한국어 비정형 회의공지 이메일 문장에서의 장소정보 추출을 위한 자질선택문제

김경렬○ 최동현 김은경 최기선
한국과학기술원 시맨틱웹첨단연구센터
{barnabas, cdh4696, kekeeo}world.kaist.ac.kr, ks.choi@cs.kaist.ac.kr

Feature Selection for Meeting Location from non-itemized Meeting Email Announcement in Korean

Kyoung-Ryol Kim○ Dong-Hyun Choi Eun-Kyung Kim Key-Sun Choi
Semantic Web Research Center, KAIST

1. 서론

이메일은 오늘날 각종 정보를 주고받는 수단으로 우리 생활에 이미 자리매김하였고, 다양한 종류의 정보들이 이메일을 통하여 매일 수신되고 있다.

그 중에서도 각종 회의와 약속 등의 정보가 담겨있는 회의공지메일은 특별히 중요한 정보를 담고 있기 때문에 많은 사람들은 이를 별도로 분류하여 메일에 있는 약속 시간과 장소 등의 정보를 확인하고 개인적인 수첩에 메모를 해두곤 하는데, 최근에는 각종 캘린더 프로그램을 이용하여 간편히 일정을 관리하는 사람들이 늘어나고 있다. 만약 이메일에서 일정정보를 자동으로 추출할 수 있다면 이메일을 읽고 수동으로 캘린더 프로그램에 일정을 등록하고 각 참석자에게 일정에 대한 소식을 알리는 수고를 프로그램이 자동으로 처리하게 할 수 있을 것이다. 본 연구에서는 지도학습방법을 사용하여 한국어로 작성된 회의공지 이메일 내 비정형 문장에 포함되어 있는 일정정보 중에서 중요한 것 중 하나인 '회의장소'정보를 추출하고자 한다.

2. 본론

정보추출을 위한 기계학습방법으로는 Hidden Markov Models, Maximum Entropy Markov Models, Conditional Random Fields 등이 주로 사용되는데, [1]에서 CRFs 는 HMMs 와는 다르게 조건부 확률을 구하는 특성 때문에 Independent assumption을 필요로 하는 HMMs 보다 성능이 뛰어나다고 실험적으로 증명이 되었고, MEMMs 과 비교했을 때 Label bias problem 이 해결되어 CRFs 는 두 가지 모델보다 나은 성능을 보인다고 기술했다. 본 논문에서는 CRFs를 기계학습 모델로 이용하였고, 추출할 정보타입은 다음의 5가지로 분류하였다. 형태소 단위로 분리된 후 해당되는 정보타입을 갖는다. IHA는 회의장소를 의미하는 관계명인 isHeldAt 의 줄임말로 사용하였다.

- IHA_I : 단일형태소 자체로 하나의 정보타입이 될 수 있는 단어
- IHA_S, IHA_C, IHA_E : 그 자체로 정보타입이 될 수 없지만 각각 정보타입의 시작(S), 중간(C), 끝(E)부분에 위치하여 다른 단어와 결합함으로써 정보타입을 구성할 수 있는 단어
- NoE : 정보타입이 아님

실험을 위해 회의공지에 대하여 자체 수집한 이메일은 1,011 개이고, 모든 문장(총 14,172개)에 대하여 개체명과 개체명간 관계를 태깅하였다. 이 중 비정형 문장으로 회의장소가 포함된 이메일은 116개이며, 문장수로는 233개이다.

실험에는 공통적으로 형태소 태그와 형태소 1단계수준의 태그를 함께 사용하였다. 형태소 분석의 결과는 4단계까지 출력이 되는데, 확률에 따라 여러 가지로 해석될 수 있기 때문에 그 모호성을 줄이고자 형태소 태그의 가장 상위수준인 1단계수준의 태그를 함께 사용하였다.

각 정보타입에 대한 추출결과를 보면, 단일정보타입인 IHA_I 가 가장 낮은 인식률을 보였다. 단어빈도수 자질을 이용했을 때 F1값이 30% 이상으로 오르고, 사전자질과 함께 사용되었을 경우 44.44% 까지 수치를 보였다. 그 다음으로 인식률이 낮은 정보타입이 시작부에 해당하는 IHA_S 인데, F1값이 34.92%부터 최대 54.55%를 넘지 못했다. 현재 주어진 자질에서는 시작부의 경계를 인식하는데 한계가 있음을 보여준다. 반면에 중간부와 끝부분인 IHA_C, IHA_E 에 대한 성능은 상대적으로 다소 높았다. 사전 자질을 적절히 활용했을 때, 두 타입 모두 F1값이 80%대에 진입했음을 볼 수 있다.

8가지 문자타입 자질이 적용되었을 때, 단일 자질에 대한 적용으로 봤을 때 다소 높은 성능향상을 보였지만 사전자질과 함께 쓰였을 때는 오히려 성능의 저하요인으로 관측된다. 특별히 문자타입 자질과 사전 자질이 함께 쓰였을 때 IHA_S 와 IHA_C 의 성능저하가 발생하는데, 이는 문자타입 자질이 장소 정보타입의 시작부 경계 인식에 오히려 혼선을 주기 때문으로 판단된다.

관측되는 또 다른 결과로는, 우편번호DB와 지하철역DB로 구성된 장소사전1에 35,396개의 데이터로 69.96%의 F1값을 보인 반면, 장소사전2~4를 모두 합친 데이터의 수는 13,413개로, 데이터 수만 보았을 때 장소사전1의 약40%의 수준이지만 F1값은 70.23%로 오히려 약간 더 높은 성능을 보이고 있어, 무조건 많은 장소의 이름을 가지고 있는 것보다 장소를 나타내는 어휘와 적절한 지명을 가진 사전 자질이 성능에 중요한 영향을 미침을 확인할 수 있다. 국립국어원에서 발표한 국어 어휘 분류목록[2]에 나오는 집/건물/장사 등의 장소에 대한 다양한 어휘들로 구성된 사전이 시스템의 인식률을 높였다.

3. 결론

본 논문에서는 한국어 비정형 회의 이메일에서 회의장소를 추출하기 위한 자질 선택 방법을 제안하였다. 형태소 태그와 1단계수준 형태소 태그를 기본으로 사용하고, 문자타입과 단어빈도수 자질을 추가적인 자질로 두었으며, 장소 추출을 위하여 제작된 4가지 종류의 사전을 여러 형태로 조합해 실험한 결과, 단어빈도수 자질과 사전2~4 자질을 적용 했을 때, 기본 수준의 분석결과(47.96%) 보다 22.27% 향상된 70.23%의 F1값을 보였다. 단일형태의 정보타입과 시작부에 대한 인식률을 높이기 위한 자질 선택연구가 추후 진행될 예정이다. 회의장소 외에도 회의공지메일에서 추출해야할 중요한 정보들이 있는데, 이들 정보타입에 대한 연구가 추가로 진행될 것이고, 모든 정보타입 추출모듈이 하나로 통합했을 때 발생하는 클래스 분류문제 등에 대해서도 앞으로 해결해야 할 과제이다.

감사의 글

이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0022444)

본 연구는 행정안전부 국가기록원의 지원을 받아 기록물 보존기술 연구개발(R&D) 사업의 일환으로 이루어졌으며, 이에 감사드린다.

참고문헌

[1] 이창기, 황이규, 오효정, 임수종, 허정, 이충희, 김현진, 장명길, Conditional Random Fields를 이용한 세부 분류 개체명 인식, 2006년도 제18회 한글 및 한국어 정보처리 학술대회, pp.268-272, 2006

[2] 국립국어연구원, "최종연구보고서 : 국어 어휘의 분류목록에 대한 연구", 1993