한국어 구문 분석기를 이용한 예문기반 유사 영문 선택에 관한 연구

권영훈∪ 윤영호 한광록
호서대학교 벤처전문대학원
zodiac@mail.hoseo.ac.kr
saw05@hanmail.net
krhan@office.hoseo.ac.kr

A study of the selection of similar English sentence based on example using the Korean parser

Young-Hoon Kwon∪ Young-Ho Yoon Kwang-Rok Han
Dept. of GSV, Hoseo University

요 약

본 연구는 예문을 이용하여 한국어 문장과 가장 유사한 영어 문장을 선택하기 위한 기존 연구보다 예문 제시의 정확도를 향상하고 기존의 문제점이었던 문장성분 선택의 불일치성을 제거하기 위해 한국어 구문 분석 시스템을 추가한 형태를 갖추고 있다. 한국어 구문 분석 시스템을 사용하는 이유는 한 문장을 하나의 프레임으로 구조화시킬 때 서술부가 문장의 의미를 나타내는 가장 중요한 역할을 하므로 서술부를 헤더로 선택하고 단순히 조사 정보를 사용하여 각 문장성분을 추출하는 방법의 문제점을 제거하고 서술부 연결 관계를 기초로 프레임의 슬롯을 확보할 수 있기 때문이다.

유사 영문이 필요한 한국어 문장이 입력되면 입력 문장에 대한 형태소 분석과 한국어 구문 분석을 통하여 한국어 문장에서 서술부와 연결되는 주요 성분을 분리하여 프레임 구조를 생성하고 생성된 프레임과 이미 구축된 예문 데이터베이스 사이의 가중치와 유사도를 계산함으로써 한국어 문장과 유사한 영어 문장의 예를 제시하여 영작에 이용할 수 있는 시스템을 구현한다.

1. 서론

급속도로 발전하는 정보화 시대에 발맞춰 국제간 기업 제휴가 빈번해 짐에 따라 번역 업무에 이용할 수 있는 기계번역 시스템의 필요성이 급속도로 증가하고 있다. 현재 국내에서 가장 많이 연구되고 또한 보급되어 있는 기계번역 시스템은 주로 일본어와 영어를 대상으로 하고 있다. 컴퓨터가 다루기 어려운 인간의 언어를 그 처리 대상으로 하고 있다는 점에서 기계번역은 상당히 어려운 연구 분야 중 하나이다. 그러나, 일본어를 대상으로 하는 일한 기계번역 시스템은 양 언어의 문법적 성질과 어순의 유사성으로 인해 현재의 기술 수준으로도 일정한 번역 품질을 확보할 수 있다. 한편, 영어를 그 처리 대상으로 하고 있는 영한기계번역이나 한영기계번역 시스템은 아직 이러한 수준에 도달하지 못하고 있다. 한국어-일본어와는 달리 한국어-영어는 양 언어의 문법적 성질과 문장 구조, 어순 등이 전혀 다르기 때문에 일한 기계번역에서와 같이 비교적 단순한 처리 방법으로는 영한, 한영 번역을 달성할 수 없기 때문이다[2].

현재의 한국어 처리를 위해서는 문장을 형태소 별로 분리하여 각각의 품사를 결정하는 형태소 분석과 태깅 과정, 문장의 문법적 구조를 분석하는 구문 분석 과정, 어의의 중의성을 해결하기 위한 의미해석이 필요하다. 그러나 한국어 처리가 기본 적으로 요구되는 한영, 영한 기계번역에는 한국어 문장과 영어 문장의 문장 구조 차이와 단어 선택 문제를 극복하기 힘든 어려움이 존재한다.

따라서 본 연구는 기존 한영 번역의 문장 구조와 단어 선택의 문제점을 극복하기 위해 한국어 문장을 영어 문장으로 변환하기 용이하도록 예문을 선택하는 방법을 제시한다. 즉, 한국어 문장과 정확히 대응되는 영어 문장을 확보하고, 한국어 문장만을 형태소 분석 및 구문 분석한 후 한국어 문장내 주요 키워드(서술어, 주어, 목적어 등)를 분리하여 저장하고 가중치를 부여하여 기본 데이터베이스를 구축한다. 이후 영어 문장으로 변환하고자 하는 한국어 문장이 입력되면 데이터베이스 구축 과정과 유사한 과정을 거쳐 이미 구축된 데이터베이스에서 가장 유사한 한국어 문장을 선별하여 선별된 한국어 문장에 정확히 대응되는 영어 문장을 제시함으로써 영한 번역 기법과 정보검색 기법을 함께 사용하는 영문 선택 지원 시스템이 될 수 있다.

2. 관련연구

1) 다국어 정보검색을 이용한 한영 문서 작성 지원 시스템[7]

이 논문에서는 기본적인 한국어 처리와 정보검색 기법을 응용하는 다국어 정보 검색에 의한 영문 작성 지원 시스템을 주제로 한다. 한국어 형태소 분석, 자동색인, 영한 사전, 시소러스, 정보 검색 기술, 예문 데이터베이스 등을 활용함으로써 입력된 한국어 문장과 유사도가 가장 큰 영어 예문의 패턴을 검색해 내는 시스템으로 기존의 기계 번역 과정과 같은 자연언어 처리 과정과 상당히 상이한 점을 특징으로 들 수 있다. 이 시스템은 한국어 문장을 분석하여 키워드를 추출한 후 한영사전을 사용하여 각 키워드에 해당하는 영어 단어를 선택하게 된다. 즉, 사용자가 선택하는 키워드에 따라 예문을 제시하는 정보 검색 기법이 많은 부분 적용되었고 영어 문장을 중심으로 색인되기 때문에 한국어 문장의 수많은 표현들이 적용되기 어려운 문제점이 있다.

2) 예제 기반 단어 선택[1]

이 논문은 일본어 예문을 기반으로 하여 영일 번역에 사용할 수 있는 시스템에 대한 연구이다. 번역시에 가중치를 높게 두어야 할 단어를 선택하는 방법에 초점이 맞춰져 있다. 문장에서 의미를 결정하는 가장 중요한 성분(서술어)을 추출한 후 다른 문장 성분을 분리하는 방법으로 서술어의 일치가 예문 선택에 있어서의 핵심이 되는 연구이다. 기계 번역에 있어서 영어문장과 일어 문장 구조의 차이는 상당히 중요한 부분임에도 불구하고 구조 자체보다는 단어의 의미 일치만을 고려했다는 단점이 있다.

3) 예문을 이용한 유사 영문 선택

이 논문은 앞에 설명한 "다국어 정보검색을 이용한 한영 문서 작성 지원 시스템"과 "예제 기반 단어 선택"의 문제점을 절충하여 제시된 논문으로, 한영 문서 작성 지원 시스템에서는 한국어 표현에서의 융통성을 유지할 수 있도록 한국어 문장만을 색인하며 여러 가지 명사의 결합 및 동사(형용사)의 결합을 고려했다. 색인 시에 선택될 키워드 추출 방법은 "예제 기반 단어 선택"에서 제시한 방법을 채택하고 있으나 조사 정보와 패턴 정보를 사용하기 때문에 문장의 구조도 고려한 논문이라 할 수 있다. 그러나 단문의 경우에는 간단하게 분석할 수 있겠지만 중문이나 복문의 경우 키워드 선택의 문제점은 또다시 나타나게 된다.

본 논문은 "예문을 이용한 유사 영문 선택에 관한 연구"의 확장으로 키워드를 선택하기 위한 방법으로 한국어 구문 분석기(chart parsing)를 사용하게 된다. 한국어 구문 분석기를 사용할 경우의 장점은 단문만이 아닌 중문을 분석하더라도 문장의 포함 관계가 명확해질 수 있기 때문에 더욱 효과적인 키워드 선택을 할 수 있게 된다[5].

3. 설계 및 구현

1) 한국어 문장 분석 및 데이터베이스 구축

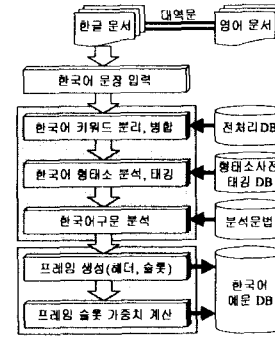

그림 3.1 데이터베이스 구축 구성도

그림 3.1은 한국어 문장을 분석해서 데이터베이스를 구축하는 시스템 모듈별 구성도로 과정은 다음과 같다. 한글문서와 동일한 내용의 영어문서를 수집하여 한글문서의 문장을 입력으로 한다. 이후 전처리 데이터베이스를 참조하여 어절을 분리하거나 병합하여 형태소 분석 및 태깅을 실행하게 되면 복합어 및 동사(형용사)+보조동사(보조형용사)의 형태, 접미사에 의한 품사 전성 등을 대부분 처리해 줄 수 있다[3][4]. 이 과정을 마친 후에 구문분석기의 입력의 형태로 가공하여 구문분석 모듈을 통과하게 된다[6][9]. 구문분석의 예문 결과는 다음과 같다.

-예문 : 나는 휴대전화에 관한 자료를 받고 싶다.

```
(SUBJ NP(baseform "나")(repform "는")(reptag "JO")...+topic)
(VV
  (OBJE ALLN
    (ANCL VV
      (UNKN NN(baseform "휴대전화")(repform "에")(reptag "JO")...)
      (VV(baseform "관하")(repform "ㄴ")(reptag "EM")...+relzer))
    (NN(baseform "자료")(repform "를")(reptag "JO")...))
  (VV
    (VV(baseform "받")(repform "고")(reptag "EM")...+connec)
    (AX(puncform".")(baseform"싶")(repform "다")(reptag "EM")...)))
```

위 예문의 결과는 올바르게 파싱된 결과로 "나는"이 주어, "자료를"이 목적어, "받고 싶다"가 서술어로서 명확한 분리가 가능하다. 파싱 결과에서 나타나는 구조적 중의성은 예문 제시에도 재사용하기 위해 데이터베이스에 모두 저장한다. 다음 과정으로 프레임 생성은 "받고 싶다"라는 서술어를 채택하고 슬롯(1.주어, 2.목적어, 3.보어, 4.부사격 후치사구, 5.불확실한 체언, 6.연속된 동사)에는 "나는"과 "자료를"이 각각 채택된다. 하나의 한국어 문장이 이와 같이 한 프레임을 형성하며 데이터베이스에 기록된다. 슬롯의 가중치는 구축된 데이터베이스 내에 존재하는 동일한 필드값의 통계로 계산된다.

2) 한국어 문장과 유사한 영어 예문 제시

그림 3.2는 한국어 문장이 입력 후 영어 예문이 제시되기까지의 전체 시스템 구성도로 그림 3.1과 프레

임 생성 단계까지는 동일하나 이후 단계에서만 차이가 있다. 입력된 한국어 문장에 대한 프레임 생성이 완료된 후 벡터 유사도 계산을 통해 가장 유사도가 높은 프레임을 선택하여 영어 문서를 검색, 영어 예문을 제시하게 된다[8].

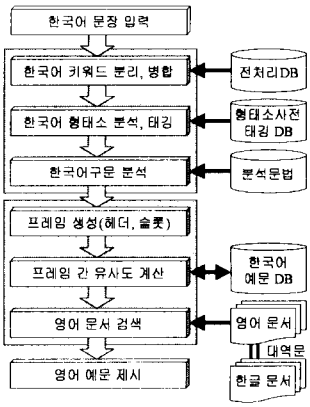

그림 3.2 영어 예문 제시 시스템 구성도

3) 구현된 시스템의 인터페이스

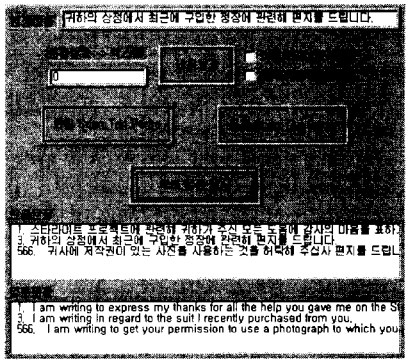

그림 3.3 구현된 시스템의 결과 화면

위 그림 3.3은 구현된 시스템의 간략화된 결과 인터페이스이다. 데이터베이스 구축 시스템과 영문 제시 시스템을 결합한 형태로 파일에서 한국어 문장을 입력받거나 화면상에서 문장을 입력받아 데이터베이스를 구축할 수 있다. 현재는 파일 단위의 영어 예문 선택은 불가능하고 문장 입력을 통해 영어 예문 제시가 가능하다. 이유는 영작 시 필요한 정상적인 영문 문법 구조를 제시하는데 목적을 두기 때문이다.

4. 결과 및 결론

1) 실험 결과

실험은 3950개의 한국어 문장과 정확히 번역된 영어 문장을 사용하였다. 3950개의 문장 중 정확히 구문 분석이 된 문장은 2966개로 약 75%의 분석 정확도를 보였다. 차트 파싱 알고리즘을 사용하고 있기 때문에 약간의 분석 오류가 나타날지라도 분석 후보에 등록될 수 있다. 구문 분석시의 오류를 포함한 분석 결과는 3950개의 문장 중 3795개로 약 96%였다. 오류를 포함한 결과를 제시한 이유는 유사한 한글 문장일 경우 구문 분석 결과가 같아질 수 있기 때문이다. 예문으로 채택된 한국어 문장이 모두 분석 문법 내에 포함되어 있지 않고, 완성되지 않은 문장도 예문에 포함했기 때문에 분석율이 낮게 나타났다. 영어 예문 제시 결과는 다음과 같다. 데이터베이스에 포함되지 않은 400개의 한글 문장을 실험한 결과 평균 제시 예문 수는 1.32개로 예문 제시에 실패하는 경우는 데이터베이스에 헤더와 슬롯이 구축되어 있지 않은 경우였다.

2) 결론

실험 결과를 살펴보면 예문 제시가 실패하는 경우는 대부분 유사한 예문이 프레임 형태로 구축되지 않은 경우였다. 데이터베이스에 방대한 양의 데이터가 구축된다면 제시되는 영어 예문의 수는 증가하게 된다. 한국어 구문 분석 시스템의 정확도를 향상시키고 방대한 데이터베이스를 효과적으로 관리하며 시스템의 전반적인 속도를 증가하게 되면 영작을 필요로 하는 많은 사람들에게 본 연구에서 구현한 유사 영문 선택 시스템의 활용도가 상당히 높아 질 것이다. 향후 연구 과제로는 서술어 부분의 다양성을 유연하게 대처하고, 현재의 일차원적인 프레임 구조를 다차원화하여 복잡한 문장도 세밀하게 분석할 수 있는 구문 분석 시스템으로 개선해야 한다.

5. 참고 문헌

[1]. Satoshi Sato, "Example-Based Word Selection", 日本 人工知能學會 Vol.6, No.4, pp.592-600, 1990.

[2]. 박철제, 김태완, "기계번역 시스템", 정보처리학회지, Vol.5, No.5, pp.29-36, 1998.

[3]. 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 대학원 컴퓨터공학과 박사 학위 논문, 1993.

[4]. 강승식, 권혁일, 김동렬, "한국어 자동 색인을 위한 형태소 분석 기능", 한국 정보과학회 학술발표논문집, 제22권, 1호, pp.930-932, 1995.

[5]. Makoto Nagao저, 황도삼 외 3인 공역, "Natural Language Processing", 홍릉과학출판사, 1998.

[6]. 서영훈, 이하규 외, "한국어 구문 Tagged Corpus 구축 및 구문 분석 데이터 사전 개발", 한국 전자 통신 연구소 최종 연구 보고서, 1998.

[7]. 황미자, "다국어 정보검색을 이용한 한영 문서 작성 지원 시스템", 호서대학교 대학원 컴퓨터공학과 석사 학위 논문, 1998.

[8]. William B. Frakes, Ricardo Baeza-Yates, "Information Retrieval", Prentice Hall, 1992.

[9]. 양재형, "KONAN Korean Grammar Rules", 1998.