지지 벡터 기계를 이용한 질의 유형 분류기

안영훈°, 김학수, 서정연
서강대학교 컴퓨터학과 자연어처리 연구실
{cyllian, hskim}@diquest.com, seojy@ccs.sogang.ac.kr

A Question Type Classifier Using a Support Vector Machine

An, Young Hun°, Harksoo Kim, Jungyun Seo
Natural Language Processing Lab., Dept. of Computer Science, Sogang University

요약

고성능의 질의응답 시스템을 구현하기 위해서는 사용자의 질의 유형의 난이도에 관계없이 의도를 파악할 수 있는 질의 유형 분류기가 필요하다. 본 논문에서는 문서 범주화 기법을 이용한 질의 유형 분류기를 제안한다. 본 논문에서 제안하는 질의 유형 분류기의 분류 과정은 다음과 같다. 우선, 사용자 질의에 포함된 어휘, 품사, 의미표지와 같은 다양한 정보를 이용하여 사용자 질의로부터 자질들을 추출한다. 이 과정에서 질의의 구문 특성을 반영하기 위해서 슬라이딩 윈도 기법을 이용한다. 또한, 다량의 자질들 중에서 유용한 것들만을 선택하기 위해서 카이 제곱 통계량을 이용한다. 추출된 자질들은 벡터 공간 모델로 표현되고, 문서 범주화 기법 중 하나인 지지 벡터 기계(support vector machine, SVM)는 이 정보들을 이용하여 질의 유형을 분류한다. 본 논문에서 제안하는 시스템은 질의 유형 분류 문제에지지 벡터 기계를 이용한 자동 문서 범주화 기법을 도입하여 86.4%의 높은 분류 정확도를 보였다. 또한 질의 유형 분류기를 통계적 방법으로 구축함으로써 lexico-syntactic 패턴과 같은 규칙을 기술하는 수작업을 배제할 수 있으며, 응용 영역의 변화에 대해서도 안정적인 처리와 빠른 이식성을 보장한다.

1. 서론

기존의 정보검색(information retrieval, IR)은 사용자의 질문에 대한 응답으로 대량의 문서를 검색하고 순위화하는데 초점을 맞추어 왔다. 그러나 많은 사용자들은 명확한 의도를 가지고 질문을 하며, 정보 검색 시스템이 대량의 문서를 찾아주기 보다는 정답들을 곧바로 찾아 제시해 주기를 바란다[1]. 이러한 요구를 만족시키기 위하여 질의응답(question answering, QA)이라는 개념이 출현했으며, 많은 연구들이 AAAI[2]와 TREC[3]을 중심으로 수행되어 왔다.

질의응답 시스템이 정보 검색 시스템과 다른 점 중 하나는 질의 처리 과정(question processing)에 있다. 질의 처리 과정은 질의에서 사용자의 질의 의도를 파악할 수 있는 질의 유형(question type)이나 키워드(keyword) 등의 정보를 질의로부터 추출하는 것이다. 특히 질의 유형은 질의응답 시스템이 문서에서 정답이 될 수 있는 정답 후보(answer candidate)들을 추출하는데 중요한 정보를 제공한다.

질의응답 시스템이 인터넷과 같은 실용적 환경에서 사용될 경우, 실제 사용자의 질의는 다양한 유형으로 나타나게 된다. 따라서 실용적인 시스템에서 사용되는 질의 유형 분류기(question type classifier)는 문장의 형

태나 단어의 쓰임에 관계없이 같은 의도를 가진 질의를 같은 유형으로 분류해 낼 수 있어야 한다. 예를 들어 "올해의 프로야구 우승팀은?" 과 "어디가 올해 프로야구에서 승리했죠?" 는 형태는 다르지만 같은 의도를 가진 질의다. 또한, 시스템이 한 응용 영역(application domain)에서 다른 영역으로 옮겨질 때 응용 영역 간의 이식이 쉽고, 응용 영역의 변화에 따른 성능 차이가 없어야 한다. 이러한 사항들을 만족시키기 위하여 본 논문에서는 문서 범주화(text categorization)에 많이 이용되는 지지 벡터 기계(support vector machine, SVM)를 이용한 질의 유형 분류기를 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 질의 유형 분류 분야에 수행되었던 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 질의 유형 분류기에 대하여 설명한다. 그리고 4장에서 실험 및 평가를 하며, 5장에서는 결론 및 향후 과제를 기술한다.

2. 관련 연구

질의 유형은 사용자의 질의 의도를 특정한 범주(category)에 할당하는 것으로 질의응답 시스템 연구의 한 분야로 진행되어 왔다[4, 5, 6, 7, 8].

기존의 질의 유형 분류 기법은 규칙에 기반한 방법(rule-based method)과 통계에 기반한 방법(statistical method)으로 나뉘어 진다[4, 5, 6, 9, 10, 11, 12]. 규칙에 기반한 질의 유형 분류를 채택하고 있는 시스템들은 일반적으로 lexico-syntactic 패턴을 구축하고, 이러한 패턴을 유한 상태 오토마타와 매치(match)하여 질의 유형을 분류한다. 일반적으로 규칙에 기반한 접근 방법을 채택한 질의응답 시스템들은 다음과 같은 장점을 가지고 있다.

- 질의 유형 분류 과정이 유한 상태 오토마타로 구현되므로 사용자의 질의에 대해서 즉각적으로 질의 유형을 분류해 낼 수 있다
- 응용 영역이 정해져 있을 경우 간단한 튜닝(tuning)으로 성능을 향상시킬 수 있다.

그러나 규칙에 기반한 질의 유형 분류 방법은 규칙을 수정하기 위해서 전문적인 지식을 가진 사람들의 노력이 필요하고, 규칙과 일치되지 않는 질의가 들어 왔을 때는 질의 유형을 분류할 수 없는 문제점을 가지고 있다. 그리고 규칙이 많아질수록 좋은 성능을 내기 위한 튜닝이 점점 더 어려워지게 된다. 또한 시스템이 다른 응용 영역에서 사용될 경우에는 기존의 규칙들을 모두 수정하거나 재작성 해야 하는 문제점이 있다.

통계적 방법에 기반한 질의 유형 분류는 수동으로 분류된 대량의 학습 데이터로부터 추출한 통계 정보를 이용한다. 통계에 기반한 질의 유형 분류는 다음과 같은 장점이 있다.

- 대량의 학습 데이터를 이용한 통계 모델을 사용하기 때문에 안정적으로 질의의 유형을 분류할 수 있으며, 응용 영역의 변화에 대해 크게 영향을 받지 않는다.
- 자동화된 통계적 방법을 사용함으로써 시스템 구축을 쉽게 할 수 있다.

그러나 통계적 접근 방식은 가끔 사용자가 질의에서 의도하지 않은 결과를 정답으로 출력하는 경우가 있다. 예를 들어 "작년 프로야구는 누가 우승했나요?" 는 우승팀을, "최근 국제 마라톤 대회에서 누가 우승했나요?" 는 우승한 사람을 정답으로 요구하는 질의다. 하지만 두 질의는 구조적으로 매우 유사하므로 질의 유형을 제대로 분리하기 어렵다. 규칙에 기반한 시스템은 이러한 문제를 보완할 수 있는 규칙을 쉽게 수정하거나 추가할 수 있지만, 통계적 방법의 경우에는 보완이 쉽지 않다.

3. 질의 유형 분류기

본 논문에서 제안한 질의 유형 분류기는 크게 질의 학습 과정과 질의 분류 과정으로 나뉘어 진다. [그림 1]

은 질의 유형 분류기의 전체 구성도이다.

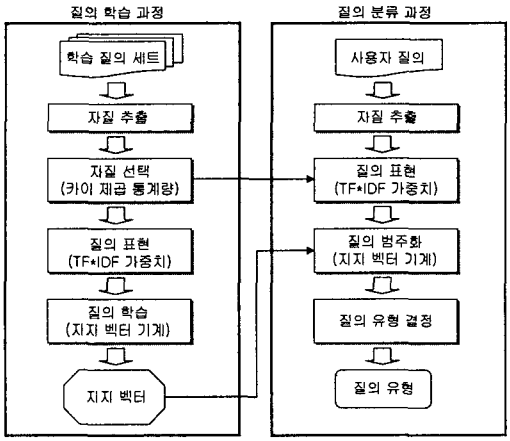

그림 1. 질의 유형 분류기의 전체 구성도

본 장에서는 실험에 사용된 질의 유형 분류 체계와 개체명 사전(named entity dictionary)에 대해서 설명하고, 질의 유형 분류기의 각 부분에 대해서 자세히 설명한다.

3.1 질의 유형과 개체명 사전

본 논문에서는 사용자의 질의 유형을 105개의 의미 범주(semantic category)로 구분하고, 그것에 따라 질의 유형을 분류한다.

표 1. 의미 범주의 일부

계층 1	계층 2		
animal	bird person	fish reptile	mammal
location	address continent town	*building* country	*city* state
date	day weekday	month year	season
time	*hour*	*minute*	*second*
organization	company group team	department laboratory	family school

[표 1]과 같이 105개의 의미 범주는 2개의 계층으로 이루어지며 첫 번째 계층에 속한 의미 범주들은 두 번째 계층에 속한 것들보다 넓은 의미를 지닌다. 본 논문에서는 105개의 의미 범주를 결정하기 위하여 TREC에 참가한 질의응답 시스템들의 의미 범주를 참고하였고, 상업용 정보 검색 시스템[13]에서 수집한 질의 로그(log)를 분석하였다.

질의 유형 분류기는 질의에서 유용한 자질들을 추출하기 위해서 개체명 인식기를 사용한다. 개체명 인식기는 PLO 사전이라 불리는 개체명 사전(named entity dictionary)을 사용해서 질의에 나타난 개체명들을 인식한다. PLO 사전은 표제어와 그에 해당하는 의미 표지로 구성되며 다음과 같은 네 종류의 엔트리를 가진다.

- 고유 명사: 인명, 국가명, 도시명, 기관명 등
- 일반 명사: 직책, 직위, 취미 등
- 단위 명사: km, m, cm, kg, g, mg 등
- 기타: 질의응답 시스템에 필요한 특수한 단어

3.2 자질 추출

자질 추출 과정은 입력된 질의에서 범주화를 위해 필요한 자질들을 추출한다. [그림 2]는 자질 추출 과정을 도식화한 것이다.

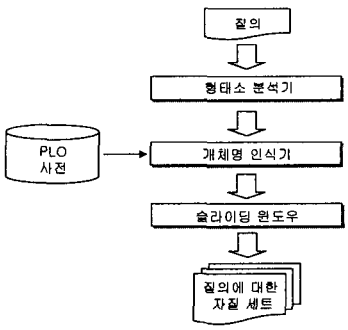

그림 2. 자질 추출 과정

자질 추출 과정은 우선 형태소 분석기를 이용해서 질의

에서 형태소를 추출하고 품사를 결정한다. 개체명 인식기는 분석된 형태소 중 PLO 사전에 존재하는 엔트리에 대해서 의미 표지를 할당한다. [표 2]는 "서강대학교의 총장실 전화번호는?" 에서 분석된 형태소, 품사, 의미 표지를 보여 준다. [표 2]에서 *none*은 해당 형태소에 대한 개체명 인식기의 결과(의미 표지)가 없음을 의미한다. '서강대학교' 에 해당하는 의미 표지는 *@location*과 *@organization*으로, 이 단어가 위치나 조직명으로 쓰일 수 있음을 나타낸다. 이와 같이 개체명 인식기가 하나의 형태소에 대해서 두 개 이상의 의미 표지를 부착할 경우, 자질 추출 과정은 의미 표지의 가능한 모든 조합으로 자질 세트를 생성한다.

표 2. 자질 추출 예제

형태소	품사	의미 표지
서강대학교	nq_loc	(@location|@organization)
의	j	none
총장실	ncn	@location
전화번호	ncn	%tel_num
는	j	none
?	sf	none

개체명 인식기를 거쳐서 추출된 자질 세트는 슬라이딩 윈도 방법을 이용하여 자질 패턴으로 만들어진다. [그림 3]은 슬라이딩 윈도의 처리 방법을 나타낸 것이다.

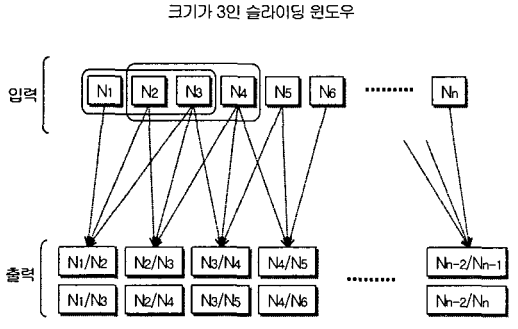

그림 3. 슬라이딩 윈도 방법

슬라이딩 윈도는 형태소 단위로 처리되며, 형태소에 의미 표지가 부착되어 있다면 의미 표지를 이용하고, 의미 표지가 없다면 형태소 분석 결과를 슬라이딩 윈도의 입력으로 이용한다. 이 때 형태소의 품사에 따라서 다음과 같이 다르게 처리한다.

- 체언, 용언, 외국어, 미등록어 : 형태소(또는 의미 표지)와 품사를 결합해서 이용
- 기호, 수식언을 제외한 기타 품사 : 품사 정보만을 이용
- 기호, 수식언 : 사용하지 않음

3.3 자질 선택

Yang은 [14]에서 여러 가지의 자질 선택 방법을 사용하여 실험을 한 결과 카이 제곱 통계량과 정보 획득량을 사용하는 것이 범주 할당 문제에 가장 효과적임을 보였다. 본 논문에서는 이를 바탕으로 비교적 구현이 쉽고 고빈도 단어에 친화적인 카이 제곱 통계량을 사용하여 자질을 선택한다. 카이 제곱 통계량을 구하기 위한 수식은 (식1)과 같다.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

(식1)에서 A는 범주 c에 속해 있는 질의 중 용어 t를 포함하고 있는 질의의 수, B는 범주 c에 속하지 않은 질의 중 용어 t를 포함하고 있는 질의의 수, C는 범주 c에 속해 있는 질의 중 용어 t를 포함하지 않은 질의의 수, 그리고 D는 범주 c에 속하지 않은 질의 중 용어 t를 포함하지 않은 질의의 수이다.

(식1)의 계산이 끝나면 (식2)에 따라 각 범주별로 얻어진 카이 제곱 통계량 중에 가장 큰 값을 해당 용어의 자질 값으로 할당한다. 그리고 이 값을 순위화하여 내림차순으로 정렬한다.

$$\chi^2_{\max}(t) = \max_{i=1}^{m} \{\chi^2(t, c_i)\} \quad (2)$$

카이 제곱 통계량에 의해 얻어진 순위를 바탕으로 정해진 순위 안에 들지 못한 자질들은 질의 유형 분류를 위한 의미 있는 정보를 제공해 주지 못하는 것으로 판단하여 자질에서 제외시킨다.

3.4 질의 표현

질의 표현을 위해서 가장 일반적으로 사용하는 문서 표현 방법은 벡터 공간 모델이다. 이것은 문서 전체에 나타난 각 자질의 빈도를 이용하여 문서를 하나의 벡터로 표현하는 것으로, 보통 자질의 빈도(TF)와 역문헌빈도(IDF) 혹은 역범주빈도(ICF)를 이용하여 가중치를 줌으로써 문서를 표현한다. 본 논문에서는 질의에서 추출된 자질들을 문서로 취급하고, (식3)과 같은 tfc-가중치(tfc-weighting) 방법을 이용하여 각각의 질의를 벡터화한다.

$$a_{ik} = \frac{f_{ik} \times log\,(\frac{N}{n_k})}{\sqrt{\sum_{j=1}^{K} [f_{ik} \times log\,(\frac{N}{n_j})]^2}} \quad (3)$$

(식3)에서 f_{ik}는 i번째 질의에서 추출한 자질 k의 출현빈도이고, N은 전체 질의의 수이며, n_k는 자질 k가 출현한 질의의 수이다. K는 질의를 표현하는 데 사용되는 전체 자질의 수이다.

3.5 질의 범주화

제안한 질의 유형 분류기는 지지 벡터 기계를 이용하여 질의를 범주화한다. 질의 학습 과정에서는 학습 질의를 바탕으로 지지 벡터 기계를 학습한다. 학습의 결과로 질의 유형 분류 체계에 대한 지지 벡터들이 생성되며, 생성된 지지 벡터들을 이용해서 질의 분류 과정에서는 입력된 질의의 범주를 구분한다. 지지 벡터 기계를 구현한 프로그램은 Joachims에 의해 구현된 SVM-light[15]를 사용하였다.

3.6 질의 유형 결정

자질 추출 과정에서 질의에 대한 자질 세트가 두 개 이상 생성된 경우, 각 자질 세트에 대해서 범주화 결과가 생성된다. 질의 유형 분류기는 [알고리즘 1]과 같은 방법으로 생성된 범주화 결과 중 가장 적합한 것을 해당 질의의 질의 유형으로 결정한다. 예를 들어 "월드컵의 개최 요일은?" 질의에 대해서 3개의 질의 세트가 생성되어 *date*, *day*, *weekday*로 범주화 되었다면, 두 번째 계층에 속한 *day*나 *weekday* 중 범주화 확률이 높은 쪽을 질의의 질의 유형으로 결정하게 된다.

1. 범주화 결과를 의미 범주별로 그룹화(grouping)한다.
2. 각 그룹에서 가장 높은 범주화 확률을 가지는 값을 해당 그룹의 범주화 확률로 선택한다.
3. 범주화 확률이 가장 높은 그룹을 선택한다.
4. 선택한 그룹이 의미 범주의 첫 번째나 두 번째 계층만으로 구성되어 있다면 범주화 확률이 가장 높은 값을 질의 유형으로 결정한다.
5. 선택한 의미 범주가 두 계층 모두 값을 가지고 있다면 두 번째 계층에 속한 결과 중 범주화 확률이 가장 높은 값을 질의 유형으로 결정한다.

알고리즘 1. 질의 유형 결정 과정

4. 실험 및 평가

4.1 실험 데이터

제안된 질의 유형 분류기의 성능을 실험하기 위해서 사용된 데이터는 서강대학교(www.sogang.ac.kr)와 코리아인터넷닷컴(korea.internet.com)과 같은 실제 웹사이트에서 수집한 사용자 질의 로그이다. 수집된 질의는 수작업으로 미리 정의된 질의 유형에 따라서 분류하였다.

수집된 데이터는 총 78개의 질의 유형으로 구분된 7,726개의 질의로 구성된다. 각 질의는 하나의 질의 유형만을 가지며 중복 할당을 허용하지 않았다. 질의 유형 분류기에서 사용하는 의미 범주는 105개이지만, 27개의 의미 범주에 대해서는 실제 수집된 질의에서 나타나지 않았다. 그리고 33개의 의미 범주에 대해서는 수집된 질의에서의 출현 빈도가 10회 미만이었다.

실험 데이터를 학습 데이터와 테스트 데이터로 분리하기 위해서 질의 유형별 분포에 맞춰서 임의로 10%의 데이터를 추출해서 테스트 데이터로 사용하고, 나머지 90%는 학습 데이터로 사용하였다.

4.2 실험

자질 추출 과정에서 슬라이딩 윈도의 크기가 성능에 미치는 영향을 알아보기 위해서 윈도의 크기를 2에서 10까지 늘려가면서 실험하였다. [그림 4]는 슬라이딩 윈도의 크기에 따른 분류 정확성을 비교한 그래프로, 슬라이딩 윈도 크기 6 이후에는 더 이상의 성능 향상이 없거나 성능이 떨어지는 경우가 있음을 확인할 수 있다. [그림 4]에서 person, desc, loc는 각 의미 범주에 대한 실험 결과를 나타낸다. total은 모든 의미 범주에 대한 실험 결과를 나타낸다.

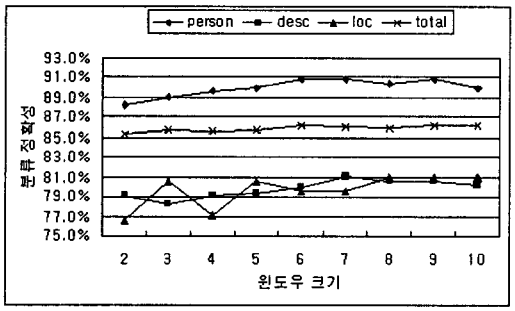

그림 4. 슬라이딩 윈도의 크기에 따른 비교 그래프

제안한 질의 유형 분류기는 슬라이딩 윈도의 크기를 6으로 설정(추출한 모든 자질 이용)하고 2차 함수를 이용한 비선형 결정면으로 지지 기계 벡터를 학습했을 때 가장 좋은 분류 성능을 보였다. 이 때의 분류 정확도는 86.4%이다. 자질의 수를 5,000개로 제안을 하였을 경우의 분류 정확도는 85.9%이다. 자질의 수가 5,000개가 넘어 가면서 성능 향상은 미미하였다.

본 논문에서 제안한 질의 유형 분류기의 성능을 평가하기 위해서 기본 시스템과 규칙 기반 시스템을 이용하였다. 기본 시스템은 다음과 같이 질의에 나타난 의문사와 초점 단어의 의미 표지에 따라서 휴리스틱(heuristic)으로 질의 유형을 결정한다.

- 질의에 '누구', '언제', '어디서' 같은 의문사가 포함되어 있다면 의문사에 의거해서 질의 유형을 결정한다. 예를 들어서 "삼성전자의 사장은 누구죠?" 라는 질의에 대해서 *person*을 질의 유형으로 결정한다.
- 질의에 의문사가 포함되어 있지 않다면, 질의의 마지막 초점 단어의 의미 표지에 의거해서 질의 유형을 결정한다. 예를 들어서 "삼성전자의 사장은?" 이라는 질의에 대해서, '사장'의 의미 지표가 *@person*이므로 *person*으로 질의 유형을 결정한다. 이 때 초점 단어가 여러 개의 의미 표지를 가진다면 제일 처음의 것을 선택한다.

규칙 기반 시스템으로는 한국어 질의응답 시스템인 [16]에서 채택한 방법을 사용하였다. [표 3]은 각 시스템의 성능을 비교한 것이다.

표 3. 질의 유형 분류기의 성능 비교

시스템 종류	분류 정확도(%)
기본 시스템	61.8
규칙기반 시스템	84.0
제안한 시스템	86.4

[표 3]에서 보듯이 제안한 질의 유형 분류기는 기본 시스템에 비해서 24.6% 향상된 성능을 보였다. 또한 규

칙 기반 시스템보다 2.4% 정도 더 좋은 성능을 보였다. 이것은 제안한 방법을 사용하면 복잡한 규칙을 기술하지 않고도 충분히 높은 분류 정확도를 얻을 수 있다는 사실을 보여 준다.

5. 결론

본 논문에는 질의 유형 분류 문제에 문서 범주화 기법을 도입한 자동 질의 유형 분류기를 제안하였다. 질의 유형 분류기는 질의로부터 자질들을 추출하기 위해서 형태소 분석기와 개체명 인식기에 기반한 슬라이딩 윈도 기법을 이용한다. 그리고 문서 범주화 기법에서 좋은 성능을 내는 지지 벡터 기계 모델을 이용하여 질의 유형을 분류한다.

실험은 실제 웹 사이트에서 수집한 7,726개의 질의를 이용하여 수행되었다. 실험을 통하여 슬라이딩 윈도 크기를 6으로 하고, 지지 벡터 기계를 2차 함수를 이용한 비선형 결정면으로 학습했을 때, 가장 좋은 성능을 내는 것을 알 수 있었으며, 그 결과 86.4%의 높은 질의 분류 정확도를 얻었다.

6. 참고 문헌

[1] Voorhees E. and Tice D. M., "Building a Question Answering Test Collection", In *Proceedings of SIGIR 2000*, pp. 200-207, 2000.

[2] AAAI Fall Symposium on Question Answering, *http://www.aaai.org/Press/Reports/Symposia/Fall/fs-99-02.html*

[3] TREC (Text REtrieval Conference) Overview, *http://trec.nist.gov/overview.html*

[4] Moldovan D., Harabagiu S., Paşca M., Mihalcea R., Goodrum R., Gîrju R. and Rus V., "LASSO: A Tool for Surfing the Answer Net" , In *Proceedings of The Eighth Text REtrieval Conference (TREC-8)*, from http://trec.nist.gov/pubs/trec8/t8_proceedings.html, 1999

[5] Prager J., Radev D., Brown E. and Coden A., "The Use of Predictive Annotation for Question Answering in TREC8" , In *Proceedings of The Eighth Text REtrieval Conference (TREC-8)*, from http://trec.nist.gov/pubs/trec8/t8_proceedings.html, 1999

[6] Kupiec J., "Murax: A Robust Linguistic Approach for Question Answering Using an On-line Encyclopedia" , In *Proceedings of SIGIR'93*, 1993

[7] Vicedo J. L. and Ferrández A., "Importance of Pronominal Anaphora resolution in Question Answering systems" , In *Proceeding of ACL 2000*, pp. 555-562, 2000

[8] Prager J., Brown E. and Coden A., "Question-Answering by Predictive Annotation" , In *Proceedings of SIGIR 2000*, pp. 184-191, 2000

[9] Hermjakob U., "Parsing and Question Classification for Question Answering" , In *Proceedings of the ACL Workshop Open-Domain Question Answering*, pp. 17-22, 2001

[10] Ittycheriah A., Franz M., Zhu W. and Ratnaparkhi A., "IBM's Statistical Question Answering System" , In *Proceedings of the Ninth Text REtrieval Conference,* http://trec.nist.gov/pubs/trec9/t9_proceedings.html, Maryland, 2000.

[11] Ittycheriah A., Franz M., Zhu W. and Ratnaparkhi A., "Question Answering Using Maximum Entropy Components" , In *Proceedings of NAACL*, 2001

[12] Mann G. S., "A Statistical Method for Short Answer Extraction" , In *Proceedings of the ACL Workshop Open-Domain Question Answering*, p

p. 13-30, 2001

[13] diquest, http://www.diquest.com

[14] Y. Yang and J. O. Pederson, "A comparative study on feature selection in text categorization", In *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[15] SVM-light, http://ais.gmd.de/~thorsten/svm_light

[16] Kim H., Kim K., Lee G. G. and Seo J., "MAYA: A Fast Question-answering System Based On A Predictive Answer Indexer", In *Proceedings of the ACL Workshop Open-Domain Question Answering*, pp. 9-16, 2001