한영 질의어 변환을 위한 공통 중간개념 구축

최용석, 서충원, 신사임, 김재호, 최기선
한국과학기술원 전산학과/전문용어언어공학연구센터/첨단정보기술연구센터
{angelove, cwseo, miror, jjaeh, kschoi}@world.kaist.ac.kr

Conceptual Interlingua Construction for Korean-English Query Translation

Yong-Seok Choi, Chung-Won Seo, Saim Shin, Jae-Ho Kim, Key-Sun Choi
Department of Computer Science KAIST/KORTERM/AITrc

요 약

질의어 변환 방법은 다국어 정보검색을 위한 방법중에 효율적인 방법이다. 양질의 질의어 변환을 위해서, 사전, 온톨로지, 병렬 코퍼스 등과 같은 자연언어 자원이 필요하다. 이러한 자연언어 자원은 양질로 대량으로 구축하려면 많은 비용이 든다는 단점이 있다. 본 논문에서는 한영 질의어 변환에 적용할 수 있는 공통 중간개념 구축방법을 제안한다. 공통 중간개념은 동사들의 축으로 이루어지며, 동사들은 기본동사들의 조합으로 표현할 수 있다고 가정한다. 공통 중간개념은 적은 자연언어 자원을 효율적으로 이용할 수 있도록 한다. 본 논문에서는 기본 동사 축을 특이값 분해(singular value decomposition) 방법으로 구하고, 그 기본 동사 축을 이용해서 질의어 변환하는 방법을 보여준다.

1. 머릿글

다국어 정보검색(Multilingual Information Retrieval)의 정의는 서로 다른 언어로 이루어진 정보들로부터 원하는 정보를 검색하는 것을 말한다[맹성현 1999]. 사용자가 언어에 구애받지 않고 여러 언어의 문서를 검색해서 원하는 정보를 얻게 해주는 것이다. 예를 들어 한글로 검색하면 한국어 문서뿐만 아니라 일본어, 중국어, 영어 문서를 모두 사용자에게 제시해 줄 수 있도록 하는 것이다. 교차언어 정보검색(Cross-Language Information Retrieval)은 다국어 정보검색의 일부분으로 자국어를 사용해서 다른 언어로 이루어진 문서를 검색하는 것을 말한다.

다국어 정보검색 시스템을 구성하기 위한 기본 기술로서 질의어 변환 기술이 필요하다. 사용자의 질의를 데이터에 맞는 언어로 바꾸어서 다국어로 이루어진 문서 집합내에서 검색하는 방법이다. 서로 다른 언어들을 다루기 위해서는 의미정보가 필요하다. 기존 방법으로는 사전과 같은 온톨로지를 이용하는 방법, 병렬 글모둠을 이용하는 방법, 은닉의미 색인 방법 등을 사용한다. 이러한 방법들로 만들어진 시스템들의 가장 큰 약점은 신조어에 대한 적응성이 떨어진다는 것이다. 또한, 병렬 글모둠과 같이 확보하기 어려운 자원을 필요로 하거나, 특정 영역에서만 우수한 성능을 발휘하는 등의 약점도 있다.

본 논문에서는 확률벡터를 사용해서 공통 중간개념(conceptual interlingua)을 생성하는 방안을 제안한다. 동사를 축으로 하는 공간에 명사들을 표현해서 언어의 변화에 맞추어 신조어를 표현할 수 있도록 하는 적응성을 갖춘 질의어 변환 모델이다. 명사를 벡터로 표현하기 위해서 단일 글모둠을 사용하며, 서로 다른 언어간의 관계는 수동으로 동사들을 연결시켜 같은 공간으로 만들어 준다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구와 문

제점에 관해서 다룬다. 3장에서는 의미정보인 공통 중간개념 구축방법에 대해서 다루고, 4장에서는 구축한 공통 중간개념을 통한 실험에 대해 다룬다. 마지막으로 5장에서 결론과 향후 과제에 대해서 살펴본다.

2. 관련 연구

정보검색을 위해서 사람이 생각하는 언어구조로 이루어진 의미 구조를 기본으로 생각했었다. 은닉의미 색인(Latent semantic indexing) 방법은 기존 생각에서 벗어나 미지의 의미 구조가 있다고 생각하고 그 감춰진 의미로 색인을 하는 것이다. 따라서, 어떤 언어로 이루어진 문서라도 그것을 은닉의미 형태로 표현할 수 있다면, 같은 좌표공간에 색인해서 같은 방법으로 다룰 수 있다.

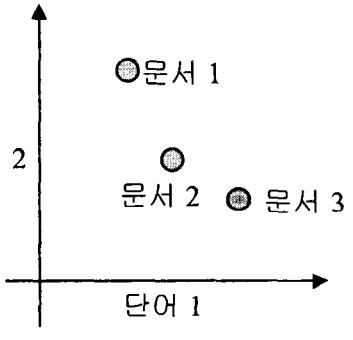

그림 1 기본 벡터 모델

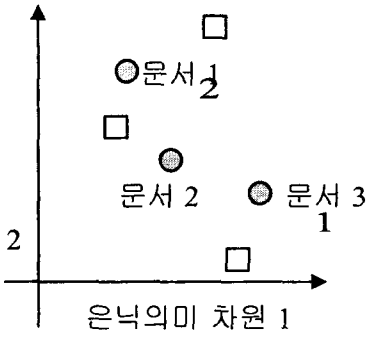

그림 2 은닉의미 색인 모델

그림 1은 일반 벡터공간 모델의 형태를 보여준다. 벡터 공간 모델에서는 단어들이 차원이 되고 문서가 어떤 단어로 이루어졌는지 표현하게 된다. 그림 2는 은닉의미 색인을 이용해서 새로운 차원에서 단어와 문서를 한꺼번에 표현하는 것이다.

문서와 단어를 한 공간에 표현하기 위해서 문서를 색인어와 문서의 행렬로 표현한다[Dumais 1997]. 이 행렬은 특이값 분해(singular value decomposition) 방법을 통해 축소시켜진다. 이 과정에서 동의어나 유사한 문서들이 서로 묶여지면서 줄어든 행렬의 크기로 벡터공간을 설정해서 그 벡터공간에서 유사도를 비교하게 된다. 이 벡터공간에 다른 언어로 이루어진 문서도 같이 표현할 수 있다.

이 방법을 사용하여 벡터공간만 잘 만들어 준다면 어려운 번역과정을 거치지 않고도 사용자가 원하는 문서를 검색할 수 있다. 하지만, 이 방법을 사용하기 위해서 학습과정에 병렬 글모둠(parallel corpus)이 많이 필요한데, 대량의 병렬 글모둠을 확보하기는 어렵다. 또한, 이 방식은 계산하는 비용이 굉장히 많이 들어가며 특정 영역에서는 잘 적용할 수 있으나 일반적인 영역에서 적용하기에는 적합치 않다.

또 다른 다국어 정보검색 방법은 언어변환으로 질의어 변환을 하는 방법이다. 사전과 병렬 글모둠을 이용하는 혼합방법[Choi 2000]은 사전의 성능에 의존적이고, 은닉의미 색인과 마찬가지로 대량 구축이 어려운 병렬 글모둠을 사용하는 단점이 있다. 다국어 온톨로지를 이용하는 유로 워드넷[Gilarranz 1997]은 독어, 이태리어, 스페인어, 영어 등 유럽 4개 언어의 단어들에 대해 서로간의 의미적 관계를 연결하는 중간언어 색인(interlingual index)을 만들어 개념기반의 문서검색을 행한다. 다국어 온톨로지를 사용함으로써 좋은 효과를 거둘 수 있으나, 다국어 온톨로지를 수동으로 구축하는데 오랜 기간이 소요된다..

3. 공통 중간개념 구축

공통 중간개념이란 기계번역에서 모든 언어를 표현할 수 있는 중간상태인 언어(interlingua)가 있다고 가정하는 것과 같이 모든 개념을 표현할 수 있는 방안을 뜻하는 이상적인 다국어 온톨로지를 말한다.

본 논문에서는 공통 중간개념을 구성하기 위해서 명사들을 동사 축위에 표현한다. 기존 은닉의미 색인 방법에서는 알 수 없는 의미축에 단어와 문서들을 배열했으나, 제안하는 방법은 사람이 이해할 수 있는 동사를 기본 축으로 삼아서, 신조어가 많이 발생하는 명사들을 의미 공간에 벡터로 표현하려고 한다. 동시에 같은 공간에 표현된다고 생각하는 영어 동사들을 축으로 설정해서 영어 명사들도 벡터로 표현한다. 동사는 기본적으로 신조어가 적기 때문에 공간의 기준인 축으로 적합한 성질을 가지고 있다.

명사를 동사로 표현하겠다는 것은 명사의 속성은 기본적으로 동사로 표현할 수 있다는 것을 가정한다. 일반적으로 어떤 명사를 상대방에게 알아내도록 하는 방법 중에 하나로

"스무 고개"와 같은 방법을 사용한다. 즉, "먹는 것", "입는 것" 이러한 정보가 명사의 속성을 표현하며, 사람은 이러한 명사의 속성을 통하여 명사의미를 파악할 수 있다.

3.1. 확률 벡터 모델

동사를 축으로 명사를 표현할 때는 동사와 명사의 공기정보를 이용하여 명사의 벡터를 구성한다. 개념은 확률 벡터모델[Wong 1987]을 기본으로 한다. 이 벡터간의 비교를 위해서 교차 엔트로피(Cross Entropy)를 사용한다. 확률 벡터(Probability Vector)는 다음과 같이 정의된다.

n차 벡터 $\vec{P}=(p_1, p_2, \cdots, p_n)$가 다음 식을 만족하면 n차 확률벡터이다.

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0, \quad i = 1, 2, \cdots, n$$

분야 하나를 확률벡터 $\vec{P}$로 본다면, 각 p_i는 i번째 점수 단어의 분야에서 나타난 빈도를 확률로 나타낸다.

확률벡터 $\vec{P}$에 대한 엔트로피를 다음과 같이 정의한다.

$$H(\vec{P}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

확률벡터 $\vec{P}$의 각 요소(element) p_i들의 불확실성(uncertainty)을 $-\log_2 p_i$의 값으로 측정할 수 있다. 그러므로, 엔트로피는 확률벡터 $\vec{P}$의 정보 불확실성(information uncertainty)에 대한 기대값이다. $H(\vec{P})$는 모든 요소들의 확률이 같을 때 최대값을 가지며, 한 요소만이 1이고 나머지는 0일 때 $H(\vec{P})$는 최소값 0을 가진다.

확률벡터 $\vec{P_1}$와 $\vec{P_2}$가 같은 차원의 벡터일 때, $\lambda \in [0,1]$에 대해서 벡터 $\vec{P}=\lambda\vec{P_1}+(1-\lambda)\vec{P_2}$도 역시 확률벡터이다. 이 때, 확률벡터 $\vec{P}$를 $\vec{P_1}$와 $\vec{P_2}$의 복합 확률벡터(Composite Probability Vector)라고 한다. 그리고, $\vec{P_1}$와 $\vec{P_2}$를 $\vec{P}$의 구성 확률벡터(Components Probability Vector)라고 한다.

확률벡터 $\vec{P}=(p_1, p_2, \cdots, p_n)$와 $\vec{Q}=(q_1, q_2, \cdots, q_n)$가 주어졌을 때, 복합 확률벡터 $\frac{1}{2}\vec{P}+\frac{1}{2}\vec{Q}$와 이 벡터의 구성 확률벡터들 사이의 엔트로피의 차이를 교차 엔트로피(Cross Entropy)라고 하며 다음과 같이 정의한다.

$$\beta(\vec{P}, \vec{Q}) = H(\frac{1}{2}\vec{P}+\frac{1}{2}\vec{Q}) - \frac{1}{2}[H(\vec{P})+H(\vec{Q})]$$

이 때, β는 다음의 부등식을 항상 만족한다.

$$0 \leq \beta(\vec{P}, \vec{Q}) \leq 1$$

β의 값은 두 개의 확률벡터가 복합될 때, 불확실성의 증가 정도를 나타내고 있다. 만약 두 확률벡터가 관련되어 있으면 각각의 확률벡터 요소들의 확률 분포(probability distribution)가 유사하며 두 확률벡터가 관련이 많을수록 β의 값은 작아진다. 즉, 불확실성의 증가 정도가 적어진다. 이러한 β의 값은 두 확률벡터의 관계 차이를 나타낸다. 그러므로, β를 비관련도 계수(dissimilarity coefficient)로 해석할

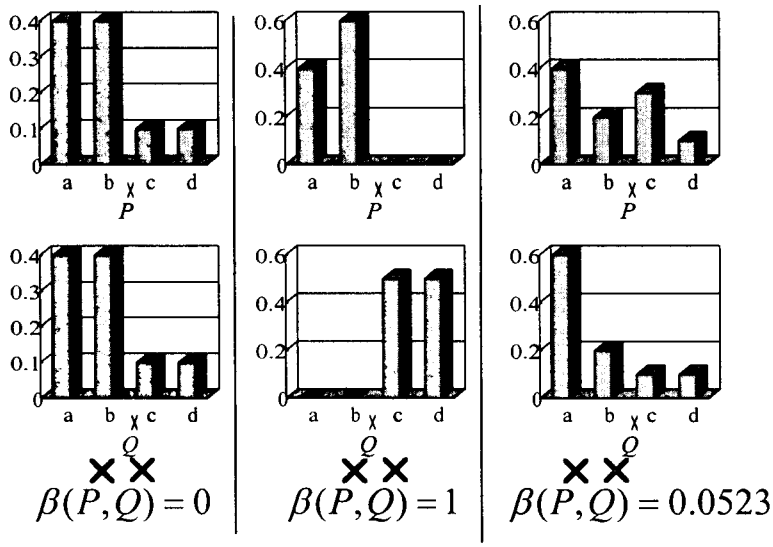

그림 3 확률벡터간의 의미거리 예

수 있다. β를 단어간의 의미거리(concept distance) 값으로 사용한다.

그림 3은 확률벡터 모델에서 두 확률벡터간의 의미거리에 관한 예를 보인 것이다. 첫 번째 경우는 완전히 벡터가 일치할 의미 거리는 0이 된다. 두 번째 경우는 완전히 상이한 벡터인 경우로 이 때 의미거리는 1이 나온다. 세 번째는 비슷한 벡터를 비교하면 의미거리 값이 0에 가까이 나옴을 알 수 있다. 따라서, 확률벡터로 표현한 각 명사들간의 의미가 얼마나 가까운지 의미거리 값으로 측정할 수 있다.

3.2. 기본동사 정의

동사를 기본축으로 설정하기 위해서는 기본동사 정의가 필요하다. 모든 동사들을 축으로 사용하면 보다 명확한 명사의 의미를 표현할 수도 있겠으나, 벡터의 크기가 너무 커지고, 자료 희귀문제(data sparseness)에 따른 부작용이 발생할 수 있다. 명사를 동사로 표현하기 위해서는 꼭 필요한 동사들을 파악해야 하며, 이를 기본동사라 한다. 술어들은 상태, 활동, 목적을 가진 사건의 세부류로 나누어지는데, 도우티(Dowty)는 이러한 모든 술어들은 상태를 나타내는 술어들과 3개의 양상 연산자 DO, BECOME, CAUSE들만 사용해서 의미속성을 표현할 수 있다고 했다[Chierchia 1993]. 따라서, 명확한 기본동사만의 축으로 모든 술어의 속성을 표현할 수 있다.

기본동사를 선정하기 위해서 형태소 분석기[이운재 1999]의 동사, 형용사 사전을 이용했다. 이 사전은 동사 3468개 형용사 1626개로 5097 용언류 단어가 들어있다. 형태소 분석기 사전에 들어있는 용언류가 기본적으로 필요한 단어라 할 수 있다. "하다"류의 동사를 제외하기 때문에 사전에 나오는 단어수와 많은 차이가 난다. 이 단어들에 수동으로 중요도 점수를 부여한 결과를 사용했다[최용석 2000]. 실제 벡터계산에 사용하기 위해서 기본동사만으로 벡터를 구성하기 어려웠다. 글모둠에 나타나는 빈도순으로 212개의 동사를 기본동사에 추가하였다. 추가로 인하여 교집합을 제외한 기본동사는 품사구분을 통하여 구분할 수 있는 동사까지 다른 동사로 구분하여 총 500개가 구성되었다.

선정한 한국어 기본동사 500개에 대하여 각각 가장 적합하다고 여겨지는 영어 동사를 수동으로 부착하였다. 이 때, 언어의 차이로 인하여 적합한 동사를 부여할 수 없는 경우 기본동사에서 제외시켰다. 품사별로 구분이 되었으나 형태적으로 같은 동사는 모두 하나로 다룬 후에, 영어와 1:1 대응이 가능한 한국어 기본동사는 총 455개가 선정되어 한국어 명사와 영어 명사를 455차원 벡터로 같은 공간에 표현할 수 있었다. 온톨로지로 사용하기 위해서 같은 공간에서 한국어 명사와 영어 명사간의 의미거리 계산을 하여, 질의어에 출현한 한국어 명사를 의미거리가 가까운 영어 명사로 대체할 수 있다.

이렇게 455개의 동사를 사용해서 명사를 표현했다면, 동사는 서로 독립적이라는 기본 가정을 하고 명사들간의 의미거리를 계산한 것이다. 그러나, 동사들은 서로 독립적이지 않고 서로간에 관련있는 의미요소들의 조합으로 이루어졌다. 예를 들어 다음과 같은 두 문장이 있을 경우 "팔다"라는 동사의 의미에는 "주다"라는 기본 의미요소는 포함되어 있고, 세부적 의미요소인 "돈같은 대가를 받았다."라는 의미까지 포함된 것이다[Genter 1982].

"영희가 철수에게 시계를 주었다."

"영희가 철수에게 시계를 팔았다."

"팔다"와 "주다"는 서로 독립적인 동사가 아닌 것이다. 이럴 경우 두 동사 모두를 축으로 할 필요성이 있는가를 조사해야 한다. 축들이 서로 독립적이지 못하다면 정보의 중복뿐만 아니라, 명사간의 유사도를 계산할 때에도 서로 의존적인 정보들이 유사도 계산에 부정적인 영향을 끼친다. 본 논문에서는 이 단점을 극복하기 위해 기저벡터를 구하는 그램 슈미트 직교화(gram schmidt orthogonalization)방법을 사용하여 동사축들간의 독립성을 확보하는 방법을 제안한다. 실질적으로 컴퓨터에서 구현은 특이값 분해(singular value decomposition) 방법을 사용하여 기저벡터를 구하는 형식으로 이루어 진다.

기저벡터를 구한다면 서로 관련있는 의미요소들은 제거되고, 독립적인 의미요소들만 축으로 남게 된다. 기저벡터에 필요한 동사가 독립적 의미요소에 필요한 기본동사라는 것을 알 수 있다. 기저벡터를 구하기 위해서 기본적으로 동사와 명사의 공기정보를 이용하여 구성한 공간을 전체공간으로 가정하고 계산했다. 기저벡터를 구하는 과정에서 사라지는 동사가 있다면, 그 동사의 의미요소들은 모두 다른 동사들에게 포함되어 있다는 것을 알 수 있다. 기본동사에 대한 공기정보를 통하여 공통 중간개념을 구성하고, 공통 중간개념을 통하여 질의어 변환을 실험해서 결과를 관찰한다.

4. 실험

한국어 공기정보는 97년판 과기원 글모둠 중 구문표지 부착 글모둠을 이용하여 구해졌다. 이 글모둠을 12만 어절 수준의 글모둠이다. 영어 공기정보는 동사와 목적어관계를 가지는 Penn TREEBANK II 1994년도판을 이용하여 추출했다.

한국어 명사 6147개에 대해서, 영어 글모둠에서 많이 출현한 명사 1000개와의 의미거리를 구하여 가장 가까운 영어 명사를 추출하였다.

한국어 명사 중 많은 부분이 기본 동사로 설정한 455개의 동사와 함께 출현하지 않아서 의미거리를 구할 수 없었으며, 정보의 부족과 한국어, 영어의 글모둠의 분야 차이로 인해 올바르지 않은 추출결과도 많았다. 하지만, 단지 455개 동사를 설정하고 그 공기정보만으로 의미거리를 구한 결과로 "무엇"이 "anything"에 대응되기도 하고, "간호, 확장, 수습, 결심" 등 "하다"를 붙일 수 있는 많은 수의 명사들이 "work"와 대응되었다.

프로그램 오류를 제거하면서 "가격"이 "price"랑 대응되었다. 전체적으로 영어 공기정보는 경제에 관련된 문서에서 추출하였기 때문에, 대응결과가 좋지 않았다.

455개 동사 중, 한국어 공기정보에 출현한 359개의 동사에 대하여 기저벡터를 구한 결과 358차원으로 한 차원 줄일 수 있었다. 그리고, 축의 가중치로 나온 값이 0.000001인 경우는 5 경우였고, 0.000002인 경우는 2 경우 발생했다. 이런 가중치 값에 임계치를 두어 8차원의 축을 사라지게 할 수 있다 사실을 알 수 있다. 이는 단순 동사로 이루어진 축들은 서로 의존적인 측면이 있다는 것을 알 수 있게 해 준다. 줄어드는 축을 관찰해 보면 (나다: -0.03652, 나누다: -0.11569, 다루다: -0.120764, 매다: 0.036515, 미치다: -0.060631, 믿다: -0.144294, 싸다: -0.036515, 쓰다: -0.314025, 알리다: 0.49459, 얹다: 0.036515, 일으키다: 0.084494, 잃다: 0.060631, 짓다: 0.115686)의 의미를 가지고 있는 축이다.

확률벡터를 이용한 계산은 벡터의 요소로 음수값이 나오면 적용 불가능하다. 따라서, 내적값을 계산해서 두 벡터 사이의 코사인 값을 유사도값으로 이용하는 실험이 필요했다. 이 실험에서는 병렬 글모둠을 사용하는데 우루과이 라운드 협정문을 사용했다. 우루과이 라운드 협정문의 특성은 표1과 같다.

표 1 우루과이 라운드 협정문 특성

항목	영어	한국어
구역	4968	4968
어절	13만 9265	7만 9290
구역당 평균 어절	28.03	15.96
개념어	6만 5844	6만 5653
유일한 개념어	2681	3847

실험 결과로는 455개 동사 중, 한국어에 대해서는 130개의 동사만이 출현하며 기저벡터를 구하는 방법으로 한 차원도 줄어들지 않았다. 다만, 축의 가중치 중 최소는 0.000057로 이 축을 제거해서 한 차원 줄일 수 있었다.

이에 영어 공기정보의 동사부분을 한국어 동사로 대치한 후에 한국어 공기정보와 같이 기저벡터를 구하는 실험을 했다. 이 때는 208개의 동사가 출현했다. 208개의 축에 가중치별로 계산을 달리했다. 각각의 정확도에 대한 결과는 아래 표와 같다.

표 2 가중치에 따른 축의 변화와 정확도

가중치(이하)	0.000000	0.000001	0.000010	0.000050
차원	208	207	205	200
정확도(%)	10	10	11	9

가중치 0이하인 축은 없었으므로, 가중치 0이하인 축을 제거시키고 한 계산은 원래 아무 가공없이 한 계산과 같다. 정확도는 100개의 단어를 임의로 추출한 후에 그 100개의 단어에 대해서 대역어가 올바로 선택되었는지 사람이 판단하였다. 대부분의 대역어가 만족할 만한 값으로 변환되지는 않았다. 하지만, 차원을 축소시켜도 정확도의 변화가 거의 없음을 알 수 있었다. 차원 변화에 대해서 대역어가 변하기도 하였으나 틀린 대역어에서 다른 틀린 대역어로 변하므로 해서 정확도 값에 변화를 주지는 못했다.

5. 맺음글

본 논문에서는 한국어-영어 질의어 변환을 위한 적응성을 가지는 공통 중간개념 구축방법을 제안하였다. 벡터 모델의 축으로 사용할 기본 동사를 결정하는 과정, 기본 한국어 동사와 영어 동사의 수동 관계설정 과정, 기본 동사의 의미 중첩을 해소하기 위한 기저 벡터 추출 과정 등을 제시했다. 또한, 실험을 통하여 벡터의 차원을 의미있게 축소할 수 있음을 보여주었다. 차원 축소 이후에도 유사도 값에 큰 변화가 없음을 알 수 있었다.

제안하는 방법을 사용하여 공통 중간개념을 구축해서 다국어 정보검색에 온톨로지로써 사용할 수 있다. 신조어에 대해서 빠르게 자동으로 대응할 수 있으며, 대량의 병렬 글모둠을 확보하지 않고, 대량의 단일 글모둠만으로 공통 중간개념의 정보의 양과 질을 향상시킬 수 있을 것으로 기대한다.

향후 의미 중첩 해소를 통한 기본 동사 확보 과정에 대해 보다 상세한 연구가 필요하며, 실제 다국어 정보검색 시스템에 적용하는 일도 향후연구로 필요하다. 의미구분과 격틀 사용을 통한 표현의 명확화 연구도 성능향상에 도움을 줄 것이다. 또한 벡터모델을 확장시켜 요소 동사들을 표현할 수 있는 방법을 연구할 수 있다.

감사의 글

문서 정렬에 도움을 준 이주호에게 감사의 마음을 표한다.

참고문헌

[맹성현 1999] 맹성현, "Cross-Language Information Retrieval", 정보과학회 한국어 정보처리 연구회 '99 자연언어 처리 튜토리얼, 53-85쪽, 서울 고려대, 1999년 8월 27일

[이운재 1999] 이운재, 김선배, 김길연, 최기선, "모듈화된 형태소 분석기의 구현", 한글 및 한국어 정보처리 학술대회-형태소 분석기 및 품사태거 평가 워크숍, 123-136쪽, 전주, 1999년 10월 8-9일

[채영숙 1999] 채영숙, "구문분석을 전제로 한 전자사전 구축", 한국과학기술원 전문용어언어공학연구센터, 센터내부 메모, 1999년

[한글학회 1991] 한글학회, "우리말 큰 사전", 어문각, 1991년

[최용석 2000] 최용석, 이운재, 최기선, "말모둠에서 동사분포 연구", 한글 및 한국어 정보처리 학술대회, 169-175쪽, 전주, 2000년, 10월 13-14일

[Chierchia 1993] Gennaro Chierchia and Sally McConnell-Genet, "Meaning and Grammar: An Introduction to Semantics", MIT Press, pp 350-360, 1993

[Choi 2000] Yong-Seok Choi, Junghoon Chun, Key-Sun Choi, "A Study on Dynamic Threshold for Korean English Query Translation", The 3rd International Conference of Asian Digital Library, Seoul Korea, December 6-8, 2000.

[Dumais 1997] S.T. Dumais, T.A. Letche, M.L. Littman, and Landauer T.K., "Automatic cross-language retrieval using latent semantic indexing", 1997 AAAI Symposium on Cross-Language Text and Speech Retrieval, American Association for Artificial Intelligence, March 1997.

[Genter 1981] Dedre Genter, "Verb Semantic Structures in Memory for Sentences: Evidence for Componential Representation", Cognitive Psychology 13, pp. 56-83, 1981.

[Gilarranz 1997] Julio Gilarranz, Julio Gonzalo and Felisa Verdejo, "An Approach to Conceptual Text Retrieval Using the EuroWordNet Multilingual Semantic Database", In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, March 1997.

[Wong 1987] S.K.M. Wong and Y.Y. Yao, "A Statistical Similarity Measure", Proceeding of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3-12, 1987.