Knowledge based적인 자연언어 분류기법을 이용한 통합적 자연언어 이해 시스템에 대한 연구

이 원 부

동국대학교 정보관리학과

본 연구에서는, 기계에 의한 자연언어 text의 이해(natural language understanding) 증진을 위해 자연언어 분류(natural language classification) 기법과 자연언어 이해 기법과의 통합적 연결성을 중점적으로 연구했다. 현재 및 종전에 개발된 대부분의 natural language understanding 기법들은 자연언어 검색및 분류 방법들과의 연결성을 전혀 고려하지 않고 language understanding 자체의 기법 연구에만 몰두해 왔다.

예를 들면 자연언어 이해과정의 전제 조건으로서, 1)자연언어 text들은 사전적으로 공히 일정 검색주제들에 관해 먼저 분류가 되어 있어야 한다는 것과 사용자들은 2)대부분의 경우 주어진 분류 text의 내용에 따라 사전적으로 선택된 이해기법의 테두리안에서만 수동적으로 자연언어 이해를 시도해야 한다는등의 비현실적인 가정을 설정하고 있다. 따라서 현재 개발중이거나 연구되는 대부분의 자연언어 이해기법들은 이해대상자연언어 text들의 내용및 서술형태의 다양성에 따른 이해과정의 variation을 탄력적으로 그리고 real time적으로 처리할 수가 없다.

따라서 본 연구에서는 상기 언급된 자연언어 기법들의 내재된 문제점을 극복하는 방편으로서, 사전적인 자연언어(text)의 classification 과정과 분류된 text들의 정리 요약을 위한 사후적인 자연언어(text)의 understanding 과정을 real_time적으로 연결시킬수있는 논리적 model을 고찰하여 보았다. 본

연구에서 이 논리적 model은 사용자위주의 자연언어 이해구조로써 PASCAL과 C를 이용하여 프로그램화 되어, AP통신에서 제공하는 자연언어로 된 texts(news story)들을 대상으로 하여 실제적인 자연언어 검색및 이해과정이 실험적으로 평가되었다. 실험의 결과 다음과 같은 주목할만한 결론이 도출되었다.

첫째 기계에의한 자연언어 이해 과정은, 대상 text들의 내용또는 서술형식 그리고 사용어휘(vocabulary)의 정형성(formality)에 영향을 받는다. 이해대상 text들의 내용의 정형성이 높을수록 기계에 의한 자연언어 이해과정은 효과적으로 수행이 될 수 있다.

둘째 자연언어 text의 정형성을 높이기 위해서는, 이해대상 자연언어(text)들이 내용에 따라 사전적으로 정밀하게 구별 되어야 한다. 이를 위하여, 특정주제에 관해 대상 text들의 상대적인 관련성(relevance)를 인지할수 있는 fuzzy theory를 이용한 자연언어 분류 검색방법이 사용될 수있다. Fuzzy 검색및 분류방법은 일정주제에 대한 각 text들의 relevance value에따라 전체 text들을 등급을 매겨 group화 시키게 되므로 궁극적으로 자연언어 분류및 검색의 정밀화를 달성할 수 있게 된다.

셋째 검색및 이해 대상 topic(주제)들의 homogeneity가 궁극적으로 자연언어 text들의 검색 및 이해에 영향을 미친다. 주제의 homogeneity란, 일정주제에 관한 text들에서 사용되어지는 vocabulary들이나 전개 내용 그리고 문장 서술 형태들의 다양성(variation)을 의미한다. 따라서 이해 주제(topic)의 homogeneity가 높을수록, 이해대상 text들의 문장 구성 형식이나 내용이 다양해지므로 fuzzy classifiacation을 이용한 구별적이고 정밀한 text들의 분류가 더욱 중요하게 요구되어진다.

넷째 본 연구에서는 자연언어 분류시, 높은 cut_off value는 비교적 높은 homogeneity를 보여주었다. 즉 fuzzy이론을 이용한 자연언어 분류시에 사용된 높은 θ value는 분류대상 자연언어들의 높은 formality를 보장하게 되어 궁극적으로 자연언어의 기계적 이해의 성과를 높여 주었다.

다섯째 자연언어 이해과정에 있어, 우선 사용자의 목적에 따라 적절한 검

색 및 평가 기준(recall 이나 precision)이 선정되고 이 검색 기준에 따라 먼저 대상 text들이 효과적(recall)으로 또는 효율적(precision)으로 선정이 된 후 사후적으로 적절한 자연언어 기법의 선택및 적용이 권장된다.

결론적으로 보면, 기계에의한 자연언어 분류및 이해 과정을 성공적으로 달성하는 데에 있어서는 사용자들이 자연언어 분류 기법과 이해 기법을 real time적으로 동시연결하여 궁극적인 text들의 정형성 확보가 가장 중요한 전략적 point가 된다. 이를 위해 fuzzy 이론을 이용한 자연언어 분류기법의 사용이 요구되어진다. 또한 fuzzy이론 이외에, 본연구에서는 아직 가설로써만 설정되어 있지만(실험적 뒷받침은 되지 못하고 있지만), text에 사용되는 vocabulary들의 관련 synonym을 고려 한다면 궁극적으로 자연언어 text들의 정형성 향상에 큰 도움이 될것으로 고려된다. 그 이유로써는 자연언어 text의 정형성에 영향을 미치는 요인으로서, 무엇보다도 text의 내용표현을 위해 사용되어지는 vocabulary들이 제일중요하고 그 vocabulary들은 관련된 synonym에 의해 쉽게 정형화가 될수있기 때문이다. synonym을 사용한 vocabulary들의 정형성의 향상은 궁극적으로 전체 text들의 정형성의 향상에 연결이 될 수 있다.

본연구에서 도출된 실험적 결론들은 test sample의 수적인 제약성에 기인된 통계적인 검증의 취약성을 가지고 있다. 앞으로 sample의 증대를 통한 통계적 확인 검증절차가 필요하다. 또한 본연구에서는 자연언어 이해 system의 시스템적인 불완전성으로 인해 pilot test의 선행이 불가피 했으나 향후 시스템의 개발완료에 따른 main test의 실행이 따르게 된다.