한국어와 영어의 명사구 기계 번역

조희영†, 서형원†, 김재훈†, 양성일‡
†한국해양대학교 컴퓨터공학과
‡한국전자통신연구원
{serensis,hide90,jhoon}@bada.hhu.ac.kr†, siyang@etri.re.kr‡

Korea-English Noun Phrase Machine Translation

Hee-Young Cho†, Hyung-Won Seo†, Jae-Hoon Kim† and Sung-Il Yang‡
†Department of Computer Engineering, Korea Maritime University
‡Electronics and Telecommunication Research Institute
{serensis,hide90,jhoon}@bada.hhu.ac.kr†, siyang@etri.re.kr‡

요 약

이 논문에서 통계기반의 정렬기법을 이용한 한영/영한 양방향 명사구 기계번역 시스템을 설계하고 구현한다. 정렬기법을 이용한 기계번역 시스템을 구축하기 위해서는 많은 양의 병렬 말뭉치(corpus)가 필요하다. 이 논문에서는 병렬 말뭉치를 구축하기 위해서 웹으로부터 한영 대역쌍을 수집하였으며 수집된 병렬 말뭉치와 단어 정렬 도구인 GIZA++ 그리고 번역기(decoder)인 PARAOH(Koehn,2004), RAMSES(Patry et al., 2002), MARIE(Crego et al.,2005)를 사용하여 한영/영한 양방향 명사구 번역 시스템을 구현하였다. 약 4만 개의 명사구 병렬 말뭉치를 학습 말뭉치와 평가 말뭉치로 분리하여 구현된 시스템을 평가하였다. 그 결과 한영/영한 모두 약 37% BLEU를 보였으나, 영한 번역의 성공도가 좀더 높았다. 앞으로 좀더 많은 양의 병렬 말뭉치를 구축하여 시스템의 성능을 향상시켜야 할 것이며, 지속적으로 병렬 말뭉치를 구축할 수 있는 텍스트 마이닝 기법이 개발되어야 할 것이다. 무엇보다도 한국어 특성에 적합한 단어 정렬 모델이 연구되어야 할 것이다. 또한 개발된 시스템을 다국어 정보검색 시스템에 직접 적용해서 그 효용성을 평가해보아야 할 것이다

1. 서 론

1960년대 이후 많은 연구자들이 자동번역시스템을 개발하기 위한 연구를 진행하였으나(Hutchins and Somers, 1992), 아직 사용자들은 만족할 만한 자동번역 시스템은 거의 없다고 해도 과언이 아니다. 미국 국립 표준 기술원(NIST)[1]의 보고에 따르면 가장 좋은 성능을 가진 자동번역시스템의 성능이 0.5137 BLEU[2]이다(NIST, 2005). 물론 이 측도가 인간의 이해도와 완전히 일치한다고는 생각되지 않는다. 다만 이 결과로 볼 때, 약 반세기 동안 꾸준히 연구되고 있지만, 아직도 번역의 질은 그다지 높지 않다는 것이다. 그러나, 자동번역시스템은 완전한 번역보다는 인간의 번역 작업을 도와주거나 외국문서를 이해하는 데는 많은 도움을 주고 있다. 더구나 최근 통신기술의 급속한 성장으로 다양한 언어로 의사를 전달할 필요성이 점점 더 늘어나고 있으며, 자동번역에 대한 수요도 급증하고 있다. 자동번역에 대한 응용 분야를 살펴보면 매우 다양하다. 예를 들면, 단순한 문서번역, 번역업체의 초벌 번역, 웹 문서 번역, 기술 문서 번역, 전자우편 번역, 방송자막 번역, (휴대폰/PDA) 자동 통역,

1) NIST: National Institute of Standards and Technology

2) BLEU : BiLingual Evaluation Understudy의 약자로서 번역의 질을 자동으로 측정하기 위한 하나의 측도이다(Papineni et al., 2001). 이 측도는 N-gram의 공기빈도를 이용하여 번역의 질을 측정한다.

다국어 정보검색 등이 있다.

많은 인터넷 사용자들은 검색엔진을 통해서 필요한 정보를 찾고 있으며, 유용한 정보들은 다양한 언어로 기술되어 있다. 모국어가 아닌 다른 언어로 기술된 문서를 찾기 위해서 주로 다국어 정보검색 시스템을 사용한다(장명길 외 1998). 다국어 정보검색 시스템의 필수적인 요소 기술은 질의어 번역이며, 대부분의 질의어는 명사로 구성된다. 따라서 명사구번역이 질의어 번역의 핵심 기술이 된다. 일반적으로 명사구번역은 개별 명사의 번역을 이용하여 쉽게 번역될 수 있을 것으로 생각되나 많은 명사들은 영역과 주변의 단어에 따라 서로 다르게 번역된다. 이와 같은 문제를 효과적으로 다루기 위해서 이 논문에서는 통계기반 기계번역 방법을 이용해서 한영/영한 명사구번역을 설계하고 구현하였다.

이 논문의 구성은 다음과 같다. 2장에서 통계기반 기계번역의 요소 기술들에 대해서 간략히 소개하고, 3장에서 한영/영한 명사구번역 시스템을 기술한다. 4장에서 제안된 한영/영한 명사구번역 시스템의 성능을 평가하고 분석한다. 마지막으로 5장에서 결론을 맺고 향후 연구 과제를 기술한다.

2 관련 연구

2.1 통계기반 기계번역

통계기반 기계번역은 기계번역이 처음 시작된 50년대부터 시도되었으나, 컴퓨터 기술과 성능의 한계로 크게 성공하지 못했다. 그러나 최근에 와서 컴퓨터의 기술이 발전되고 또 대량의 병렬 말뭉치들이 쉽게 구축할 수 있게 되면서 다시 활발한 연구가 시작되었다. 통계기반 기계번역의 기본적인 개념은 Shannon의 정보이론에 기반을 두고 있으며, 정렬기법을 이용하여 기계번역을 모델링하였다. (식 1)은 정렬기법을 이용한 통계기반 기계번역 모델이다.

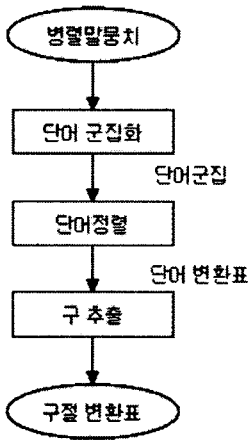
$$\hat{e} = \arg\max_{e} P(e|f) = \arg\max_{e} P(e) \times P(f|e) \quad \text{(식 1)}$$

(식 1)에서 $P(e)$를 언어모델(language model)이라 하고, $P(f|e)$를 번역모델(translation model)이라고 한다. $\arg\max_e$는 통계 기반 기계번역 시스템으로 일반적으로 번역기라고 한다.

통계기반 기계번역을 위한 많은 모델들이 제안되었다(Brown et al, 1990; Manning & Schütze 1999). 그러나 실용화나 상용화를 위해서는 좀더 많은 연구가 필요실정이다. 현재 국내의 경우에는 통계기반 기계번역에 대한 연구가 활발하게 이루어지지 않았다. (신중호, 1996)에서 (Brown et al., 1993)을 근간으로 한영 단어정렬 모델을 제안하였고, (Huang & Choi, 2000)와 (리금희 외, 2002)에서 한중 단어정렬 모델을 제안하였다.

2.2. 단어정렬

단어정렬은 병렬 말뭉치로부터 단어 단위의 대역을 찾는 것을 말한다. IBM Model 1-5(Brown et al. 1990)를 시작으로 많은 연구들이 진행 되었으며 대체로 자율학습을 통한 단어 정렬 모델을 학습한다. 단어의 정렬은 1:1이 아닐 뿐만 아니라 단어간 위치가 뒤집혀 정렬이 될 수 있어 모델 자체가 복잡하다. 그래서 특별한 언어들 사이에는 사전 지식을 이용하여 통계기반 기계번역 모델을 개선하였다. 단어 정렬은 이 자체로 많은 응용 분야를 가지고 있는데, 대역 사전의 자동 생성(Smadja et al., 1996), 의미모호성해소(Diab, 2000) 등이 있다. 단어 정렬의 도구로는 GIZA++ (Och and Ney, 2000), k-vec(Fung and Church, 1994), PWA (Ahrenberg, et al, 2000) 등이 있으며, 이 논문에서는 GIZA++를 사용하였다. 아래에서 GIZA++에 대해서 간략히 설명할 것이다.



(그림 1) GIZA++의 실행과정

GIZA++ 는 GIZA(SMT 툴인 EGYPT의 일부분)의 확장판이다. GIZA는 미국 Johns-Hopkins 대학의 1999년 Center for Language and Speech Processing의 여름 워크샵에 통계기반 기계번역 팀에 의해 개발된 병렬 말뭉치로부터 단어정렬 모델을 학습하는 프로그램이다. GIAZ++은 Franz Josef Och에 의해 구현되었으며, IBM 모델 4, 5와 HMM 모델이 추가되었으며, 단어 군집화 기

법을 이용해서 학습의 속도가 크게 개선되었다. (그림 1)에서 GIZA++의 실행 과정을 보여주고 있다. GIZA++은 mkcls라는 도구를 이용해서 각 언어에 속한 단어를 군집화하고, 그 결과를 이용해서 단어를 정렬한다. 정렬된 단어의 결과는 단어 변환표에 저장되며, 이 변환표에는 단어의 변환 확률도 포함되어 있다. 번역의 정확성을 높이기 위해서 단어 변환표를 이용하여 구절(phrase)를 추출하여 구절 변환표에 저장되며 이 변환표에도 구절의 변환 확률이 포함되어 있다.

2.3 언어 모델

언어 모델은 단어들이 어떤 방법으로 결합하여 문장을 구성할지를 기술하는 방법을 말하며, 일반적으로 언어 모델이라는 말은 통계 언어모델을 의미한다. 이 논문에서는 n-gram 언어 모델을 사용하며, SRILM[3](Stolke et al., 2002)을 이용해 언어 모델을 학습하여 구축하였다. 아래에서 n-gram과 SRILM에 관하여 간단히 살펴볼 것이다.

N-Gram 언어 모델은 가장 널리 사용되는 모델이며 문장의 확률을 가장 일반적으로 표현하는 모델로서 (식 2)와 같이 표현된다.

$$Pr(\omega_1^n) = Pr(\omega_1)Pr(\omega_2 \mid \omega_1)Pr(\omega_3 \mid \omega_1\omega_2) \cdots (Pr(\omega_n \mid \omega_1^{n-1})) = \prod_{i=1}^{n} Pr(\omega_i \mid \omega_1^{i-1}) \quad (식\ 2)$$

확률 $Pr(\omega_1^n)$은 각 단어 ω_1^n에 대한 조건 확률 $Pr(\omega_i \mid \omega_1^{i-1})$로 정의될 수 있다. N-gram 은 여러 가지 방법으로 개선되었으며 지금도 꾸준히 연구 되고 있다.

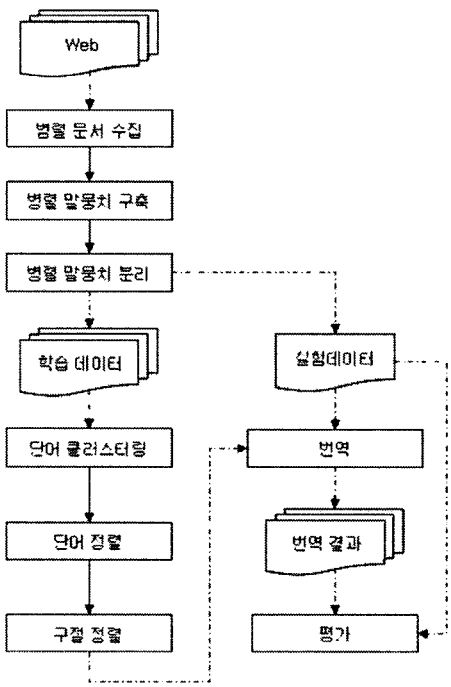
SRILM은 본래는 C++로 만들어진 음성 인식과 통계적 태깅과 분할에서 사용하기 위해 만들어진 통계기반 언어모델을 추정하기 위한 도구이다. SRILM은 1995년 SRI Speech Technology and Research Laboratory에서 만들어 졌다. 90년 최초 등장 당시에는 단순 n-gram만이 되었으나 지금은 다양한 n-gram 모델들도 추정이 가능하게 되었다. SRILM은 언어모델을 학습할 뿐 아니라 문서를 분할하는 등의 다양한 기능도 포함되어 있다.

2.4 번역기

번역기는 학습된 언어모델과 단어 정렬 말뭉치로부터 추정된 단어 정렬 모델을 이용해서 주어진 원시 문장을 목적 문장으로 번역한다. 통계기반 번역기는 탐색문제로서 주로 A* 알고리즘이나 탐욕알고리즘을 이용해 구현한다(Germann, 2003). 또 다른 방법인 A* 알고리즘의 일종인 빔 탐색법은 계산된 확률값이 허용된 범위내에 있을 경우에만 탐색을 한다. 그리고 또 다른 방법으로는 FST를 이용하는 방법과 동적프로그래밍을 이용하는 방법들이 있다. 이 논문에서는 번역기로 PARAOH(Koehn,2004), RAMSES(Patry et al., 2002), MARIE(Crego, et al., 2005) 세 가지를 사용한다.

PHARAOH는 A* 알고리즘의 일종인 빔 탐색법을 이용하며 RAMESE는 PHARAOH를 재구성한 공개프로그램이다. MARIE는 n-gram 언어 모델을 이용한 번역 모델이다.

3. 시스템 설계



(그림 2) 한영/영한 명사구 번역 시스템

이 논문에서 구성한 시스템은 (그림 2)와 같으며 크게 학습시스템과 평가시스템으로 나눌 수 있다. (그림 2)의 왼쪽은 학습시스템이며 오른쪽이 평가시스템이다.

학습시스템은 먼저 웹으로부터 병렬 문서를 수집한

3) SRILM : The SRI Language Modeling Toolkit

다. 이 논문에서는 명사구 번역으로 웹에 공개된 전공용어 사전 및 일반 사전으로부터 한영명사구 쌍이 포함된 문서를 수집한다. 수집된 병렬 문서에서 html 태그 등 불필요한 요소들을 제거하여 병렬 말뭉치를 구축한다. 수집된 말뭉치는 영한, 한영 순서를 뒤집으면 영한, 한영 병렬 말뭉치가 되므로 이것을 학습데이터로 하였다. 병렬 말뭉치에서 각 명사구는 단어 단위로 분리된다. 한국어의 경우 조사가 있을 경우 조사 앞에 기호 '+'를 표시하여 일반적인 단어와 구별될 수 있도록 하였다. 이와 같은 방법으로 구축된 병렬 말뭉치는 평가를 위해서 학습 말뭉치와 실험 말뭉치로 분리된다. 실험 말뭉치의 크기는 전체 말뭉치에 10% 규모로 하였으며 임의로 추출하였다. 구축된 병렬 학습 말뭉치와 GIZA++을 이용하여 단어 변환표와 구절 변환표를 학습하였다. GIZA++ 학습에서 한영 모델과 영한 모델은 모두 같은 방법으로 학습되었으며 M(5:0:3:3:0:0)[4]+HMM(5)으로 학습하였다. 이 논문에서는 같은 병렬 말뭉치를 이용해서 한영과 영한 단어정렬 모델을 학습하였다. 각 목적언어의 언어모델은 SRILM을 이용해서 학습말뭉치로부터 구한다. 이 때 자료부족 현상을 완화하기 위한 방법으로 KNdiscount 방법(Kneser-Ney discounting)을 이용하였으며 n-gram은 3차까지 이용하였다.

평가시스템은 평가말뭉치에 속한 원시언어 문장을 번역기(PHARAOH, RAMSES, MARIE)를 이용해서 번역한 결과를 이용한다. 번역된 문장을 평가말뭉치에 속한 목적언어 문장과 비교하는 방법으로 평가된다. 기계번역의 평가는 매우 어렵다. 왜냐하면 번역된 문장이 얼마나 원시언어 문장의 의미를 전달하는지를 평가해야 하기 때문이다. 최근에 널리 사용되는 평가방법으로 BLEU가 있다. 이 방법은 번역된 문장과 정답문장[5]으로부터 구한 n-gram이 얼마나 많이 교차하는지를 측도로 사용한다.

4. 실험 및 평가

4.1 실험

이 논문에서 웹으로부터 90,270개의 명사구 대역쌍을 수집하였다. 이를 학습 말뭉치와 평가 말뭉치로 분리하였으며 각각 81,243개와 9,027개의 대역쌍을 포함하고 있다. 학습 말뭉치를 이용해서 영한, 한영 번역 모델을 학습하였다. 번역의 성능의 측도로는 앞에서 언급했듯이 BLEU를 사용하며 그 결과는 <표 1>과 <표 2>와 같다.

<표 1> 한영 명사구 번역의 성능

번역기	PHARAOH	RAMSES	MARIE
24,372 (30%)	25.38	27.35	31.72
36,558 (45%)	31.79	35.26	39.19
48,745 (60%)	34.22	38.20	46.77
60,931 (75%)	39.02	42.67	50.27
73,117 (90%)	41.27	45.78	55.92
81,243 (100%)	43.31	48.07	59.41

<표 2> 영한 명사구 번역의 성능

번역기	PHARAOH	RAMSES	MARIE
24,372 (30%)	7.63	7.85	11.96
36,558 (45%)	8.97	9.38	14.43
48,745 (60%)	9.02	9.45	16.03
60,931 (75%)	10.09	10.61	17.54
73,117 (90%)	10.37	10.81	18.17
81,243 (100%)	11.27	11.76	18.67

이 논문에서는 구절 단위로 정렬된 병렬 말뭉치를 이용하여 GIZA++를 통한 자동 학습 방법으로 모델의 파라미터를 학습하였다. 그러므로 결과 값의 정확도는 학습 데이터의 양과 특성에 크게 의존된다. 학습 말뭉치의 크기와 번역 결과의 관계를 파악하기 위해 학습 말뭉치의 일정 비율로 증가시키면서 학습하였으며, <표 1>과 <표 2>에서 보아 알 수 있듯이 모든 시스템이 학습 말뭉치가 커지면 번역 성능이 증가됨을 알 수 있다. 그러나 모든 학습 말뭉치를 사용하였지만 번역 성능이 그다지 좋지 않았다. 이는 아직 학습 말뭉치가 크게 부족함을 간접적으로 말해주고 있다.

<표 1>과 <표 2>에서 보아 알 수 있듯이 번역기의 종류에 따라서 성능의 차이를 보인다. MARIE의 성능이 PHARAOH와 RAMSES에 비해서 우수하게 좋았다. PHARAOH를 개선한 RAMESE는 PHARAOH보다 좀더 나은 성능을 보였다. 그리고 영한 번역보다 한영 번역 결과가 더 좋았다. 이후 좀더 많은 말뭉치를 사용하면 좀더 정확하고, 좋은 성능을 얻을 수 있을 것으로 기대한다.

4) M(5:0:3:3:0:0): IBM 단어정렬 모델은 1~6까지 있는데 각 모델의 반복횟수를 표시하고 있다. 즉 모델 1은 5회 반복하고 모델 2는 수행하지 않음을 나타낸다.

5) 정답문장: 평가말뭉치에 속한 목적언어 문장을 말한다.

5. 결론 및 향후 과제

이 논문에서 통계기반의 정렬기법을 이용한 한영/영한 양방향 명사구 번역 시스템을 설계하고 구현하였다. 단어정렬을 위해서는 GIZA++을 사용하였으며 번역기로는 PHARAOH, RAMSES, MARIE를 사용하였다. 약 4만개의 명사구 대역쌍을 병렬 말뭉치를 구축하여 시스템의 성능을 평가하였다. 한영/영한 명사구 번역 시스템 모두가 최대 59.41%/18.67%의 BLEU를 보였다. 이는 그다지 좋은 성능을 아니라고 생각된다. 이는 학습 말뭉치의 양이 절대적으로 부족하기 때문이다. 특히 영한 명사구 번역의 결과가 특히 떨어진다.

앞으로 좀더 많은 양의 병렬 말뭉치를 지속적으로 병렬 말뭉치를 구축할 수 있는 텍스트 마이닝 기법이 개발되어야 할 것이다. 무엇보다도 한국어 특성에 적합한 단어 정렬 모델이 연구되어야 할 것이다. 또한 개발된 시스템을 다국어 정보검색 시스템에 직접 적용해서 그 효용성을 평가해보아야 할 것이다.

참고 문헌

[1] Ahrenberg, L., Merkel M., Sågvall Hein, A., and Tiedemann, J., (2000). "Evaluation of Word Alignment Systems", *Proceedings of LREC 2000*, pp 1255-1261.

[2] Brown P. F., Cocke, J., Della Pietra, S., Della Pietra, D., Jelinek, F., Mercer, R. and Roossin, P. (1990) "Statistical Approach to Machine Translation", *Computational Linguistics*, vol. 16, no 2, pp. 79-85.

[3] Brown, P. F., Della Pietra, S. A, Della Pietra, V. J., and Mercer, R. L., (1993). "The Mathematics of statistical Machine Translation: Parameter Estimation", *Computational Linguistics*, vol. 19, no.2, pp. 263-311.

[4] Crego, J. M., Marino, J. B., and Gispert, A., (2005). "An Ngram-based Statistical Machine Translation Decoder", *Proceedings of the 9th European Conference on Speech Communication and Technology. Proceedings of IEEE*, vol. 88, no. 8, pp. 1270-1278.

[5] Diab, M. (2000) "An unsupervised method for multilingual word sense tagging using parallel corpora: A preliminary investigation". *Proceedings of the ACL-2000 Workshop on Word Senses and Multilinguality*, pp.1-9.

[6] Fung, P. and Church, K. W., (1994). "K-vec: A New Approach for Aligning Parallel Texts", *Proceedings of the Fifteenth International Conference on Computational Linguistics*, pp.1096-1102.

[7] Germann, U. (2003). "Greedy Decoding for Statistical Machine Translation in Almost Linear Time", *Proceedings of HLT-NAACL-2003*, pp.72-79

[8] Hiemstra, D.(1998). "Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus", *Perer-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds,) Proceedings of the 8th CLIN meeting*, pp. 41-58.

[9] Huang J.-X., Choi, K.-S. (2000). Chinese-Korea "Word Alignment Based on Liguistic Comparison", *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 392-399.

[10] Hutchins, W.J. and Somers, H.L. (1992) *An Introduction to Machine Translation, Academic* Press, London. Academic Press.

[11] Koehn, P., (2004), "Pharaoh: a beam search decoder for phrase based statistical machine translation models", *Proceedings of the 6th Conf. of the Association for Machine Translation in the Americas*, pp. 115-124

[12] Manning, C.D. & Schütze, H. (1999), *Foundations of statistical natural language processing*, Cambridge, MA: MIT Press.

[13] NIST (2005), *Machine Translation Evaluation Official Results*, http://www.nist.gov/speech/tests/summaries/2005/mt05.htm

[14] Och, F. and Ney, H. (2000), "Improved Statistical Alignment Models". *Proc. of the 38th Anneual*

Meeting of the Association for Computational Linguistics, pp. 400-477.

[15] Papineni, K. A, Roukos, S., Ward, T., and Zhu, W.-J. (2002), "Bleu: a Method for Automatic Evaluation of Machine Translation", *Proceedings of the ACL-02,* pp. 311–318.

[16] Patry, A., Gotti, F. Langlais, Ph., (2006). "Mood at work: Ramses versus Pharaoh, Proceedings of the Workshop on Statistical Machine Translation", Proceedings of HLT-NAACL, pp. 126-129, New-York, USA

[17] Rosenfeld, R. (2000). "Two decades of statistical language modeling: where do we go from here?", *Proceedings of the IEEE,* vol. 88, pp.1270-1278.

[18] Simões, A. M., Almeida, J. J. (2003). "NATools - A Statistical Word Aligner Workbench", P*rocessamiento del Lenguaje Natural, 31th,* pp. 217–224.

[19] Smadja, F., McKeown, KR, and Hatzivassiloglou, V. (1996). "Translating collocations for bilingual lexicons: A statistical approach". *Computational Linguistics,* vol.22, no.1, pp.1-38.

[20] Stolke et al., (2002). "SRILM - An Extensible Language Modeling Toolkit," Proc of the, ICSLP, pp. 901-904.

[21] 리금희 , 김동일, 이종혁 (2002). 중-한 대조분석정보를 이용한 단어정렬, 제14회 한글 및 한국어 정보처리 학술발표 논문집, pp. 40-46.

[22] 신중호 (1996). 한국어/영어 병렬 말뭉치에 대한 단어단위 및 구단위 정렬 모델 , 한국과학기술원 전산학과 석사학위 논문.

[23] 장명길, 김영길, 박영찬,(1998), 다국어 정보검색, 정보과학회지, vol. 16, no. 8, pp. 21-31