의료 문서의 특성을 고려한 단어 모호성 해소 연구

Word Sense Disambiguation for Coarse-grained Medical Corpus

송사광, Sa-kwang Song*, 장재원, Jaewon Jang*, 임명은, Myungeun Lim*,
맹성현, Sung Hyon Myaeng**, 박수준, Soo-Jun Park*

*한국전자통신연구원 바이오인포매틱스팀, **ICU

요약 진료 기록 문서(CDA)가 의사들에 의해 작성되기 때문에 많은 전문용어, 약어, 숫자, 기호 등을 포함하고 있다. 본 논문에서는 이러한 특성을 고려하여 문서 내에서 여러 의미로 해석될 수 있는 약어, 중의어 등의 단어 모호성을 해소하고자 의미적 등가 부류를 이용하여 모호성을 해소하였다. 특히 의료문서가 많은 비율의 숫자, 기호를 사용하고 있고 문서 내에서 많은 의미적 유의성을 포함하고 있기 때문에 이들을 불용어로 처리하지 않고 의미적 등가 부류에 포함시킴으로써 진료문서 특성을 반영하였다.

핵심어: 단어 모호성, 진료 기록 문서, CDA 문서, 텍스트 마이닝

1. 서론

기계가 처리할 수 있는 문서들이 기하급수적으로 늘어감에 따라 정보검색, 텍스트 마이닝, 질의응답시스템 등의 연구 중요성이 매우 크게 부각되고 있다. 이러한 연구의 성공적인 수행을 위해 다양한 자연어처리 연구가 선행되어야 하는데 특히 단어의 중의성 해소도 매우 중요한 기본 연구 중의 하나라고 할 수 있다. 특히, EMR(Electronic Medical Record), EHR(Electronic Health Record) 관련 연구 및 사업들의 확대로 인해 진료기록 문서의 폭발적인 증가를 예상할 때 바이오메디칼 도메인의 텍스트마이닝 기술은 그 필요성이 더해가고 있다. 그러나, 논문, 신문기사, 백과사전 등 잘 구성되어진 문서와 달리, 진료기록 문서(CDA: Clinical Document Architecture)[11]는 의사들이 많은 전문용어, 약어, 숫자, 기호 등을 사용하여 직접 작성하고, 또한 작성된 문서가 많은 중의어를 포함하고 있기 때문에 텍스트 처리에 어려움을 겪는다.

본 논문에서는 진료기록 문서 내의 빈출하는 약어뿐만 아니라 일반 단어를 포함한 중의성 문제를 짚어보고 그 해결책으로써 의미적 등가부류를 이용한 방법을 제시하고, 숫자나 기호 등이 중요한 역할을 하는 진료기록문서와 같은 거친(Coarse-grained) 문서 등에서 이들의 중요성과 이를 활용할 수 있는 방법을 제시하고자 한다.

2. 관련 연구

단어의 의미 중의성 문제는 그 영향이 상위 응용인 정보추출(Information Extraction) 시스템, 정보 검색(Information Retrieval) 시스템, 자동 요약(Automatic Summarization) 시스템, 대화 시스템(Dialog System) 시스템 등의 성능에 영향을 미치므로 이를 처리하지 않은 경우, 응용 연구의 효율이 떨어지거나 성능향상이 이루어지지 않을 수 있다. 비록 중의성에 대한 연구가 꽤 오랫동안 진행되어 왔지만, 진료기록 문서와 같이 잘 정리되지 않은 문서에 대한 연구는 거의 전무한 상태이다.

YU et al. [7]은 SVM 기계학습 방법과 One Sense per Discourse Hypothesis를 이용하여 Medline 요약문서집합을 대상으로 자동 약어 모호성 해소시스템을 개발하였다. 그러나 Medline 집합이 언어처리가 용이하도록 정형화된 언어로 쓰여진 것을 감안하더라도 약 84% 정도의 정확도를 얻었다.

Lee et al. [15] 또한 SVM 알고리즘을 이용하여 일반적인 영어 단어들의 의미 모호성을 해소하고 동시에 영어단어를 인두어로 번역하는 연구를 수행하였다. SVM을 위한 특징으로써 unigram, collocation, POS 태그, 의미적 관계 등을 이용하였다.

또한, Ngai et al. [16] 은 FrameNet 온톨로지 카테고리에 어절들을 할당하는 분류 작업처럼 의미적 역할 레이블링하는 supervised 접근방식을 제안하였다. 그는 몇몇 추출된 특징들 뿐만 아니라 FrameNet에 있는 다양한 어휘적이고 문법적인 특징들을 Boosting, SVM, Decision Tree 등과 같은 기

계학습을 이용하였다.

Liu et al. [4] 은 두 개의 medical 문서집합과 한 개의 일반적인 영어 문서집합에 대해 다양한 분류방식을 적용하여 그 성능을 비교하였다. 그들의 분류 방법에는 전통적인 결정 리스트와 그들이 수정한 방식들, 나이브 베이스 분류방식, 그리고 그들이 개발한 혼합된 학습방식 등이 포함된다. 그들이 사용한 특징들에는 local co-occurring words, collocations, 품사태그와 의미 분류에서 변형된 특징들, 그리고 관심어절 주위의 윈도우 사이즈의 변화 등이 있다.

비록 이러한 연구의 효율이 좋은 편이고, 그 중 하나는 95%의 정확성을 얻었지만, 그들의 방법이 본 연구에서 대상으로 하는 진료기록 문서인 CDA문서(의사에 의해 환자의 증상, 처치, 병력 등이 기록된 문서)와 같은 많은 양의 기호나 숫자가 포함되고 언어 분석이 용이하지 않는 문장구조를 갖는 문서집합에서도 적용이 용이하리라고는 확신할 수 없다.

3. 본론

3.1 진료기록 문서의 특성

표 1과 같이 진료기록 문서가 많은 양의 의사들의 전문용어와 약어, 숫자, 기호 등을 포함하고 있고 문서 자체도 일반인이 문장 단위로 분리하기 어려울 정도로 전문화, 특성화되어 있다. 무엇보다도, 많은 빈도로 사용되는 약어의 경우 문장 내에서의 컨텍스트에 따라 그 의미가 달라지는데, 형태소 분석 등의 과정을 거치면 대부분 알려지지 않은 명사로 취급되어서 상위 언어처리 작업 시에 오류가 발생하게 된다. 따라서 이러한 약어 및 일반 어절의 중의성 해소가 선행되지 않는다면 이러한 문제가 상위 어플리케이션의 성능에 파급되어 문제를 발생시킬 확률이 매우 크다.

표 1 진료기록 문서 예제

R/O COPD, R/O pneumothorax 로 check 한 chest PA 상 hilar enlargement 있고 mass shadow 있어 R/O lung cancer 로 w/u 위해 내원함 PMHx > DM/HT/Tb/Hepatic ds(-/-/-/-) Social Hx > smoking(+): 1-2pack/day * 50yr alcohol(+): heavy S/R > G/W(+), E/F(+), f/c(-/-), c/s/r(+/+/-), HA/Dz(-/-) wt. change(+): 10kg loss, dyspnea(-), chest discomfort(-) indigestion(-), epi. soreness(-) A/N/V/D/C(+/-/-/-/-), H/M/H(-/-/-) abdominal pain(-), dysuria(-), foamy urine(-) 123-4.7-90-22 BUN/Cr 53/11.7 CBC 6160-6.9-246k peritosol cell (-) prot 58 LD 3 Iron/TIBC 121/207 ferritin 372 CRP 0.1

표 1에서와 같이 진료기록 문서는 일반적인 문서에서 보기 어려운 많은 약어, 기호, 숫자를 포함하고 있고, 문장이나 단락을 구분하기 조차 어려운 것이 현실이다. 특히 일반적인 언어처리에서 무시될 가능성이 매우 많은 숫자, 기호가 10% 이상을 차지하고 있다.

3.2 약어 통계

그림 1은 서울대학교병원[17]의 진료기록문서인 CDA 문서집합에서 추출한 고빈도 약어에 대한 통계를 나타낸다. 약어는 많은 수가 여러 의미로 해석될 수 있다. 따라서 각 약어가 가지는 의미들의 수를 나타낸 것이다. 그림 1에서 3가지 의미를 가지는 약어가 가장 많은 빈도를 나타냄을 알 수 있다. 심지어는 하나의 약어가 8가지 의미로 해석되는 경우도 있음을 확인할 수 있다. 평균적으로는 한 개의 약어는 3.267개의 의미를 포함하고 있다. 이러한 약어의 중의성이 의료정보 시스템이나 의료정보 사용자 인터페이스의 효율을 낮추는 매우 큰 역할을 한다.

예를 들어, FC라는 약어는 Fronto temporal, Free thyroxine, Fallot tetralogy, function test, full term, foot, flexor tendon 등 7가지 의미를 가지고 있는데, 의료정보 시스템의 검색시 "FC"라는 질의 또는 "Function Test" 질의를 입력 받았을 때 사용자의 필요(needs)와 문서내의 의미의 연결이 어려워 제대로 된 결과를 제시하기 어렵다. 이때 문서내에 포함된 FC라는 약어의 중의성이 해소된다면 사용자에게 보다 명확하고 유용한 정보를 제공할 수 있다는 것은 명백한 것이다. 특히 진료기록 문서와 같이 약어나 기호가 많은 비중을 차지하는 곳에서는 더욱더 중요하다고 할 수 있다.

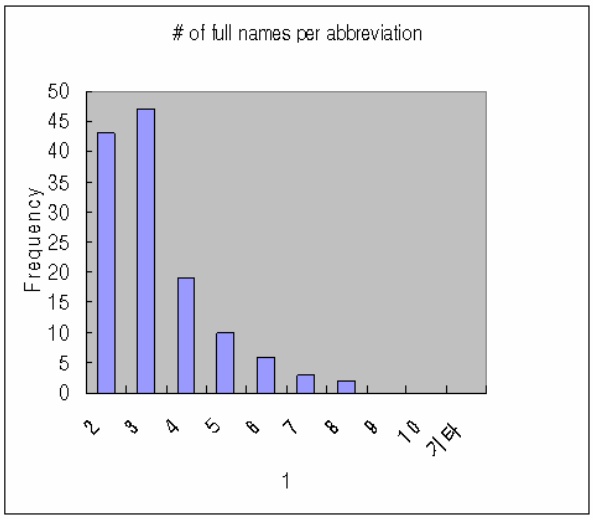

그림 1 약어당 원래 의미의 개수 통계

아래 표 2은 몇몇 약어의 원래의 수와 그 원래 의미, 그리고 문서내의 빈도수를 나타낸 것이다. 예를 들어 ACA는 3가지 의미를 가지고 있는데 Adenocarcinoma, anterior cerebral artery, anterior communicating artery 이고 CDA 문서집합에서 1364번 출현했다는 의미이다.

표 2 약어당 원래 의미의 개수 예

Abbreviation	#	Full Name	N
ACA	3	Adenocarcinoma, anterior cerebral artery, anterior communicating artery	1364
AF	4	Atrial Fibrillation, Atrial flutter, abnormal frequency, acid-fast	418
AS	5	apgar score, activated sleep, anal	109

		sphincter, ankylosing spondylitis, aortic stenosis	
DS	5	dental surgery, dead air space, dead space, deep sleep, down's syndrome	220
FC	7	Fronto temporal, Free thyroxine, Fallot tetralogy, function test, full term, foot, flexor tendon	1038
IMP	6	Internal medicine pulmonary, Inosine monophosphate, idiopathic myeloid proliferation, impression, improved, Important	2950

3.3 의미 등가 부류

일반적으로 단어의 중의 성을 해소를 위해 전후 어절의 패턴을 이용하게 되는데, 본 논문의 대상인 진료기록 문서는 기호, 숫자 등의 빈출로 인해 그 성능을 보장받을 수 없다. 따라서 이러한 기호, 숫자, 빈출 어절 등의 패턴을 분석하여 동일한 의미적 유사성을 갖는 패턴을 하나의 등가부류에 할당함으로써 다양한 어절의 변화를 통일시켜 중의성 해소의 성능을 높이고자 한다.

아래 표 3은 진료기록 문서인 CDA문서 내에 자주 출현하는 패턴의 예들이다, 대부분이 전문용어나 기호 숫자 등으로 쓰여져서 일반인들이 알고 이해하기는 매우 어렵지만, 나름대로 의미적인 패턴을 포유하고 있어서 이러한 패턴을 무시하고 일반적인 언어처리 작업을 수행한다면 매우 많은 정보를 잃어버리는 결과를 낳게 된다.

표 3 고빈도 패턴의 예

T/RT(- /-)
Dorsalis pedis a. (++/++)
Carotid bruit(+ suspicious/-)
F/U/D(++/+/-)
Hx(+ : lt. MRM)
intact to pain DTR : ++,++/++,++ A/C(-/-), Babinski(-/-)
155-5.5/3.2-0.5-65-16/18
15,480-9.0-342K
4700-16.5-52.8-144K
U/A : alb 3+, glu 1+, Bld 2+
11.5%
FEV1 2.02(67%) 2.11 1.92 1.74 1.77 1.80 1.50
0.5 tab tid ++ 1 일 3회 * 14일
95/70 - 72 - 18 - 37.3 ℃
LNE(-/-) V/E(-/-) T/E(-)

따라서, 이러한 고빈도 패턴을 하나의 의미적 등가부류로 태깅하여 정보 손실을 최소화하기 위해 표 4와 같이 의미적 등가부류를 위한 패턴과 역할과 의미 등을 Regular Expression[8] 형태로 표현하였다.

표 4 의미적 등가부류의 예

정규식	역할	의미
₩d+(₩.₩d+)?	{num}	Number
₩d+	{int}	Integer
₩s{0,2}	{s}	Space
([₩x80-₩xff][₩x80-₩xff])+	{kor}	Korean Token
[ ]*[0-2]?[0-9]:[0-6]?[0-9]	TIME	Time
{s}{num}({s}-{s}{num})*{s}((₩(.+?₩))|(.*%))	ABBA	ABGA pattern
{s}{num}{s}({s}-{s}{num}{s})*(/{s}{num}{s}({s}-{s}{num}{s})*){2,5}	PERF	Performance
[a-z]+{s}[a-z0-9]+{s}{int}{s}일{s}{int}회{s}₩*{s}{int}일	DRUG	Drug Dosage

4 실험

진료기록 문서는 POS 태깅, UMLS 태깅, 불용어 제거 과정 등의 전처리 작업을 거친 후 약어나 중의어 전후의 특징값을 추출한 후 SVM과 HMM을 사용하여 중의성 해소 작업을 진행한다. 이때, 특징값은 인접한 단어 빈도수, 가중치 부여 방법, 의미 등가 부류 방법 적용 등의 과정을 거치는데 그림 2은 이 과정을 모식적으로 나타낸 것이다. 특징 추출 방법에 대한 자세한 설명은 다음 절에 설명하고 있다.

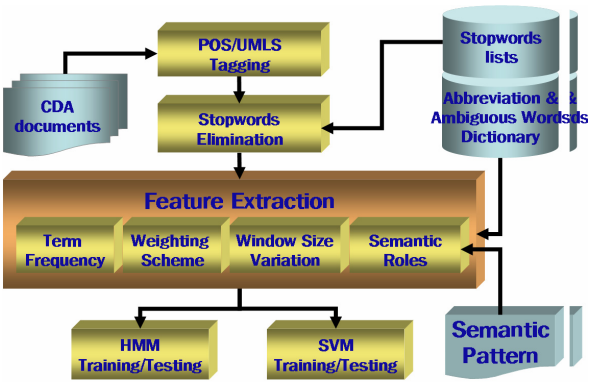

그림 2 시스템 구조도

4.1 Bag-of-words

한 약어 또는 중의어의 의미를 문장 내에서 판단하기 위해서는 보통 전후의 문맥을 활용한다. 이때 가장 간단한 방법이 전후에 인접한 단어들을 활용하는 것인데 단순이 이들의 집합을 이용하는 방법이 Bag-of-words 방식이다.

예를 들어 아래와 같은 문장 내에 'AMBIG'라는 중의어가 있다면 전후의 모든 어절 w1부터 w6까지의 빈도수를 이용하여 특징 벡터를 구성하는 것이다.

w1, w2, w3, w4, (AMBIG), w1, w5, w3, w6

4.2 가중치 방법

위에서 추출된 어절들의 빈도수만 이용하는 대신 각 어절

의 문장 내에서의 가중치를 부여하는 방식을 사용하였다.

$$WGT(t_i) = \frac{TF(t_i)}{\arg max(TF(t_1)...TF(t_N))} * \log(\frac{N_s}{n_s})$$

수식 1 가중치 방법

수식 1에서 $TF(t_i)$ 는 어절 t_i의 빈도수를, N_s는 의미의 총 수를, n_s는 t_i가 포함된 곳에서의 의미의 수를 나타낸다.

4.3 POS(Part of Speech) 태깅

일반적인 품사 태깅 정보를 활용하는 방법이다. 문서가 한글 외에 많은 빈도로 영어 단어/구절을 포함하고 있으므로 한글과 영문 품사태깅 정보를 활용하였다.

4.4 UMLS(Unified Medical Language System) 태깅

대상 문서가 진료기록 문서이므로 의학 용어 및 그와 관련된 표현들이 매우 많이 사용된다. 따라서 이러한 의학 용어 대한 태깅을 위하 UMLS 태거를 수정하여 본 연구에 활용하였다.

UMLS 자체는 매우 방대한 양의 어휘를 포함하고 있고, 그에 대한 의미 분류를 수행할 수 있다. 또한 다중 어절을 포함한 용어의 검출도 가능하기 때문에 매우 유용한 툴이라 할 수 있다.

4.5 의미 등가 부류

3.3 절에서 설명한 의미 등가 부류 표 4에서와 같이 Regular Expression을 활용하여 문서를 태깅하고 태깅된 정보를 활용하는 것이다. 단순히 어절 및 그 빈도를 이용하는 대신 등가 부류 정보와 그 빈도수를 활용함으로써 문맥의 의미 손실을 최소화하여 보다 정확한 성능을 보장할 수 있을 것이다.

4.6 윈도우 사이즈

중의어의 전후를 고려할 때 전후 얼마까지의 어절을 고려할 것인가가 또한 성능에 영향을 미친다. 따라서 본 실험에서는 중의어를 전후로 인접한 단어들의 수인 윈도우 사이즈를 조절해 가면서 실험을 수행하였다. 문장의 시작이나 끝인 경우는 더 이상의 윈도우 확장을 하지 않고 문장이 연속되는 부분만 윈도우 사이즈를 확장하였다.

4.7 실험 세팅

실험은 성능 비교를 위해 4가지 세팅에 대해 가각 진행하였는데, 단순 특징 어절(K), 특징어절(K)+ 태깅정보(T), 특징어절(K)+ 태깅정보(T)+ 의미등가(S), 특징어절(K)+ 태깅정보(T)+ 의미등가(S)+ 가중치부여(W)로 구분된다. 특징어절이라 함은 Bag-of-words와 같이 인접한 어절만을 특징으로 사용했다는 의미이고, 태깅정보는 한글 및 영어 POS 태깅정보 및 UMLS 태깅정보의 사용, 의미등가는 본 논문에서 제안한 의미등가 부류를 사용했다는 의미이며, 가중치는 4.2에서 설명한 가중치식을 사용해서 특징별 가중치를 부여했다는 의미있다.

학습 및 테스트를 위한 문서집합은 아래 표와 같다. 데이터는 프로그램을 통한 자동 수집과 수작업을 통한 수집으로 진행되었다. 수작업은 서울대학교 병원의 대학원생들이 수행하였는데, 하나의 중의어 당 평균 46.5개의 컨텍스트 문장을 수집하였고, 프로그램을 통한 자동 수집은 평균 209.3개의 문장을 수집하였다. 수집된 집합의 2/3을 학습을 위해 사용하였고, 나머지 1/3을 테스트에 활용하였다.

표4 학습 데이터 집합

	Manual Finding	Automatic Finding	Sum
Training Set	31	141.5	172.5
Test Set	15.5	67.8	83.3
Total	46.5	209.3	255.8

5 실험 결과

앞의 4가지 세팅에 대해 실험 결과는 중의성 해소 정확도는 그림 2와 같이 나타났다. 결과를 요약하면 아래와 같다.

- KTSW 세팅의 성능이 가장 좋았다.
- RBF 커널을 이용한 SVM 적용시 window 크기 6에서 최고 성능: window크기에 비례해서 성능 증가하지 않음
- SVM 사용한 방법이 HMM보다 5.2%정도 좋다.
- 태깅정보(T)+ 의미등가(S) 가 정확도 향상에 10% 정도 기여

표 5 SVM을 적용한 실험 결과

Window Size	K (Baseline)	K+T	K+T+S	K+T+S+W
2	0.835	0.871	0.909	0.912
4	0.856	0.907	0.918	0.923
6	0.879	0.918	0.929	0.939

8	0.87	0.905	0.918	0.932

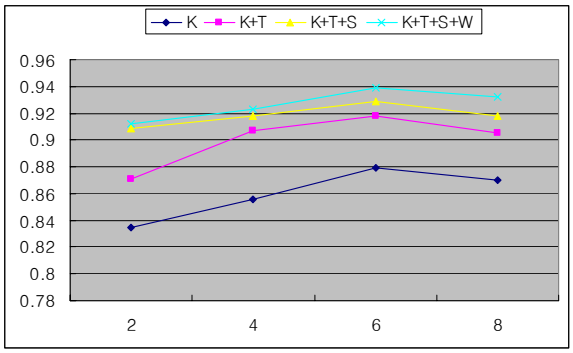

그림 2 실험 결과: window 크기별 각 세팅의 정확도 그래프

아래 표6은, SVM 알고리즘을 사용할 때 다양한 커널을 적용하였으나 윈도우 사이즈가 최대인 30에서 Linear 커널의 실험 결과가 가장 좋게 나타난 것을 보인다. 그러나 추가적인 실험을 통해 윈도우 사이즈 별로 모든 커널의 효율을 비교해 본 결과, RBF 커널이 윈도우 사이즈 6에서 최고 성능을 나타낸 것을 확인할 수 있었다.

Linear 커널의 경우는 윈도우 사이즈가 커짐에 따라 정확도가 향상하는 결과를 보여준 반면, RBF 커널을 적용한 경우는 윈도우 사이즈 6을 경계로 정확도가 상승하지 않고 오히려 감소하는 현상을 보였다. 이러한 현상을 중의어에서 멀리 떨어져 있는 어절이 오히려 RBF 커널을 통한 중의성 해소에 도움이 되지 않는다는 것으로 단순히 인접한 단어의 수를 증가시킨다고 전체 성능이 상승하지 않는다는 것을 의미한다. 이는 최고 성능(최고 윈도우 사이즈)을 갖는 Linear 커널 실험 결과가 윈도우 사이즈 6일 때의 RBF커널의 결과보다 좋지 못하다는 것으로 확인할 수 있다.

표 6 SVM 커널 실험

Kernel Method	Precision
Linear	0.921947
Polynomial	0.912279
RBF	0.878172
Tanh	0.835285

6. 결론

본 논문에서는 진료기록 문서와 같이 조악한(Coarse-grained) 문서에서의 단어 중의성 해소를 위해 기호와 숫자 등을 포함한 의미적 등가 부류 방법을 적용하여 향상된 결과를 얻을 수 있었다.

따라서 진료기록과 같이 일반적인 언어 분석을 통한 응용 프로그램의 적용이 어려운 곳에서도 그 효율을 높일 수 있는 방법을 제시하였고, 기호나 숫자 등이 무시할 수 있는 데이터가 아닌 매우 중요한 정보로 사용될 수 있다는 것을 보일 수 있었다.

참고문헌

[1] Joshi et al. "Supervised Word Sense Disambiguation in the Medical Domain using Support Vector Machines." JAMIA, 2004.

[2] Hongfang Liu, et al. "A Study of Abbreviations in MEDLINE Abstracts," AMIA 2002.

[3] Hongfang Liu, et al. "Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS," JAMIA 2002.

[4] Hongfang Liu et al., "Multi-aspect comparison study of supervised word sense disambiguation", JAMIA 2004

[5] Yaakov HaCohen-Kerner et al. "Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents," EsTAL 2004.

[6] Antonio Molina et al. "A Hidden Markov Model Approach to Word Sense Disambiguation," LNCS 2002

[7] Zhonghua YU et al. "Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis," SIGIR 2003.

[8] Regular Expression HOWTO Homepage by A.M. Kuchling: http://www.amk.ca/python/howto/regex/

[9] Christopher J. C. Burges et al. "A Tutorial on Support Vector Machines for Pattern Recognition," Kluwer Academic Publishers, Data Mining and Knowledge Discovery, 2, 121-167, 1998

[10] 0. Chapelle, V. Vapnik, 0. Bousquet, and S. Mukherjee. "Choosing multiple parameters for support vector machines." Machine Learning, 46(1-3): 131-159, 2002.

[11] What is CDA?: http://www.hl7.org.au/CDA.htm#CDA

[12] SVMlight: http://svmlightjoachims.org/

[13] Unified Medical Language System (UMLS) http://www.nlm.nih.gov/research/umls/

[14] Vladimir N. Vapnik, "The Nature of Statistical Leaming Theory." Springer, 1995.

[15] Yoong Keok Lee et al., "Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources," SENSEVAL-3, ACL, 2004

[16] Grace Ngai et al., "Semantic Role Labeling with Boosting, SVMs, Maximum Entropy, SNOW, and Decision Lists," SENSEVAL-3, ACL, 2004

[17] Seoul National University Hospital http://www.snuh.org/

[18] Sa-Kwang Song, “Abbreviation Disambiguation Using Semantic Abstraction of Symbols and Numeric Terms,” IEEE NLP-KE, 2005

[19] Hyeju Jang, Sa-Kwang Song, Sung Hyon Myaeng, “Semantic Tagging for Medical Knowledge Tracking”, IEEE 2006 International Conference of the Engineering in Medicine and Biology Society, 2006