문맥과 위치정보를 사용한 정보추출

민경구[O] 선충녕 서정연[*]
서강대학교 컴퓨터학과[O*] 다이퀘스트

mingk24@gmail.com[O], wilowisp@diquest.com, seojy@sogang.ac.kr[*]

Information Extraction Using Context and Position

Kyungkoo Min[O], Choong-Nyoung Sun, Jungyun Seo[*]
Department of Computer Science, Sogang University[O*], Diquest Inc.

요 약

인터넷의 발달로 전자문서가 증가함에 따라, 정보추출기술의 중요성도 함께 증가하게 되었다. 정보추출 (IE)은 다양한 형태의 문서로부터 필요한 내용만을 추출하여 정형화된 형태로 저장하는 문서 처리기술이다. SIES (Sogang Information Extraction System)는 기계학습 방법과 고정일의 수동작성 된 규칙기반의 방법론을 함께 사용하는 정보 추출시스템으로 운법에 맞지 않는 운장 등의 입력에 대해 견고한 문장분석을 위해 Lexico-Semantic Pattern (LSP)과 개체명사전 (Named Entity Dictionary)를 사용하였으며, SIES의 기계학습의 성능향상을 위해 기존에 널리 사용되는 문맥정보 외에 후보단어들의 위치정보를 고려한 특성자질과 스코어링 방법을 사용하였다.

1. 서론

정보추출 (IE)은 다양한 형태의 문서로부터 필요한 내용만을 추출하여 정형화된 형태로 저장하는 문서 처리기술이다. 인터넷의 발달로 전자문서가 증가함에 따라, 정보추출기술의 중요성도 함께 증가하게 되었다. 대량의 문서를 쉽게 살펴보기 위한 방법으로 문서요약 기술이 있다. 문서요약에서는 주로 문서내의 중요한 의미를 가지는 몇 개의 문장을 추출하는 방법을 사용하므로 요약 결과를 개별 항목으로 분류하여 저장할 수 없기 때문에 추출된 결과를 NLIDB (Natural Language Interface of Database) 나 QA(Question Answering)등과 같은 분야에서의 지식베이스로의 활용이 곤란하다는 문제가 있고, 이와 같은 목적을 위해서는 정보추출기법이 더 적합하다고 할 수 있다. 정보추출 방법을 일반 문서에 적용할 때 생기는 문제점으로는 많은 인터넷 상의 문서들이 생략, 비문, 구어체 운장 등을 많이 포함하고 있다. 따라서 전통적이고 심화된 언어분석 방법보다는 좀 더 견고하고 가벼운 언어 분석 모듈이 필요하다 [1].

본 연구에서는 스케줄을 위한 이메일 문서를 대상으로 하였다. 인터넷 문서의 특징을 반영한 견고하고 가벼운 언어분석을 위해 복잡한 구문분석대신 Lexico-Semantic Pattern (LSP)을 사용한다. 수동으로 구축된 규칙을 사용하는 정보추출방법론 [2][3][4]을 개선한 SIES (Sogang Information Extraction System)의 기계학습을 위한 특성자질 (문맥정보, 운장정보, 문서정보)을 소개한다. (그림 1)

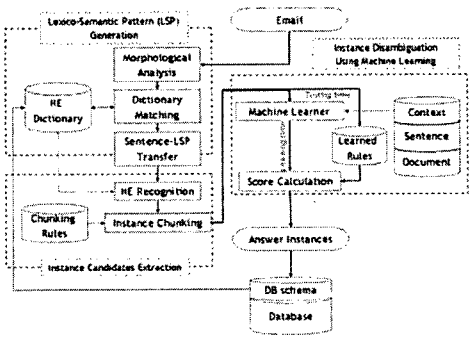

그림 1 시스템 구조도

논문의 구성은 다음과 같다. 먼저 2장에서 전체 시스템의 구조와 개요와 정답후보추출에 대한 내용을 소개하고,

3장에서 특성자질에 대해 설명한다. 4장에서 실험결과를 설명한 후, 5장에서 결론을 내린다.

2. 정답 후보 추출

본 시스템에서는 먼저 입력문서 중 정답 후보를 추출하고 기계학습을 통해 가장 적합한 후보를 판별해 정답으로 추출하는 방식을 사용한다 [2][3]. 정답 추출을 위해서 먼저 입력된 문서는 문장단위로 LSP 형태로 변환된다. LSP는 하나의 문장에서 형태소 열을 추상화하기 위해 어휘와 의미 타입을 함께 사용하는 형태를 가진다. 어휘로는 형태소 단위의 품사들을 사용하고, 의미 타입은 개체명 사전의 의미 태그를 사용한다 [5]. LSP의 의미태그 변환을 위해서 사용되는 개체명 사전은 명사 위주로 구성되어있다. 개체명 사전의 각 엔트리는 두 종류의 의미태그를 가지는데, 하나는 개념이며, 다른 하나는 개념의 인스턴스이다. LSP 형태의 입력문장은 NE 태거를 통해 후보 인스턴스를 선정하게 된다. 우리는 Maximum Entropy와 신경망, Pattern Selection 모델을 복합적으로 사용한 NE 태거를 이용하였다[6]. [6]의 NE 태거를 본 실험 영역에 맞게 수정하기 위해 15개의 세부 영역으로 구분되어 있는 NE 태그를 인명, 지명, 조직명으로 통합하고 날짜와 시간 타입 추출을 위해 간단한 오토마타를 추가하였다.

3. 기계학습을 이용한 정답 추출

NE 태거의 추출결과 하나의 문서에 동일한 타겟필드를 가지는 둘 이상의 후보 인스턴스가 발생할 경우 (예를 들어 '10시', '오후 2시'라는 두 개의 인스턴스를 '약속시간' 이라는 타겟필드로 추출하였을 경우), 정답을 판별하여야 한다. 이를 후보 인스턴스의 모호성 해소라고 하고, '문맥자질', '문장자질', '문서자질' 특성을 사용하여 학습한 결과를 정답으로 추출하게 된다.

첫 번째, 문맥자질은 후보인스턴스의 좌우에 등장하는 단어정보를 사용하는 것으로 단어의 어휘를 그대로 사용하지 않고, 앞서 변환된 LSP를 사용한다. 문맥자질을 사용함으로써 어휘레벨의 규칙에서 발생할 수 있는 희소성 (sparseness) 문제를 줄일 수 있다. 문장자질과 문서자질은 추출대상영역의 문서특성이 반영된 것으로 우리가 추출대상으로 삼는 이메일 문서의 타겟필드(시간, 날짜, 장소, 참석인)는 인접 인스턴스의 주위문맥이나 문장에서 출현하는 경향을 보인다. 따라서 우리는 후보인스턴스 주위에서 후보인스턴스와 동일한 유형/다른 유형의 후보의 출현빈도 정보를 사용하기로 하였다. 문서정보는 문장특성을 문서범위로 확장한 것으로 후보인스턴스의 문서 내 위치, 주변문장들에서의 인스턴스의 분포, 제목필드와 동일한 인스턴스의 출현 등의 정보를 특성화 한 것이다. 이러한 특성 자질을 학습하기 위해 WEKA Toolkit [7]에 포함되어 있는 Naïve Bayes 학습 [8]을 사용하였다.

기계학습 후에도 후보인스턴스들간의 모호성이 해소되지 않을 경우의 우선순위 결정을 위해 후보 인스턴스의 신뢰도를 계산하였다. 이메일 문서 목적을 위한 문서의 초점이 되는 인스턴스들은 다른 타입의 정답 인스턴스들과 함께 출현하는 경향이 있다. 이를 반영하기 위한 빈도점수와 위치 점수는 다음과 같이 후보인스턴스의 빈도점수와 위치점수의 합으로 나타내어진다 (그림 2). 빈도 점수는 주어진 후보의 문서 내에서의 중요도를 나타내며, 위치점수는 얼마나 많은 인스턴스들이 뭉쳐있는지를 반영한다.

$$Conf_i\_ML = frequncy\_Score_i + location\_Score_i \quad (2)$$

$Conf_i\_ML$: confidence score of instance$_i$ in step - 2

$$frequncy\_Score_i = \frac{\text{frequency of instance}_i}{\text{number of same type candidates with instance}_i} \quad (3)$$

$$location\_Score_i = \frac{\text{different type from instance}_i \text{ within surrounding sentences}}{\text{total number of type}} \quad (4)$$

그림 2 신뢰도 계산

그림 3은 신뢰도 계산의 한 예를 보여준다. 추출을 위한 타겟필드를 T, 각각의 T에 대한 인스턴스를 F라고 하면 T = {t_i | t_i ∈ 미리정의 된 NE 타입, i < 타겟필드의 수 }. F = {f_{ti} | f_{ti} ∈ real instance of target field t} 와 같다.

1: ___f_{a1}___ ← surrounding sentence
2: __ f_{a1}__ f_{a2}_____ f_{b1}_ f_{c1}_
3: ____ f_{a1}_____ ← surrounding sentence
4: __ f_{d2}_____

{ f_{a1}, f_{a2}, …} ∈ t_a
a, b, c, d: instance which have NE type A, B, C, D.
$Conf_{a1}\_ML$ = 0.75+0.75

그림 3 신뢰도 계산의 예

또한, 타겟 필드의 인스턴스 타입을 A, B, C, D라고 하고 각각의 인스턴스를 f_a, f_b, f_c, f_d라고 하면, 2째 줄의 'f_{a1}'에 대한 기본 0.75의 빈도점수와 ('f_a' 타입의 NE는 4개의 인스턴스를 가지는데, 'a1'은 3회 등장한다) 0.75의 위치점수를 가지며 (4종류의 타겟이 존재하고 주변 문장에서 3개의 타입이 관찰되었다.) 최종 신뢰도는 1.5 (0.75 + 0.75)를 가지게 된다.

4. 실험 및 결과

전자메일에서 스케줄 정보를 추출하는 것은 정보추출 기술이 적용될 수 있는 좋은 예이다. 실험에 필요한 문서 수집을 위해, 약속을 잡기 위한 245개의 관련메일을 수집하였고 각각의 메일은 평균 23.5개의 문장으로 이루어져있다. 추출대상항목으로는 '참석자', '장소', '시간', '날짜'의 4개를 선정하였고, 추출된 결과를 저장하기 위해 MySQL (MySQL 3.23.58 for Linux) 을 사용하였다.

NE 태거를 사용한 후보추출에서 1,376개의 NE 인스턴스 중 1,386개의 인스턴스를 추출하였고 이중 1,314개의 올바른 NE를 추출하였다. 문서 별로 보면 문서의 타겟필드당 평균 2.15 개의 후보인스턴스들이 등장하였고, 모든 문서에서 복수후보가 발생하였다.

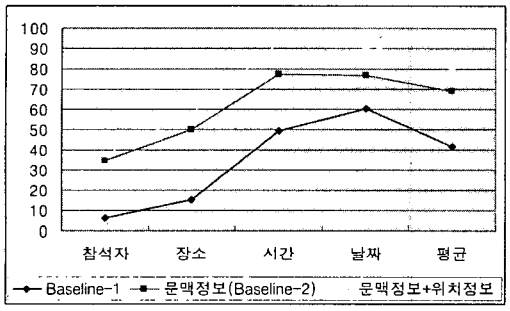

그림 4 F1-measure의 비교

문장정보, 문서정보를 사용한 방법의 유효성을 알아보기 위해, 두 개의 베이스라인을 만들었다. 첫 번째 모델은 (Baseline-1) 후보인스턴스 생성 후 첫 번째 후보를 정답으로 카운트했을 경우의 성능을 측정한 것이고, 두 번째 모델은 (Baseline-2) 후보 인스턴스 주변의 문맥정보만을 사용한 기계학습 결과이다. 마지막으로 모든 자질을 추가하여 실험하였다. 그림 4는 네 개의 타겟 필드에 대한 각각의 방법론의 결과를 보여주고 있다. 인스턴스간의 위치정보를 사용한 결과는 문맥만을 사용한 결과에 비해 평균 18%의 향상된 성능을 보여준다.

5. 결론 및 향후 과제

본 연구에서는 정보추출 시스템 SIES의 기계학습방법에 대해 살펴보고 스케줄링을 위한 한국어 이메일에 적용해 보았다. SIES는 복잡한 문장분석 대신 형태소분석결과와 의미사전을 이용한 LSP를 사용하여 견고한 언어분석을 가능하게 하였다. 또한 NE 태거를 사용하여 정답후보를 추출하고 문맥정보 외에 문장/문서 단위의 위치정보를 사용한 기계학습방법과 인스턴스의 빈도와 문서 내 위치를 고려한 스코어링 방법을 사용하였다.

현재 성능의 향상을 위한 정확도를 향상시키기 위한 방법에 대한 연구가 진행 중이며, 현재보다 문서의 구조를 잘 반영할 수 있는 방법에 대한 연구가 계속되어야 할 것이다.

6. 참고 문헌

[1] F. Ciravegna, "Adaptive information extraction from text by rule induction and generalization," IJCAI 2001

[2] K. Min, H. Jung, J. Seo, Context-Based Information Extraction on E-mail Texts," Asia Information Retrieval Symposium 2004, 2004

[3] K. Min, H. Jung, J. Seo, "Information Extraction on E-mail Texts for Personal Information Management Domain," International Association for Development of the Information Society WWW/Internet 2004, 2004

[4] H. Jung, K. Min, W. Kim, W. Sung, and D. Park, " Query Analysis Using Context-Based Information Extraction on Navigation Domain," Proceedings of the 30th Annual Conference of the IEEE Industrial Electronics Society, 2004

[5] A. Mikheev, and S. Finch, "Towards a Workbench for Acquisition of Domain Knowledge from Natural Language," In Proceedings of the 7th Conferenc4e of the European Chapter of the Association for Computational Linguistics, 1995

[6] C. Seon, Y. Ko, J. Kim, J.Seo, "Named Entity Recognition using Machine Learning Methods and Pattern-Selection Rules," NLPRS, 2000

[7] http://www.cs.waikato.ac.nz/~ml/weka/

[8] T. Mitchell, "Machine Learning," McGraw-Hill, 1997