 기계번역에서의 단어의 역어선택을 위해 일반적으로 사용되는 연어사전과 단어 유사도 기반의 접근법을 결합하는 방안을 제시하였다. 단어들 사이의 유사도를 계산하기 위해서 잠재의미분석 모델에 기반한 방법을 이용하였으며, LSA와 PLSA 두 가지 모델을 적용하였다. 이 모델들은 문법관계 등에 대한 사전지식을 고려하지 않음에도, 연어사전만 이용하는 경우에 비해 최고 15% 정도의 정확도 향상을 이루었다. 특히 PLSA의 경우가 LSA 경우보다 약간 더 높은 정확도를 보였다. 이 결과는 언어모델링 [4]이나 정보 검색 [5]등에서의 결과와 일치한다.

본 논문에서는 정보검색에서 일반적으로 사용되는 방식에 따라 단순히 벡터로 표현된 문서에 대해 두 모델을 적용하였으며, 보다 의미있는 결과를 위해서는 기계번역이나 자연언어처리 분야에서의 지식을 보다 고려할 필요가 있다. 예를 들어 스테밍 알고리즘은 정보검색에서는 일반적으로 사용되나, 기계번역과 같은 자연언어처리 분야에서는 더 정밀하고 언어학적 지식이 고려된 형태소 분석을 적용하는 것이 보다 일반적이다. 또한 기계 번역에서 역어 선택 과정에서는 4절의 표 2에서 보인 것과는 다른, '인간', '장소', '건축물' 등과 같이 보다 추상화된 수준에서의 자질이 성능 향상에 도움이 될 수 있다. 이를 위해 WordNet과 같은 시소러스에 기반한 유사 단어 선택 방법[7]과 본 논문에서의 말뭉치에 기반한 접근 방법을 결합할 수 있을 것이다.

감사의 글

본 연구는 과학기술부 뇌신경정보학 사업 (BrainTech)과 교육부 BK21-IT 프로그램에 의하여 일부 지원되었음. 또한 이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터신기술공동연구소에 감사 드립니다.

참고문헌

[1] Dagan, I. and Itai, A., "Word Sense Disambiguation using a Second Language Corpus," *Computational Linguistics*, vol. 20, pp. 563-596, 1994.

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

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

[4] Hofmann, T., "Probabilistic Latent Semantic Analysis," *Proceedings of the Fifteenth Conference on Uncertainty iin Artificial Intelligence (UAI'99)*, 1999.

[5] Hofmann, T., "Probabilistic Latent Semantic Indexing," Proceedings of the 22th Annual International Conference on Research and Development in Information Retrieval (SIGIR '99), 1999.

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

[7] Kim, Y., Zhang, B., and Kim, Y., "Collocation Dictionary Optimization using WordNet and k-nearset Neighbor Learning,", *Machine Translation*, 2002 (to appear).

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