ISSN(Print): 2287-5905

http://dx.doi.org/10.3745/KTSDE.2013.2.1.035

English-Korean Transfer Dictionary Extension Tool in English-Korean Machine Translation System

Sung-Dong Kim†

ABSTRACT

Developing English-Korean machine translation system requires the construction of information about the languages, and the amount of information in English-Korean transfer dictionary is especially critical to the translation quality. Newly created words are out-of-vocabulary words and they appear as they are in the translated sentence, which decreases the translation quality. Also, compound nouns make lexical and syntactic analysis complex and it is difficult to accurately translate compound nouns due to the lack of information in the transfer dictionary. In order to improve the translation quality of English-Korean machine translation, we must continuously expand the information of the English-Korean transfer dictionary by collecting the out-of-vocabulary words and the compound nouns frequently used. This paper proposes a method for expanding of the transfer dictionary, which consists of constructing corpus from internet newspapers, extracting the words which are not in the existing dictionary and the frequently used compound nouns, attaching meaning to the extracted words, and integrating with the transfer dictionary. We also develop the tool supporting the expansion of the transfer dictionary. The expansion of the dictionary information is critical to improving the machine translation system but requires much human efforts. The developed tool can be useful for continuously expanding the transfer dictionary, and so it is expected to contribute to enhancing the translation quality.

Keywords : English-Korean Machine Translation, English-Korean Transfer Dictionary, Out-Of-Vocabulary, Compound Noun

영한 기계번역 시스템의 영한 변환사전 확장 도구

김 성 동†

요 약

영한 기계번역 시스템을 개발하기 위해서는 언어에 대한 다양한 정보를 필요로 하며, 특히 영어 단어에 대한 의미 정보를 포함하는 영한 변환사전의 풍부한 정보량은 번역품질에 중요한 요소이다. 지속적으로 생성되는 새로운 단어들은 사전에 등록되어 있지 않아 번역문에 영어 단어가 그대로 출력되어 번역품질을 저하시킨다. 또한 복합명사는 어휘분석, 구문분석을 복잡하게 하고 사전에 의미가 등록되지 않은 경우가 많아 올바르게 번역하기 어렵다. 따라서 영한 기계번역의 번역품질 향상을 위해서는 사전에 등록되어 있지 않은 단어들과 자주 사용되는 복합명사들을 수집하고 의미 정보를 추가하여 영한 변환사전을 지속적으로 확장하는 것이 필요하다. 본 논문에서는 인터넷 신문기사로부터 말뭉치를 추출하고, 사전 미등록 단어와 자주 나타나는 복합명사를 찾은 후, 이들에 대해 의미를 부착하여 영한 변환사전에 추가하는 일련의 과정으로 구성되는 영한 변환사전의 확장 방안을 제안하고 이를 지원하는 도구를 개발하였다. 사전 정보의 확대는 많은 사람의 노력을 필요로 하는 일이지만, 영한 기계번역 시스템의 개선을 위해서는 필수적이다. 본 논문에서 개발한 도구는 사람의 노력을 최소화 하면서, 영한 변환사전의 정보량 지속적인 확대를 위해 유용하게 활용되어 영한 기계번역 시스템의 번역품질 개선에 기여할 것으로 기대된다.

키워드 : 영한 기계번역, 영한 변환사전, 사전 미등록어, 복합명사

1. 서 론

영한 기계번역 시스템을 개발하기 위해서는 영어와 한국어에 대한 언어 정보(영어 어휘규칙 정보, 영어 구문구조 정보, 영한 변환 정보, 한국어 생성 정보)와 다양한 사전 정보(영어 어휘사전, 영한 변환사전, 한국어 생성사전)를 구축하고 이들 정보를 활용하여 분석과 생성을 수행하는 알고리즘을 개발해야 한다. 영한 기계번역 시스템의 성능은 구축한 정보량과 이를 활용하는 효과적이고 효율적인 알고리즘과 밀접한 관계가 있다. 본 논문에서는 영한 기계번역 시스템의 번역품질 유지와 개선을 지원할 수 있도록 영한 변환사

※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2010-0010815).
† 종신회원 : 한성대학교 컴퓨터공학과 교수
논문접수 : 2012년 7월 6일
수 정 일 : 1차 2012년 9월 3일
심사완료 : 2012년 9월 18일
* Corresponding Author : Sung-Dong Kim(sdkim@hansung.ac.kr)

전의 정보량을 지속적으로 확장하는 방안을 제시한다.

영어 문장을 번역할 때, 영한 변환사전에 없는 단어는 영어 단어가 그대로 번역문에 나타나며, 이러한 결과에 대해서 사용자는 낮은 평가를 하게 된다. 그런데, 신조어가 빈번하게 나타나고 있어 신조어들에 대한 정보를 빠르게 영한 변환사전에 추가해야 영한 기계번역의 번역품질을 유지할 수 있다. 그리고 영한 기계번역에서 복합명사는 하나의 단위로 분석하여 번역함으로써 보다 정확하고 자연스러운 번역을 할 수 있다. 숙어기반 영한 기계번역에서 복합명사는 영한 변환사전에 숙어 등록을 하고 숙어인식을 통해 번역을 하고 있다. 복합명사도 지속적으로 새로 나타나고 있으며 사전에 없는 복합명사들은 숙어인식이 되지 않아 하나의 단위로 처리되지 못하여 분석을 복잡하게 하고 영어 원문이 그대로 출력되는 등의 문제를 유발한다. 따라서 구문분석 이전에 복합명사를 미리 찾아, 하나의 단위로 묶어서 분석을 수행해야 하며 복합명사에 대한 대역어를 지정하여 번역을 수행해야 복합명사를 포함한 문장에 대해 보다 좋은 번역을 생성할 수 있다.

본 논문에서는 인터넷 신문기사로부터 영어 말뭉치를 구축하고 말뭉치로부터 영한 변환사전에 등록되지 않은 단어들과 자주 사용되는 복합명사를 수집하여 이들에 대한 의미를 부착한 후 기존 영한 변환사전에 통합하는 영한 변환사전 확장 방안을 제안한다. 신조어의 수집을 위해 매일매일의 신문기사로부터 말뭉치를 구축하며, 말뭉치로부터 기존 영한 변환 사전에 없는 단어를 추출한다. 복합명사의 추출을 위해서는 복합명사를 구성할 수 없는 불용어를 정의하고, 문서에 나타나는 빈도수를 기준으로 복합명사를 확인하도록 한다. 즉 규칙에 의해 불용어를 확인하고 빈도수에 의해 복합명사로 판단하는, 규칙과 통계적인 방법을 혼용한 수집 방법을 적용하였다. 추출된 미등록 단어와 복합명사에 대해서 대역어를 입력하는 작업은 전문가에 의해 수행되고 결과를 영한 변환사전에 통합하여 영한 기계번역 시스템에서 사용할 수 있도록 한다. 본 논문에서는 말뭉치 수집, 사전 미등록 단어와 복합명사의 추출, 의미 부착, 영한 변환사전과의 통합 등의 일련의 과정을 지원하는 도구를 개발하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 기계번역 시스템과 변환사전의 관계 및 변환사전 구축에 관한 이전의 연구와 복합명사 인식을 위한 기존 연구들을 검토한다. 3장에서는 본 논문에서 제안하는 영한 변환사전의 확장 방안을 설명하고 4장에서는 일정 기간 동안의 사전 미등록어와 복합명사 수집결과를 제시한다. 그리고 5장에서 앞으로의 과제를 제시하며 논문을 마무리한다.

2. 관련 연구

[1]에서는 사용자 번역 사전 구축을 통해 번역품질을 개선할 수 있음을 보였다. 즉 사용자 사전 정보가 기존 번역 사전 정보와 문법 규칙을 대치할 수 있도록 하여, 사용자 사전을 통한 번역품질 개선에 대한 결과를 제시하였다. 이는 번역 시스템이 포함하고 있는 번역사전 이외에 독립적인 사전을 추가하여 번역품질을 개선할 수 있음을 보여준다. 그러나 독립적으로 운영되는 사전을 유지하는 경우, 정보의 일관성과 우선순위 등의 문제가 있을 수 있다. [2]에서는 기계번역에서 사전의 중요성을 언급하며 내용과 조직 면에서 사전이 갖추어야 할 요건을 제시하였다. 사전의 내용은 양적, 질적으로 충분하여 대역어를 선택할 수 있을 정도로 많은 단어를 포함해야 한다고 하였다. 그리고 사전 조직은 번역 시스템 설계에 영향을 미칠 수 있으므로 효율적으로 단어를 저장하여 빠른 정보 접근을 지원해야 한다고 하였다. 이 연구는 1980년대 말에 수행되었는데, 현재는 컴퓨터 하드웨어의 발전으로 사전 조직에 의한 접근속도 향상은 크게 의미 없으며 가능한 많은 단어를 포함하도록 하는 것이 번역품질을 개선하는 데 중요하다고 할 수 있다.

영한 또는 한영 기계번역 시스템을 개발하는 과정에서 변환사전 구조와 생성에 대한 연구가 있었다. [3]에서는 말뭉치로부터 연어와 숙어 정보를 추출하여 영한 변환사전을 구축하는 방법을 제시하였다. [4]와 [5]에서는 한영 기계번역을 위한 사전의 구조와 사전 생성 방법에 대한 연구 결과를 제시하였다. 또한 [6]에서는 변환방식의 기계번역 시스템을 위한 변환사전의 생성방법에 대한 특허를 제안하기도 하는 등 사전의 구조와 생성방법에 대한 연구가 있었다. 그런데 이러한 연구는 기계번역 시스템을 개발할 때, 초기 사전의 구축에 대한 연구로서 기존에 존재하는 정보를 기계번역 시스템이 활용할 수 있도록 전자화하는 과정에 대한 방법을 제시한 연구들이다.

복합어(compound words)는 두 개 이상의 단어로 구성되는 하나의 어휘 단위(lexical unit)인데,명사 복합어, 형용사 복합어, 부사 복합어, 전치사 복합어 등 복합어는 영어의 모든 품사에 해당하는 단위를 형성할 수 있다[7]. 본 논문에서는 그 중 명사 복합어, 즉 복합명사를 말뭉치로부터 추출하여 변환사전에 등록함으로써 영한 기계번역의 번역품질 개선을 도모하고자 한다. 영어에서 복합명사 인식은 명사구 추출(noun phrase extraction) 문제의 부분인데, 명사구 추출은 문장에 존재하는 모든 명사구를 인식하는 문제이며 정보검색(information retrieval)의 성능 개선을 위해 유용하게 이용될 수 있다[8]. 복합명사에 관한 연구는 개체명 인식(named-entity recognition) 방법을 적용하여 복합명사를 인식하는 방법들과 통계적 방법을 적용하여 말뭉치로부터 복합명사를 추출하는 방법, 그리고 복합명사의 대역어를 찾아내는 방법들에 관한 연구가 있었다. 은닉 마코프 모델(Hidden Markov Model)[9], CRF(conditional Random Fields)[10] 등의 기계학습 방법을 적용하여 개체명을 인식하는 연구가 있었으며, 개체명은 일반 명사와는 달리 몇몇 기사에 동시에 나타난다는 현상을 이용하여 드물게 나타나는 개체명을 확인하는 방법을 제시한 연구도 있었다[11]. [12]에서는 여러 가지 동시출현(co-occurrence) 척도와 다양한 어휘적 단서를 이용하여 복수단어 표현(multi-word expression)을 추출하는 방안을 제시하였다. 기계번역은 복

합어 인식 뿐만 아니라 이를 적절하게 번역해야 하는데, 이를 위해 복합어의 대역어를 자동으로 설정할 필요가 있다. [13]에서는 웹과 말뭉치로부터 자기-학습(self-learning) 학습 방법으로 복합어의 대역어를 획득하는 방안을 제시하였다. 여기서 제안한 방법은 언어 쌍에 독립적이므로 영어-한국어 간의 복합어의 대역어를 얻는 데에도 활용할 수 있을 것으로 판단된다.

본 논문에서는 기 구축된 변환사전의 정보량을 지속적으로 확장하여 기계번역 시스템의 지속적 개선을 지원하기 위한 방법론에 초점을 맞추었다. [14]에서는 변환사전의 정보량 확대를 보다 용이하게 할 수 있도록 변환사전 관리 도구를 개발하였다. 그런데 이 도구는 사전 전문가가 번역품질의 개선을 위해 개별적인 단어를 추가하는 작업을 지원하는 것으로서 대량의 정보를 추가하는 기능은 지원하지 않는다. 본 논문에서는 대량의 정보를 사람의 노력을 최소화하면서 변환사전에 추가하는 방법론에 대해서 연구하였다. 즉, 인터넷 신문기사로부터 영어 말뭉치를 구축하고, 이로부터 사전 미등록 단어와 자주 나타나는 복합명사를 추출하고, 여기에 의미를 부착하여 변환사전에 추가하는 과정으로 구성되는 방안을 제시하고 이를 지원하는 도구를 개발하였다.

3. 영한 변환사전의 확장 방안

본 논문에서는 영한 기계번역 시스템의 번역품질 유지와 개선을 지원하기 위해 영한 변환사전의 정보량을 지속적으로 확장하는 방안을 제안한다. 사전 미등록 단어와 자주 사용되는 복합명사를 지속적으로 수집하여 추가하는 방식으로 정보량을 확장하려고 한다.

Fig. 1은 영한 변환사전의 확장 과정을 보여준다. 첫 번째 단계에서는 데일리중앙[1], 코리아헤럴드[2], 매일경제 영문판[3] 등 3 가지 인터넷 영어 신문 사이트에서 신문기사를 추출하여 영어 말뭉치를 구축한다. 두 번째 단계에서는 구축한 말뭉치에서 기존 영한 변환사전에 등록되어 있지 않은 단어를 추출하고 자주 나타나는 복합명사를 수집한다. 세 번째 단계에서는 추출한 단어와 복합명사의 의미정보를 부착한다. 그리고 네 번째 단계에서는 기존 영한 변환사전에 새로 추출된 단어들을 추가한다. 첫 단계에서는 경제와 과학기술 분야의 신문기사들을 추출하여 말뭉치를 구축하며, 둘째 단계에서는 변환사전과의 통합을 용이하게 하기 위해 첫 글자의 알파벳 순서에 따라 정렬하여 단어를 정리한다. 세 번째 단계만이 전문가의 노력을 필요로 한다.

위에서 기술한 4 단계 과정을 지원하기 위한 도구를 개발하였으며 Fig. 1에서 점선으로 표시한 부분들이 이에 해당한다. 이 중 의미 부착 도구를 이용할 때 사람의 노력이 주로 소요되며, 다른 도구는 입력만 준비하면 손쉽게 도구를 이용하여 결과물을 생성할 수 있다.

1) http://koreajoongangdaily.joinsmsn.com
2) http://www.koreaherald.com
3) http://news.mk.co.kr/english

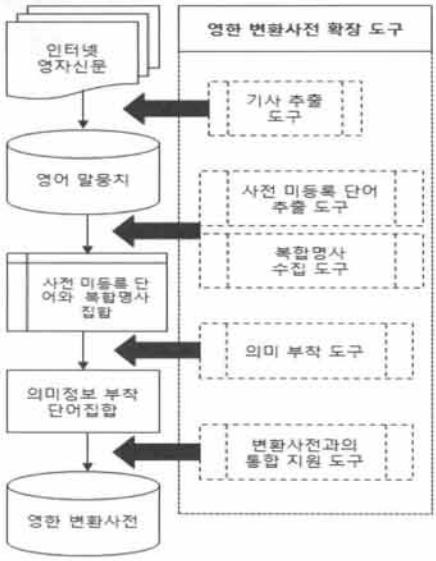

Fig. 1. Processes of English-Korean Transfer Dictionary Extension

3.1 신문기사 추출 도구

인터넷 신문기사 추출 도구는 신문기사 목록을 포함하는 페이지로부터 각 신문기사의 페이지 주소를 추출하는 URLExtractor, 신문기사 페이지 주소를 입력받아 신문기사의 문장을 추출하는 ArticleExtractor, 추출한 신문기사 파일에 포함된 문장들을 한 라인 당 한 문장씩 정렬하는 DocumentArranger 등의 모듈로 구성된다.

사용자는 인터넷 신문 사이트로 이동하여 수집하려는 영역을 지정하거나 영역의 기사목록 페이지의 주소(url)를 기사 추출 도구에 입력한다. 그러면 URLExtractor, Article Extractor, DocumentArranger 등이 차례로 동작하여 기사목록 페이지에 나타난 모든 신문기사들에 포함된 문장을 포함하는 하나의 파일이 생성된다. Fig. 2는 인터넷 신문기사의 추출 과정을 보여준다. 그림에서 점선으로 구분된 부분이 기사 추출 도구 이다.

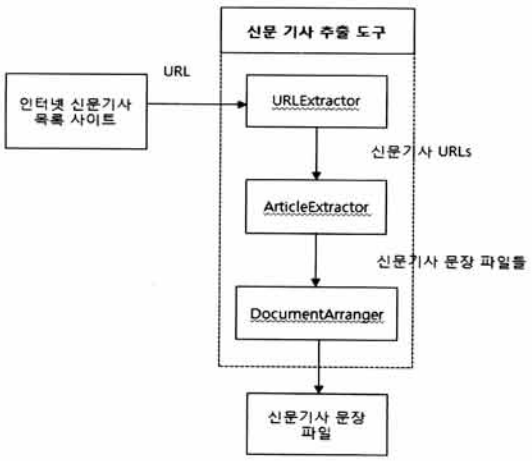

Fig. 2. Processes of Internet News Article Extraction

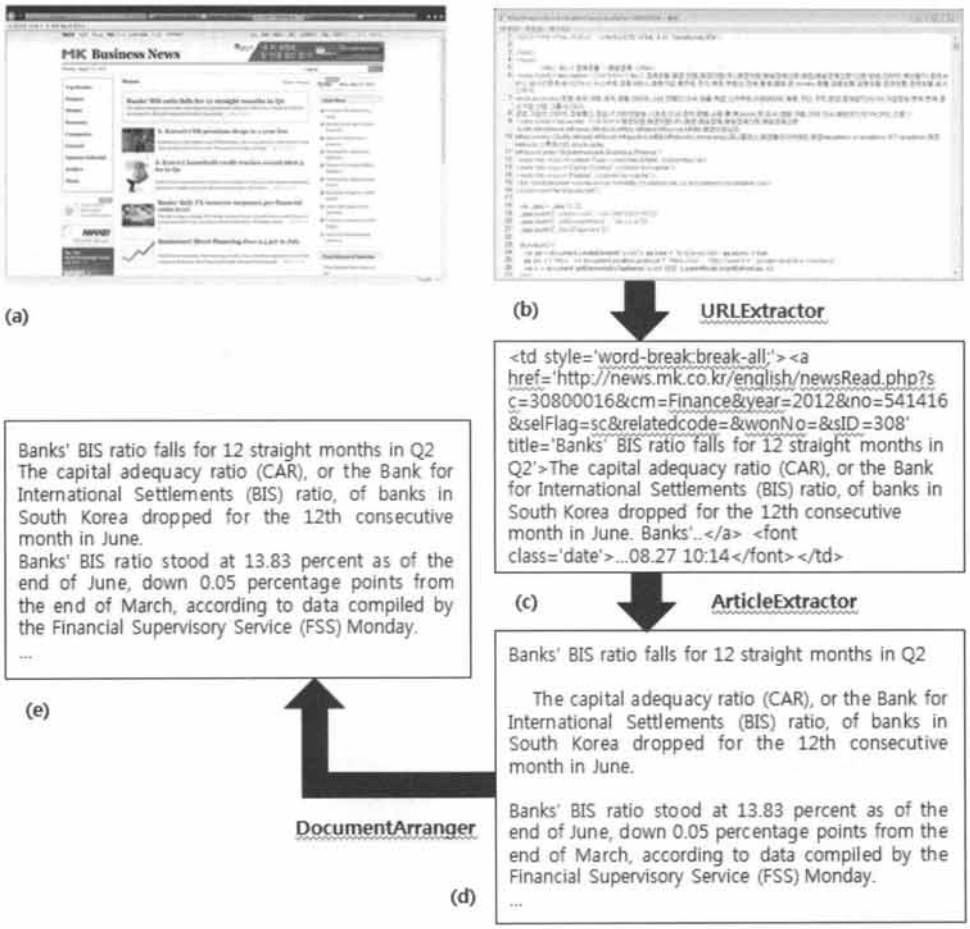

Fig. 3. Example of Article Extraction Processes ((a) MK English News Article List Page, (b) HTML Document, (c) URL Extraction Result File)

Fig. 3은 매일경제 영어 신문에서 기사를 추출하는 과정을 보여준다. 여기서는 "Finance" 영역의 기사를 수집하기 위해 "Finance" 영역을 입력한다. 그러면 URLExtractor가 기사목록을 포함하는 페이지(Fig. 3(a))의 html 문서(Fig. 3(b))에서 기사들의 주소를 추출한다(Fig. 3(c)). 다음으로 ArticleExtractor가 기사의 주소를 얻어서 해당 주소에 있는 기사를 추출하여 파일에 저장한다. ArticleExtractor는 기사들의 개수만큼 반복적으로 수행되어 모든 기사에 있는 문장을 추출하여 하나의 파일에 저장한다(Fig. 3(d)). 이 때 생성된 파일은 신문 기사에 포함된 문장을 포함하고 있으나, 한 라인에 여러 문장이 포함되어 있을 수 있다. 따라서 DocumentArranger에서 한 라인에 한 문장씩을 포함하도록 문서를 정리하고 그 결과는 (Fig. 3(e))와 같다.

3.2 사전 미등록 단어 추출 도구

영한 변환사전에 등록된 단어 목록을 만들고 이를 이용하여 사전 등록 여부를 판단하고 첫 단계에서 생성한 영어 말뭉치로부터 사전 미등록 단어를 추출한다. 영한 변환사전은 명사, 동사, 형용사, 부사, 대명사, 전치사, 접속사 등 7개의 사전으로 구성되는데, 품사별 사전 중 어느 곳에도 등록되어 있지 않은 단어를 말뭉치로부터 추출하여 첫 글자의 알파벳 순서로 정렬하여 파일에 출력한다. 단어의 의미 부착 과정을 지원하기 위해 미등록 단어의 문맥[4]을 함께 저장한다. 또한 미등록 단어의 변환사전 추가 여부를 결정하기 위해 단어의 빈도수를 함께 저장한다.

Fig. 4는 Fig. 3(e)로부터 추출한 사전 미등록 단어들의 일부를 보여주는데, 대문자로 시작하는 단어와 소문자로 시작하는 단어를 구분하였다. 단어의 문맥은 단어 이름의 파일에 별도로 저장된다. Fig. 4(a)는 대문자로 시작하는 사전 미등록 단어들이고 Fig. 4(b)는 "Asan.txt" 파일에 저장된 "Asan"이라는 단어가 나타난 문장이다. 마찬가지로 Fig. 4(c)와 Fig. 4(d)는 소문자로 시작하는 단어들과 "anti-virus"가 나타난 문장을 포함하는 "anti-virus.txt" 파일을 보여준다.

3.3 복합명사 수집 도구

복합명사의 수집은 복합명사를 구성할 수 없는 불용어의 판단과 문서에 나타나는 빈도수를 이용하여 복합명사를 추

4) 여기서는 단어가 사용된 문장

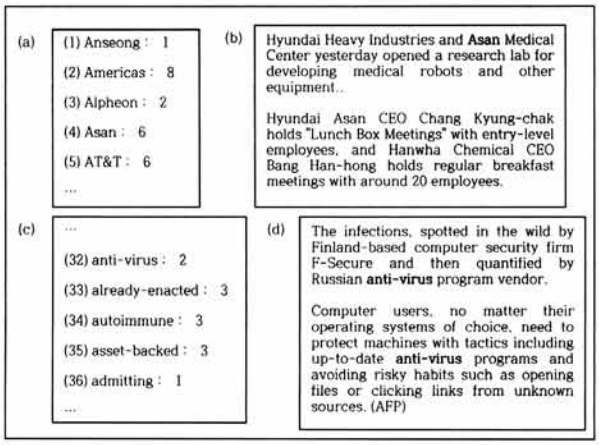

Fig. 4. Example of Un-registered Word Extraction ((a) Un-registered Words Beginning with Capital Letter, (b) "Asan.txt" File, (c) Un-registered Words Beginning with Lower Case Letter, (d) "anti-virus.txt" File)

출하는 방법을 적용한다. 즉, 불용어를 판단하기 위한 규칙을 정의하고 불용어를 제외한 연속적인 단어들을 결합하여 복합명사 후보를 추출하고 추출된 복합명사 후보들이 문서에서 나타나는 빈도수를 기준으로 복합명사로 판단한다. 간단한 방법이지만 자주 나타나는 복합명사를 수집함으로써 영한 기계번역 시스템의 번역품질 개선에 기여할 수 있다.

Fig. 5는 문서에서 복합명사 후보를 추출하기 위한 알고리즘을 보여준다. 복합명사 후보 추출 후에 문서에서 나타나는 빈도수가 일정한 기준 이상일 경우 복합명사로 판정한다. 알고리즘에서 *Document*는 복합명사 후보를 추출할 문서이고, *n*은 복합명사의 길이(2이면 두 단어 복합명사, 3이면 세 단어 복합명사), *CNTable*은 알고리즘의 수행결과 생성되는 복합명사 후보 테이블로서 복합명사와 빈도수를 저장한다. 알고리즘은 다음과 같이 동작한다. 주어진 문서(*Document*)의 각 문장을 단어 리스트(*word_list*)로 변환한다. 단어 리스트의 첫 단어부터 불용어가 아닌 경우에 인자로 주어진 복합명사 길이(*n*) 만큼의 불용어가 아닌 연속적 단어를 결합하여 복합명사 후보(*candidate_CN*)를 생성한다. 생성된 복합명사 후보가 *CNTable*에 등록되지 않은 것이라면 *CNTable*에 추가하고, 이미 등록된 것이라면 빈도수를 증가시킨다. 이 빈도수가 후에 복합명사로 판정하는 기준이 된다.

Fig. 6은 알고리즘의 입력과 출력의 예를 보여준다. 여기서 복합명사의 길이는 2로 하였다. 주어진 문서(Fig. 6(a))로부터 2 단어로 구성된 복합명사들의 테이블을 만들고 그 내용을 (Fig. 6(b))와 같이 출력한다. 불용어의 확인을 위해서 영어 어휘분석기를 이용하여 다음과 같은 정보를 확인한다: 품사, "~ly"로 끝나는 형용사 여부, "~ing"로 끝나는 형용사 여부, 비교급-최상급 형용사 여부 등.

Fig. 5의 알고리즘에서 가장 중요한 부분은 복합명사를 구성할 수 없는 불용어를 확인하는 것(checkStopWord())이며, Table 1에서 불용어 확인을 위한 9 가지 기준을 정의하여 제시한다. 반례들은 나타날 때마다 특별하게 처리할 수

```
Extract_CompoundNounCandidates(Document, n, CNTable)
{
    foreachaSentence in the Document
    {
        Convert aSentenceinto word_list;
        foreachword_w in word_list
        {
            if (checkStopWord(word_w) == true )
            {
                candidate_CN= combine_with_next_words(n);
                CNTable_entry = search candidate_CN in CNTable;
                if (CNTable_entrydoesn't exist)
                {
            make new CNTable_entry for candidate_CN;
                    addCNTable_entryinto CNTable;
                }
                else increment CNTable_entry'sfrequency;
            }
        }
    }
}
```

Fig. 5. Algorithm for Extracting Compound Noun Candidates

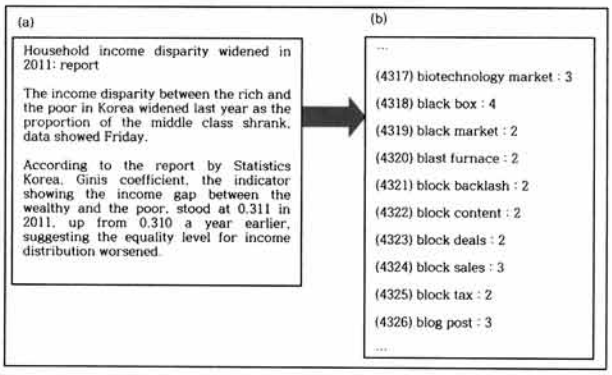

Fig. 6. Example of Input/Output ((a) Input, (b) Output)

있도록 한다. 예를 들어, worker, teacher 등의 단어는 품사를 추가로 확인하고, 4-year, 900-student 에서와 같이 하이픈을 포함하는 단어라도 숫자가 하이픈으로 다른 단어와 연결되는 경우를 확인하여 불용어 여부를 판단한다.

3.4 의미 부착 도구

의미 부착 도구는 단어의 의미를 사전에서 검색하여 의미를 부착하고, 그 품사를 지정하는 작업을 사람이 보다 쉽게 할 수 있도록 지원한다. 변환사전에 등록하더라도 명사 이외의 품사인 경우에는 어휘 사전에 품사정보가 등록되지 않으면 변환사전 검색이 수행되지 않으므로 어휘 사전에 품사정보를 등록해야 한다. 복합명사의 경우에는 해당 복합명사가 사용된 문맥을 확인하여 의미를 부착한다.

Fig. 7은 의미 부착 도구의 사용자 인터페이스의 모습을 보여준다. 의미를 선정할 단어는 [English Word] 부분에 나타나고 [Sample Sentences] 부분에서 예문을 참고하여 의미를 결정하여 [Meaning] 부분에 입력하고 단어의 적절한 품

Table 1. Criteria for Stop Word Identification

Criteria	Is Stop-word?	Examples
Verb, Adverb, Preposition, Article, Conjunction, Pronoun	Yes	
Word including hyphen(-)	No	Counter examples: 4-year, 900-student, …
Word including apostrphe(')	Yes	I'm, Mary's, …
Adjective, Adverb word ending with 'ly'	Yes	beautifully, stringently, …
Word ending with 'ing'	Yes	aborning, absorbing, …
Comparative, superlative adjective word ending with 'er' or 'est'	Yes	former, later, modest, … counter examples: worker, teacher, keeper, …
WH-words	Yes	when, where, who, …
Successive adjective words or adverb+adjective	Yes	very good, much more, …
Word beginning with non-alphabet letter[5]	Number: Yes Symbol: depending on front/next words	Nerco Oil & Gas Inc., Telephone & Telegraph Co., …

사를 [Part-of-Speech] 부분에서 선택한다. 의미를 결정할 때 고려할 정책을 [Policy] 부분에 제시하여 일관성 있는 의미 결정을 할 수 있도록 하였다. 의미 부착 도구는 사전 미등록어와 복합명사를 제시하고, 미등록 단어와 복합명사의 문맥을 제공하는 등의 기능을 갖추어 사람이 효율적으로 의미 부착 작업을 수행할 수 있도록 지원한다.

의미 부착의 결과로서 사전 미등록 단어와 복합명사에 대한 의미가 부착된 파일이 생성된다. 대문자로 시작하는 사전 미등록 단어를 포함하는 파일(Fig. 8(a)), 소문자로 시작하는 미등록 단어를 포함하는 파일(Fig. 8(b)), 그리고 복합명사 파일(Fig. 8(c)) 등 3가지 파일이 생성된다. 복합명사 파일은 모두 명사를 포함하지만 다른 파일은 다양한 품사를 가지는 단어를 함께 포함하므로 각 단어의 품사 또한 같이 나타난다.

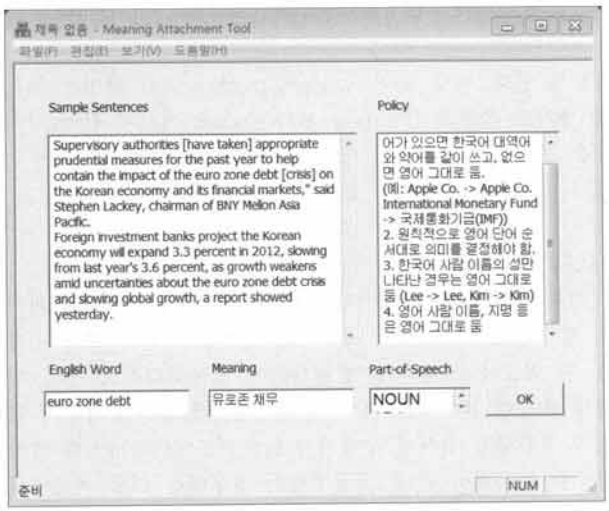

Fig. 7. Example of Meaning Attachment Tool Interface

5) 예로서 제시한 것 중 '&'가 이에 해당한다. 첫 문자가 알파벳이 아닌 단어는 숫자와 기호가 해당하며, 기호의 경우 전-후 단어를 고려해야 하므로 '&'를 전-후 단어와 함께 예로서 제시하였다.

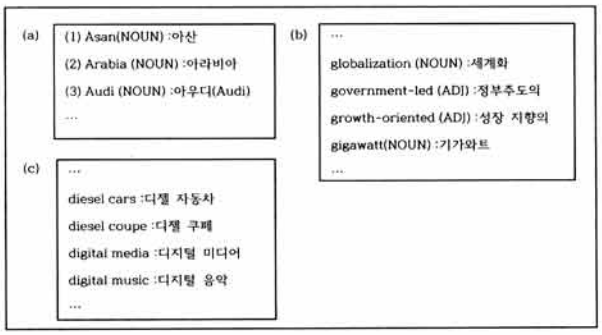

Fig. 8. Results of Meaning Attachment ((a) File for Un-registered Words Beginning with Capital Letter, (b) File for Un-registered Words Beginning with Lower case Letter, (c) Compound Noun File)

3.5 변환사전과의 통합 지원 도구

의미가 부착된 단어와 복합명사들은 기존 영한 변환사전에 통합되어야 하는데, 알파벳 순서를 유지하면서 단어들이 추가되어야 하므로 이를 자동적으로 수행하는 도구가 필요하다. 복합명사는 기존의 품사별 사전 이외에 독립된 복합명사 사전을 마련하여 복합명사 정보를 별도로 유지하도록 하였다. 따라서 3.4 절에서 생성된 복합명사 파일 자체가 복합명사 사전의 역할을 한다. 그리고 이 도구는 어휘사전에 필요한 품사 정보를 추가하는 기능을 제공하는데, 어휘사전에 있던 단어의 경우에는 품사정보만을 추가하며, 그렇지 않은 단어에 대해서는 알파벳 순서에 맞게 단어와 품사정보를 함께 추가하도록 하였다. 결과적으로 보다 많은 단어를 포함하는 확장된 변환사전과 복합명사 사전이 최종 결과물이 된다.

4. 실 험

영한 기계번역 시스템의 영한 변환사전 확장을 위해 인터넷 영어 신문기사 말뭉치를 수집하고 수집한 말뭉치로부터 사전 미등록 단어와 복합명사를 추출하였다. 신문이 신조어를 많이 포함하고 있다고 판단하여 신문으로부터 말뭉치를

구축하였으며, 경제, 과학, 기술 분야의 기사를 추출하였다. Table 2는 2012년 2월 26일부터 5월 20일까지 12주간 동안, 2명이 매주 20분씩의 작업을 통해 구축한 말뭉치에 대한 통계를 보여준다.

Table 2. Statistics of Collected English Corpus

	# of words	# of sentences
MK English newspaper	1,010,598	43,407
Korea Herald	506,449	24,507
DailyJoongang	316,047	15,409
Total	1,833,094	83,323

Table 3은 수집한 말뭉치로부터 추출한 사전 미등록어의 수를 보여준다. 사전 미등록어를 추출한 후, 일정 빈도수[6] 이상의 단어들에 대해서 특수문자의 포함여부를 검사한 후 사전 등록 대상을 선정한다. 사전과의 통합을 용이하게 하기 위해 대문자 시작 단어와 소문자 시작 단어를 구분하였다.

Table 3. The Number of Collected Un-registered Words

	# of words
Word beginning with capital letter	708
Word beginning with lowercase letter	260
Total	968

Table 4는 3.3절에서 설명한 복합명사 수집 도구를 이용하여 2, 3, 4 단어로 구성된 복합명사를 추출한 결과이다. 5 단어 이상의 복합명사도 존재하겠지만 그 수가 적을 것이라 판단하여 4 단어 이하의 복합명사만을 추출하였다.

Table 4. The Number of Compound Nouns from DailyJoongang, Korea Herald, and MK

Word\ Frequency	2	3	5	7	10
2 words	17,350	11,249	5,296	3,214	1,951
3 words	4,940	3,003	1,120	613	309
4 words	1,340	761	243	121	47
Total	23,630	15,013	6,659	3,948	2,307

위의 표들은 기준 빈도수를 2, 3, 5, 7, 10으로 정하고 기준 빈도수 이상 나타나는 복합명사의 개수를 보여준다. 기준 빈도수가 높을수록 추출되는 복합명사의 개수가 줄어드는 것은 당연하나, 10번 이상 나타나는 복합명사의 총 수는 2,307로 논문에서 제시한 방법으로 많은 복합명사를 추출할 수 있음을 알 수 있다.

12주간의 작업을 통해 사전 미등록 단어 968개와 약 2,300개 이상의 복합명사를 수집할 수 있었다. 이후 전문가에 의한 의미 부착과 영한 변환사전과의 통합이 이루어진다. 전문가는 사전 미등록 단어 중 's가 포함된 소유격 단어와 명사의 복수형 단어들은 제외하였는데, 결과적으로 74개가 제외되고 894개의 단어들만 변환사전에 추가되었다[7]. 그리고 수집된 복합명사들은 모두 복합명사 사전에 등록되었다. 전문가에 의한 의미 부착 작업 이외의 다른 과정은 모두 자동적으로 수행되기 때문에 최소한의 노력을 투자하여 지속적으로 영한 변환사전을 확장할 수 있는 기반이 마련되었다고 할 수 있다.

5. 결 론

본 논문에서는 영한 기계번역 시스템의 영한 변환사전의 확장을 위해 영어 말뭉치 구축과 이로부터 사전 미등록어와 복합명사를 수집하여 기존 변환사전에 추가하는 일련의 과정으로 구성된 영한 변환사전 확장 방안을 제시하고 이를 지원하는 도구를 개발하였다. 영한 변환사전 확장 도구는 인터넷 신문기사를 추출하여 영어 말뭉치를 구축하고 기존 영한 변환사전에 없는 단어와 일정 빈도수 이상 나타나는 복합명사를 추출하고, 의미추가 및 영한 변환사전과의 통합 등의 과정을 지원하는 모듈로 구성된다.

복합명사의 수집을 위해 복합명사를 형성할 수 없는 불용어를 정의하여 문서에서 복합명사 후보를 추출한 후, 빈도수를 기준으로 복합명사로 간주하는 간단한 방법을 적용하였다. 이는 영한 기계번역을 위해 최대한 많은 복합명사를 인식하여 최대한으로 성능을 개선하기 보다는 의미 있는 성능 개선을 약간의 노력으로 달성하려는 목적에 부합한다고 할 수 있다. 따라서 논문에서 제안한 복합명사 수집 방법은 실제로 적용될 수 있는 유용성이 있다고 판단한다.

논문에서 제안한 영한 변환사전 확장 방안은 적은 사람의 노력만을 투입하여 지속적으로 사전 정보량을 확대할 수 있도록 한다. 이를 통해 영한 기계번역 시스템의 번역품질을 유지하고 개선하는 것을 지원할 것으로 기대된다. 또한 영어 말뭉치 구축 기능은 영한 기계번역 시스템의 개선을 위해 다양하게 활용될 수 있을 것이다.

앞으로 새로 수집된 단어들에 대한 분석이 필요할 것으로 판단된다. 새로운 명칭을 나타내는 고유명사, 하이픈(-)을 포함하는 명사 또는 형용사들이 많이 나타나는데 이들에 대한 의미 부착을 자동화 또는 반자동화 하는 방안을 연구할 예정이다. 의미 부착이 자동화될 수 있다면 사전에 등록하지 않고 번역 시스템에서 직접 한국어 대역어를 생성할 수 있으므로 매번 사전에 등록하는 것보다 사전 미등록어에 대한 일반적인 해결방법이 될 것이다. 또한 추출한 복합명사에 대한 분석을 통해 불용어를 확장하고, 보다 정교한 불용어 확인 방법을 고안하여 대용량의 데이터로부터 의미있는 복합명사를 추출하는 연구가 필요하다. 그리고 숙어인식을

6) 여기서는 5번 이상 나타나는 단어를 대상으로 하였음.

7) 이들 단어에 대해서는 's를 제외한 단어나 단수형 단어가 변환사전에 포함된 여부를 고려하여 변환사전에 이미 포함된 단어는 어휘사전에만 등록하고, 그렇지 않은 경우에는 어휘사전과 변환사전에 동시에 등록한다.

이용하여 복합명사 처리하였던 기존의 방식을 개선하여 3.5절에서 언급한 것처럼 복합명사를 사전을 독립적으로 유지하고 이를 어휘분석과 구문분석 단계에서 활용하도록 하여 영어 분석의 효율성을 개선하는 연구가 필요하다. 이는 복합명사를 포함하는 문장에 대해 보다 빠르고 정확한 번역을 가능하게 할 것이며 궁극적으로 영한 기계번역 시스템의 성능에도 긍정적 영향을 미칠 것이다.

참 고 문 헌

[1] Jeff Allen, "Improved Translation Quality with Machine Translation Dictionary Building", TranslatioCafe.com, June, 2006.

[2] Mary McGee Wood, E. Pollard, H. Horsfall, N. Holdel, B, Chandler, and J. Carroll, "Dictionary Organization for Machine Translation: The Experience and Implications of the UMIST Japanese Project", Proceedings of the 3rd Conference on European Chapter of the Association for Computational Linguistics, 1987.

[3] H. S. Lee, Y. T. Kim, "Automatic Extraction of Collocations and Verbal Idioms from Corpus for a Generation of English-Korean Transfer Dictionary," Journal of KIISE: Vol.21, No.6, pp.2110-2117, 1994.

[4] S. J. Lee, S. K. Park, Y. T. Kim, "Head-based Phrase Structure Transfer Dictionary for Korean-English Machine Translation," in Proceedings of the 6th Human and Cognitive Language Technology (HCLT), 1994.

[5] C. Y Ok, "Phrase-based Transfer Dictionary for Korean-English Machine Translation," Phd. Thesis, Dept. of Computer Engineering, Seoul National University, 1993.

[6] S. M. Kim, C. W Min, S. C. Kang, J. I. Char, "Method and Apparatus for developing a transfer dictionary used in transfer-based machine translation system," Patent No. 100530154, 2005.

[7] Su Nam Kim, "Statistical Modeling of Multiword Expressions," Ph.D. thesis, University of Melbourne, Melbourne, 2008.

[8] H.-S. Bae, K.-S. Choi, "Electronic Dictionary for Performance Improvement of the Information Retrieval System," Journal of French Culture and Art Study, No.6, pp.69-82, 2002.

[9] Jansche, Martin. "Named Entity Extraction with Conditional Markov Models and Classifiers," Proceedings of Conference on Computational Natural Language Learning, pp.1-4, 2002.

[10] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons," Proceedings of Conference on Natural Language Learning, pp.188-191, 2003.

[11] Y. Shinyama and S. Sekine, "Named Entity Discovery Using Comparable News Articles," Proceedings of the International Conference on Computational Linguistics, 2004.

[12] A. Kunchukuttan and Om P. Damani, "A System for Compound Noun Multiword Expression Extraction for Hindi," Proceedings of ICON-2008, 6th International Conference on Natural Language processing, pp.20-29, 2008.

[13] Yujie Zhang and Hitoshi Isahara, "Acquiring Compound Word Translations Both Automatically and Dynamically," Proceedings of the Pacific Asia Conference on Language, Information, and Computation, pp.181-186, 2004.

[14] Sung-Dong Kim, Da-Un kang, Bohee Lee, Dorim Kim, "Development of Dictionary Management Tool for English-Korean Machine Translation System," in Proceedings of the 36th KIISE(Korean Institute of Information Scientists and Engineers) Fall Conference, Vol.36, No.2(C), pp.199-203, 2009.

김 성 동

e-mail : sdkim@hansung.ac.kr

1991년 서울대학교 컴퓨터공학과(공학사)

1993년 서울대학교 컴퓨터공학과 (공학석사)

1999년 서울대학교 컴퓨터공학과 (공학박사)

1999년~2001년 ㈜엘앤택 기술이사

2001년~현 재 한성대학교 컴퓨터공학과 교수

관심분야 : 기계번역, 자연언어처리, 데이터마이닝