공군 정보 영한 기계번역 시스템 : ALKOL *

이현아○ 임철수 최명석 강인호 김길창
한국과학기술원 전자전산학과
{halee, cslim, mschoi, ihkang}@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

English-to-Korean Machine Translation System for Air Force Intelligence : ALKOL

Hyun Ah Lee Chul Su Lim Myung Seok Choi In Ho Kang Gil Chang Kim
Dept. of Electrical Engineering & Computer Science, KAIST

요 약

본 논문에서는 공군 정보 번역을 위한 영한 기계번역 시스템 ALKOL에 대해서 소개한다. ALKOL은 어휘화된 규칙에 기반한 번역 시스템으로, 어휘화된 규칙은 어휘-분석-변환-생성의 네 단계의 정보가 연결된 형태로 사전에 저장된다. 이와 같은 사전 구조에 의해 번역 과정의 효율성을 높일 수 있고, 어휘화된 규칙에 의해 정확하고 자연스러운 번역 결과를 얻을 수 있다. ALKOL의 번역 과정은 형태소 분석, 품사 태깅, 분석 전처리, 구문 분석, 변환, 생성의 단계로 이루어진다. 각 단계에서는 전/후처리를 보강하여 실제 번역 환경에서 나타나는 문제들을 해결하고, 하나 이상의 번역 결과를 출력하여 사용자가 원하는 결과를 선택할 수 있게 한다.

1 서론

본 논문에서는 공군 정보의 번역을 위한 영한 기계번역 시스템 ALKOL에 대해서 소개한다. ALKOL은 어휘화된 규칙에 기반한 변환 방식 번역 시스템이다[1]. 어휘화된 규칙은 사전에 저장되며, 사전은 어휘-분석-변환-생성의 네 단계로 구성된다. 각 단계의 사전은 연결된 형태로 구성되어 사전 검색의 횟수를 줄일 수 있으므로 번역 과정의 효율성을 높일 수 있다. 또한, 어휘화된 규칙을 이용하므로 보다 정확하고 자연스러운 번역 결과를 얻을 수 있다. 번역은 사전에 적재된 정보를 이용하여, 형태소 분석, 품사 태깅, 분석 전처리, 구문 분석, 변환, 생성의 과정으로 이루어진다. 변환 사전에는 번역어 선택 정보와 함께 영어 구구조를 한국어 의존 구조로 대응시키기 위한 정보가 들어 있어, 구문 수준의 변환(syntactic transfer;[9])으로 번역이 수행된다.

ALKOL은 공군 정보 분야의 영어 문서 중에서 전문 잡지를 대상으로 한다. 대상 문서에서는 문장이 대체적으로 길고 애매한 경우가 많으며, 숫자나 단위 표현, 복합 명사, 삽입 어구 등이 자주 나타난다. ALKOL에서는 이러한 특성에 맞게 각 단계의 전/후처리 단계를 보강하여 실용성 있는 번역 결과를 얻게 한다. 또한, 애매성이 있는 입력에 대해서는 하나 이상의 번역 결과를 출력하여 사용자가 원하는 결과를 선택할 수 있게 한다.

ALKOL은 클라이언트-서버의 분산 환경에서 구축되어, 서버 측에서는 번역 서버와 사전 서버, 클라이언트 측에서는 사용자 인터페이스와 번역 사전에 대한 사전 편집기가 개발된다. 사용자 인터페이스에서는 문자 인식을 지원한다.

본 논문은 다음과 같이 구성된다. 2절에서는 ALKOL 시스템의 전체 개요를 설명한다 3절에서는 ALKOL의 사전 구조에 대해서 설명하고, 4절에서는 번역 서버에서의 번역 과정을 설명한다. 5절에서는 사용자 인터페이스를 보이고, 6절에서는 실험 및 평가 결과를 간단히 보이고, 7절에서는 마지막으로 결론을 맺는다

2 ALKOL의 시스템 개요

ALKOL은 PC와 HP workstation간의 클라이언트-서버의 분산 환경에서 개발된다[5]. 그림 1은 ALKOL의 구조를 보인다.

*본 연구는 공군본부 연구과제 "군사정보 기계번역 시스템 용역개발"에 의해 지원되었음.

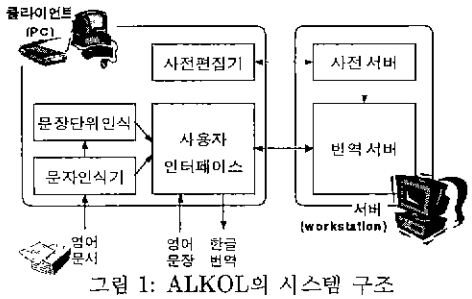

그림 1: ALKOL의 시스템 구조

PC의 클라이언트에는 사용자 인터페이스와 사전 편집기가 존재한다. 사전 편집기는 관리자나 개발자가 사전에 접근하고 사전 정보를 편집하기 위한 환경을 제공한다. 사용자 인터페이스에서는 사용자의 영어 문장을 입력으로 받아 번역 서버에 보내고 번역된 결과를 받아 사용자에게 제시한다. 또한 사용자 인터페이스는 문자 인식 기능을 포함하여 인쇄물에 대한 번역도 지원한다[8].

서버는 사전 서버와 번역 서버로 구성된다.

사전 서버는 사전의 내용을 데이타베이스로 구성하여 사전 정보의 효율적인 삽입, 수정, 삭제를 지원하며, 번역 서버의 요청을 받아 사전 정보를 제공한다. 번역 사전은 시스템 사전과 하나 이상의 사용자 사전으로 구성된다. 시스템 사전은 영한 번역을 위한 기본 사전에 해당된다. 사용자 사전은 각 사용 영역마다의 특성을 반영하는 추가적인 사전이다. 사용자 인터페이스에서는 번역 과정에서 어떤 사용자 사전을 추가적으로 사용할 지 선택할 수 있고, 번역 과정에서 사용자 사전은 시스템 사전보다 우선적으로 사용된다. 사용자 사전과 시스템 사전은 어휘-분석-변환-생성 네 단계의 동일한 구조를 가진다.

번역 서버는 사전 서버에서 제공하는 사전 정보를 이용하여 사용자 인터페이스에서 요청한 영어 입력 문장을 한국어 문장으로 번역한다. 그림 2는 번역 서버의 구조를 보인다. 번역 서버에 영어 문장이 입력으로 들어오면 영어 형태소 분석과 품사 태깅을 통해 각 단어에 대한 2순위 품사까지의 사전 정보를 얻는다. 사전 정보와 품사 정보를 이용하여 명사구 인식(NP chunking) 등의 구문 분석 전처리를 거친 결과는 구문 분석을 거쳐 영어 구구조로 분석된다. 영어 구구조는 영한 변환을 거쳐 한국어 의존구조로 변환된다. 분석과 변환에서는 일반 규칙(general grammar)과 사전의 어휘화된 규칙(lexicalized grammar)을 사용한다. 한국어 생성에서는 의존구조를 한국어 문장으로 만들어 번역 결과를 얻는다.

ALKOL 사전 구조와 번역 서버에서의 번역 과정은 다음 절에서 자세히 다룬다.

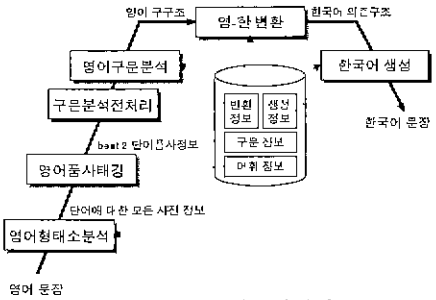

그림 2 ALKOL 번역서버의 구조

3 번역 사전의 구조와 내용

ALKOL은 어휘화된 규칙에 기반한 번역 시스템이다[1]. 어휘화된 규칙은 번역 사전에 저장된다. 번역 사전은 번역 과정의 효율성을 위해 그림 3과 같이 어휘-분석-변환-생성의 네 단계의 연결된 형태로 구성되며, 각각에 대한 고유한 이름(entry ID)를 가진다[7].

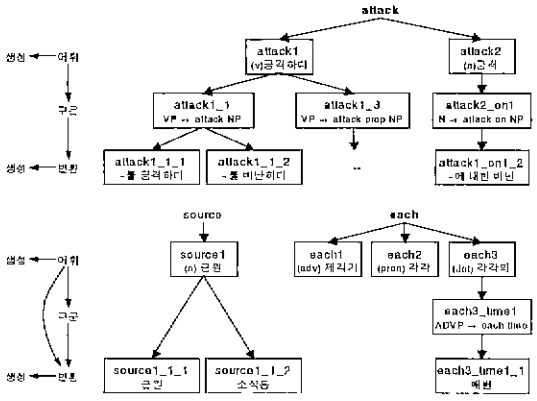

그림 3: ALKOL의 사전 구조

- 어휘 사전
  어휘 사전은 영어 어휘를 품사별로 각기 다른 엔트리로 저장한다. 예를 들어 'attack'에 대해서는 동사 엔트리 attack1와 명사 엔트리 attack2가 사전에 저장된다 어휘 사전은 각 엔트리 어휘에 대한 품사 정보, 활용 정보, 의미 자질 정보 등을 가진다. 복합어도 어휘 사전에 저장된다.

- 분석 사전
  분석 사전은 어휘화된 영어 구문 분석 규칙을 가진다 그림 3에서와 같이 분석 사전 엔트리는 규칙에 포함된 어휘의 어휘 사전 엔트리에 연결된다. 분석

사전과 변환 사전에 작성되는 어휘화된 규칙은 단순 동사구/전치사구, 단순 관용어구, 복합 관용어구, 양상 자질 관련의 네 종류로 나뉜다. 단순 동사구/전치사구 규칙은 규칙의 RHS(right hand side)가 하나의 동사나 전치사 어휘와 하나 이상의 구문분석심볼로 구성되는 경우를, 단순 관용어구는 RHS가 어휘로만 구성되는 경우를, 복합 관용어구는 RHS가 하나 이상의 어휘와 구문분석심볼으로 구성되는 경우를 나타낸다. 양상 자질 관련 규칙은 희망, 불가능 등의 양상(modal)를 처리하기 위한 규칙이다.

그림 4는 단순 동사구 VP → attack NP 규칙에 대한 사전의 엔트리 attack1_1의 예를 보인다. 분석 사전은 문맥 자유 문법(CFG) 형식의 어휘화된 구문 분석 규칙, 규칙이 적용되기 위한 조건(CONDITION), 규칙이 적용되었을 때 동작할 작용(ACTION), 각 규칙의 가중치, 연결된 변환 사전 엔트리 등의 구성 요소를 가진다. 그림 4의 분석 사전의 조건에서는 주어진 규칙이 적용되기 위해서는 'attack'의 품사가 동사이고 명사구는 목적격이어야 함을 명시한다. 조건이 성공하면 규칙이 적용되고, 작용에 의해 LHS의 동사구에 동사 'attack'이 가지는 모든 자질이 전달된다

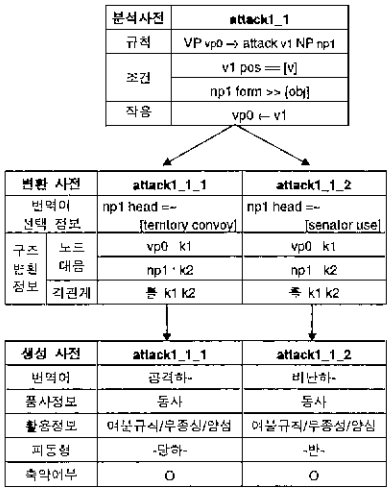

그림 4: 단순 동사구 규칙의 예제

그림 5에서는 N → attack on NP와 같은 복합 관용어구에 대한 사전 예제를 보인다. '~에 대한 공격'으로 번역되는 이 표현은 전치사 'on'의 위치에 'upon'이나 'against'가 와도 동일한 의미로 번역된다. 이 경우 N → attack prep NP로 규칙을 기술하고 조건을 기술하여, 중복되는 의미와 형태의 규칙을 하나의 엔트리로 기술할 수 있다. 마찬가지로 'make big use of' 'make great use of' 등도 VP → make NP of NP의 형태로 기술하고, 첫번째 명사구의 중심어(head word)가 'usc'라는 조건을 명시하여 하나의 엔트리로 처리한다.

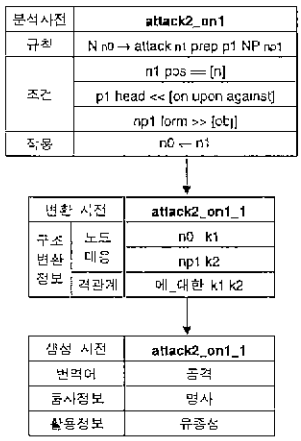

그림 5: 복합 관용어구 규칙의 예제

그림 6은 ADVP → each time과 같은 단순 관용어구나 VP → want to VP와 같은 양상 자질 관련 규칙의 예제를 보인다.

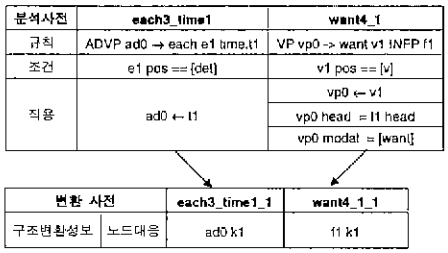

그림 6: 단순 관용어구와 양상 자질 관련 규칙 예제

- 변환 사전

  변환 사전은 영어 단어의 한국어 번역어를 선택하기 위한 정보와 분석 사전의 규칙으로 분석된 영어 구구조를 한국어 의존 구조로 변환하기 위한 정보를 가진다[2]

  원시 언어 단어에 대한 번역어를 선택하기 위해서는 문맥에서 발생하는 단어, 즉 공기 어휘 정보를 사용한다. 그림 4의 변환 사전에서는 VP → attack NP 규칙으로 분석된 영어 구조에서, 명사구 NP의 위치에 오는 단어가 'territory'나 'convoy'와 유사한 단어이면 'attack'을 attack1_1_1(공격하다)로 번역하고, 'senator'나 'use'와 유사한 단어이면 attack1_1_2(비난하다)로 번역하기 위한 변환 사전의

예를 보인다. 번역어 선택 정보로는 공기 어휘 뿐만 아니라 공기 어휘의 품사, 의미 자질 등의 정보도 사용할 수 있다.

구조 변환 정보는 노드 대응과 격관계의 두가지로 구성된다. 그림 4의 사전에서는 동사구와 명사구를 각각 하나의 한국어 노드로 대응시키고 두 노드를 목적격 관계, 즉 조사 '을/를'으로 연결하는 형태를 보인다. 그림 5에서는 attack on NP를 '~에 대한 공격' 으로 번역하기 위한 사전 형태를 보인다.

동사나 전치사에 관련된 규칙이나 관용어구 관련 규칙은 문장 구조를 결정하기 때문에 구문 분석 규칙으로 기술하지만, 일반 명사나 형용사 등에 대해서는 분석 규칙을 따로 기술할 필요가 없다. 변환 사전은 그림 3의 'attack'의 경우나 'each'의 경우 같이 분석 사전에 연결된 경우 뿐만 아니라, 'source'와 같이 어휘 사전에 바로 연결된 형태도 허용한다. 어휘 사전에 바로 연결된 변환 사전에는 번역어 선택 정보만 기술된다. 그림 7에서는 명사 'source'에 대한 번역어 선택 정보를 보인다 명사 'source'는 군사정보 영역에서 '근원'보다는 '소식통'으로 더 자주 번역된다. 사전에는 변환된 한국어 구조에서 'source'가 주격 관계 '이/가'로 연결되어지는 가상 노드에 'say'와 유사한 단어가 오면, 'source'를 source1_1_2(소식통)으로 번역하도록 사전이 기술된 예를 보인다. 이처럼 어휘 사전에 연결된 변환 사전의 번역어 선택 정보는 한국어 구조를 이용하므로, 번역 과정에서는 한국어 구조가 완성된 이후인 후변환 단계에서 사용된다.

변환 사전	source1_1_1	source1_1_2
가상 노드	e0 가	e0 가
번역어 선택 정보	e0 head =~ [be]	e0 head =~ [say]

그림 7: 명사 'source'의 번역어 선택 정보

- 생성 사전

  어휘 사전과 변환 사전은 한국어 생성을 위한 정보를 가지고 있는 생성 사전과 연결된다. 생성 사전에서는 불규칙 활용 정보나 이형태 정보, 격틀 정보, 축약 여부, 명사 특성에 맞는 단위 명사 등이 기술된다. 예를 들어 그림 4에서 attack1_1_2는 동사 어간 '비난하'로 번역되며, 여불규칙 활용을 하고, 수동태가 될 때는 '-받-'을 피동어미로 이용한다는 정보를 보인다 attack2_on1_1에 대해서는 유증성 명사 '공격'으로 번역된다는 정보가 기술된다.

  이외에 생성 사전에서는 한국어 특성을 반영하기 위한 다양한 정보가 기술된다. 일반적으로 수동태를 한국어에서 표현하기 위해서는 '-어 지-' '-받-' 등의 피동 어미를 이용하는데, 'teach'나 'hit' 등은 수동태를 만들기 위해서는 어미 변형이 아니라 단어 자체를 바꿔야 한다. ALKOL의 생성 사전에서는 'hit'를 능동태에서는 '때리다'로 번역하고 수동태가 될 때는 어미 변형을 하지 않고 '맞다'라는 단어로 번역한다는 정보를 기술한다. 그리고, 'international'(국제적인/국제)와 같이 한국어에서 형용사로 쓰이는 경우와 명사 수식어로 쓰일 때 형태가 달라지는 단어에 대하여 수식어 형태를 별도로 지정해 주는 등의 다양한 정보를 포함할 수 있다.

이처럼 ALKOL에서는 어휘 고유의 특성을 반영한 규칙을 사전에 저장하여 번역에 이용하므로 번역결과의 품질을 높일 수 있고, 각 단계가 연결된 형태의 사전을 이용하므로 사전이나 규칙의 검색 횟수를 줄여서 번역 과정의 효율성을 높일 수 있다. 또한, ALKOL에서는 대상 영역 코퍼스의 특성을 영역 분석 워크벤치를 통해 분석하고, 이를 반영하여 사전을 구축하여, 대상 영역에 최적화된 번역 결과를 얻을 수 있다[4].

4 ALKOL의 번역 과정

ALKOL의 번역과정은 그림 2와 같이 영어 형태소 분석과 품사 태깅, 구문 분석 전처리와 구문 분석, 영한 변환과 생성의 단계로 구성된다. ALKOL의 각 단계는 하나 이상의 분석 결과를 다음 단계로 전달할 수 있어 여러개의 번역 결과를 사용자에게 제시할 수 있다. 또한, 번역 과정의 효율성과 결과의 정확성을 높이기 위해 각 단계에 후처리를 보강하고 구문 분석 전처리 단계를 도입한다.

4.1 영어 형태소 분석과 품사 태깅

영어 형태소 분석기에서는 어휘 사전의 정보를 이용하여 입력된 영어 문장에 대한 형태소 분석을 수행한다. ALKOL에서는 대상 영역 코퍼스에 대한 분석 작업을 통해 고유 명사의 처리, 날짜·시간·단위 표현 등의 복합 단위 형태소 처리, 미등록어 처리 등을 강화한다[6]. 영어 형태소 분석에서는 입력 문장의 각 단어가 가지는 모든 사전 정보를 품사 태깅으로 전달한다.

영어 품사 태깅에서는 상태 기반 태깅 방법을 이용하여 각 단어가 가질 수 있는 2순위까지의 품사를 선택한다. ALKOL에서는 11개의 품사를 이용하기 때문에 펜트리뱅크 코퍼스의 품사 집합과 ALKOL의 품사 집합을 대응시켜 ALKOL의 품사에 맞는 결과를 얻도록 한다[3]. 또한, 펜트리뱅크와의 차이점으로 발생하는 문제를 해결하기 위하여 후처리를 수행한다.

두 단계를 거쳐 그림 8의 (a)의 영어 입력 문장 "The source claimed they attcked him."에 나타난 단어들의 2순위 품사까지의 사전 정보가 다음 단계로 전달된다.

4.2 구문 분석 전처리

구문 분석은 기계번역에서 가장 많은 시간을 소모하는 단계이다. ALKOL에서는 구문 분석에서의 복잡도를 감소시키고 정확도를 높이기 위해 구문 분석 전처리 단계를 도입한다. 분석 전처리 단계의 작업은 크게 다섯가지로 구성된다.

- 품사 태깅에서 결과로 낸 상위 2순위 품사 중에서 발생가능성이 희박한 품사를 제거한다. 품사 태깅에서는 번역 정확도를 높이고 복수 번역을 지원하기 위해 확률값이 상위 2등 안에 드는 품사 정보를 모두 결과로 낸다. 구문분석 전처리에서는 구문 분석의 복잡도를 낮추기 위해서 단어의 전후 품사열에 대한 휴리스틱 규칙을 이용하여 품사열에서 발생 가능성이 거의 없는 것을 제거한다.

- 정규 문법을 이용한 명사구 인식(NP chunking)을 수행한다. 정규 문법은 품사 정보와 어휘 정보를 기반으로 수동으로 구축한다. 적용률이 낮더라도 정확률이 높은 규칙을 작성하여 품사열 정보만을 이용하여 분석할 수 있는 명사구를 미리 묶어서 구문 분석의 부하를 줄인다.

- 휴리스틱 규칙을 이용하여 병렬 문장이나 문장 기호로 연결된 문장을 분리한다. 코퍼스 분석을 통해 ", and"나 ";"와 같이 문장 분리에 이용될 수 있는 패턴을 찾아 휴리스틱 규칙을 기술한다. 긴 문장을 짧은 여러 개의 문장으로 분리할 수 있으므로, 구문 분석의 복잡도를 줄이고 명료한 번역 결과를 얻을 수 있다.

- 삽입구문(parenthetical)에 대한 처리를 수행한다. 영어 문장 중간에 발생하는 ", however, ", ", as it were, "등의 삽입 구문을 문장 앞이나 뒤로 옮긴다. 삽입구의 이동으로 영어 원문의 정확한 의도가 번역문에 전달되지 못 하는 경우도 발생하지만, 구문 분석의 정확도를 향상시켜 번역 성능을 높일 수 있다.

- 괄호 표현 등에 대한 처리를 수행한다. 한 단어나 어구에 대한 부가 설명 등을 위해 이용되는 괄호 표현을 해당 단어나 어구의 자질로 저장한다. 괄호 표현만을 따로 번역한 뒤, 결과 문장에서 해당 영어 어구 뒤에 첨가시킨다.

4.3 영어 구문 분석

영어 구문 분석에서는 사전의 분석 사전에 있는 어휘화된 규칙과 일반 규칙을 이용하여 영어 문장을 구구조 트리로 분석한다.

ALKOL에서는 일반 규칙(general grammar)과 어휘화된 규칙(lexicalized grammar) 두가지의 차등화된 규칙을 이용한다[1]. 어휘화된 규칙은 일반 규칙보다 높은 가중치를 가지며 사전에 저장된다. 두가지 규칙은 같은 형식으로 기술되므로 동일한 처리 과정을 통해 사용된다. 이와 같이 차등화된 규칙을 사용함으로써, 어휘 정보를 최대한 반영하여 자연스러운 번역 결과를 기대할 수 있다. 또한, 어휘 규칙이 불완전하더라도 일반 규칙만으로도 시스템이 동작할 수 있게 규칙이 작성되어 있으므로 시스템의 견고성을 추구할 수 있다.

영어 구문 분석에서는 분석 사전의 규칙과 일반 규칙에 기술된 가중치를 이용하여 분석으로 얻어지는 각 구구조 트리의 선호도를 구한다. 분석 단계에서는 선호도가 높은 하나 이상의 구구조 트리를 변환의 입력으로 줄 수 있어, 다른 구조를 가지는 여러 개의 문장을 번역 결과로 얻을 수 있다. 그림 8의 (b)는 3절에서 소개한 사전 정보를 이용한 영어 문장 "The source claimed they attacked him."의 분석 결과를 보인다.

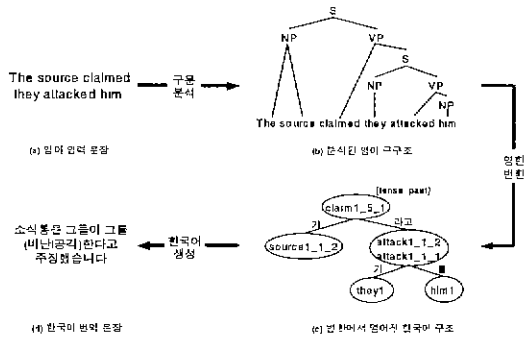

그림 8: ALKOL의 번역 예제

4.4 영한 변환

ALKOL은 구문 수준의 변환 방식(syntactic transfer) 번역 시스템이다. 영한 변환에서는 사전 정보를 이용하여 영어 구구조를 한국어 의존 구조로 변환하고 번역어 선택을 수행한다.

영한 변환은 주변환과 후변환의 두단계로 구성된다. 주변환에서는 분석 사전에 연결된 변환 사전 정보, 즉 어휘화된 규칙과 일반 규칙을 이용하여 번역어를 선택하고

구구조 트리를 한국어 의존구조로 변환한다. 후변환에서는 어휘 사전에 연결된 변환 사전 정보를 이용하여 나머지 단어들의 번역어를 선택하고, 영어와 한국어의 차이점을 해소하기 위한 후처리를 수행한다[2].

그림 8의 경우를 예로 들면, 주변환 단계에서는 그림 4와 같은 사전 정보를 이용하여 'claim'과 'attack'의 번역어를 선택하고 영어 구조를 한국어 구조로 변환한다. 후변환에서는 변환된 한국어 구조 정보와 그림 7의 사전 정보를 이용하여 'source'의 번역어를 선택한다. 번역어 선택 과정에서 입력 문장에 나타난 단어와 사전에 기술된 공기 어휘간의 유사도는 WordNet을 기준으로 [11]의 방법을 이용하여 계산한다.

후변환에서는 번역어 선택과 함께 영어와 한국어의 차이점을 해소하기 위한 작업이 수행된다. ALKOL에서는 [2]에서 제안한 부정 자질 표현 처리와 문장 주체 표현 차이 처리 뿐만 아니라, 보조사 생성, 전체부정/부분부정의 처리, a/the의 번역 등의 작업을 추가한다[6].

구문 분석에서는 하나 이상의 구구조를 출력하므로 문장 단위의 복수 후보를 제공한다. 변환에서는 단어 단위나 구 단위의 복수 후보를 생성할 수 있다. 예를 들어 그림 4의 사전을 이용하면 문장 "The source claimed they attacked him."에서 '비난하다'와 '공격하다'가 모두 'attack'의 번역어로 선택될 수 있고, 이들은 그림 8의 (c)와 같이 모두 생성의 입력으로 넘어간다.

변환 결과로 얻어지는 한국어 의존 구조는 한국어의 실질어가 노드에 대응되고 한국어의 기능어가 노드간의 링크의 값이나 노드의 자질값으로 대응되는 형태이다. 그림 8의 (c)는 (b)에 대한 한국어 의존 구조를 나타낸다.

4.5 한국어 생성

한국어 생성에서는 영한 변환에서 얻어지는 한국어 구조에 대하여 어순 처리, 격전이 등을 포함하는 구문 생성과 기능어 처리와 음운현상 처리 등을 포함하는 형태소 생성 단계를 거쳐 한국어 문장을 얻는다.

한국어 생성을 거쳐 ALKOL에서는 영어 문장 "The source claimed they attacked him."에 대하여 "소식통은 그들이 그를 {비난|공격}한다고 주장했습니다"의 결과를 출력한다. 사용자 인터페이스에서는 가장 선호도가 높은 '비난한다고'가 기본 번역으로 제시되고, 복수번역 선택 기능을 통해 '공격한다고'를 선택하거나 후편집을 통해 원하는 다른 단어로 교체할 수 있도록 지원한다. 마찬가지 과정으로 영어 문장 "I saw the boy with a telescope."에 대해서는 "나는 {망원경을 가지고|망원경을 가진|망원경과 함께} 소년을 {보았습니다|톱으로 자릅니다}"의 번역 결과를 얻을 수 있다. 사용자는 가장 선호도가 높은 "나는 망원경을 가지고 소년을 보았습니다"을 그대로 번역 결과로 받아들이거나, 복수번역 선택 기능을 통해서 "나는 망원경을 가진 소년을 보았습니다"를 선택할 수 있다.

일반적인 번역 시스템에서는 시스템에서 결정한 하나의 번역 결과를 출력한다. 하지만, 문맥 정보가 한정되어 있고 번역 지식이 불충분하기 때문에 언어 현상에서 발생하는 모든 애매성을 해결할 수 없다. 그러므로 시스템이 결정한 하나의 번역 결과의 정확도는 그리 높지 않다. 이 경우 번역 능력이 없는 사용자가 번역 시스템을 이용하는 경우 영어 문서에 대한 올바른 이해가 불가능하고, 번역 능력이 있는 사용자가 번역 시스템을 사용하더라도 결과와 원문을 비교하여 번역 결과를 선별적으로 사용하거나 사전을 뒤져 다시 번역을 하는 등의 번거로운 과정을 거쳐야 한다.

ALKOL에서는 입력 문장에 대한 하나 이상의 번역 결과를 사용자에게 제시한다. 따라서 하나만의 결과를 사용자에게 보이는 경우보다 높은 정확도를 기대할 수 있고 번역 결과의 활용도도 높아진다. 하나 이상의 번역 결과는 사용자 인터페이스를 통하여 보기 쉽게 사용자에게 제시되고, 사용자는 원하는 번역을 선택할 수 있다. 다음에서는 사용자 인터페이스에 대해서 설명한다.

5 ALKOL의 사용자 인터페이스

그림 9는 ALKOL의 사용자 인터페이스를 보인다. 사용자 인터페이스는 MS Windows 환경에서 동작한다. 그림에서 볼 수 있듯이 사용자 인터페이스는 크게 좌우의 두 개의 창으로 구성된다. 좌측은 작업 관리창에 해당한다. 번역의 대상이 되는 인쇄된 문서와 온라인 텍스트, 그리고 번역 결과 파일이 하나의 작업을 구성한다. ALKOL의 인터페이스에서는 작업을 그림 9의 좌측과 같이 나타내어, 하나의 프로그램에서 진행되고 있는 여러 작업들을 손쉽게 관리할 수 있다. 우측 창은 선택된 작업에 관련된 문서, 즉 번역 대상 문서와 번역 결과 문서 등을 보여준다.

사용자 인터페이스의 기능은 크게 기본 기능, 번역 기능, 사용자 편의 지원의 세부분으로 나눌 수 있다.

- 기본 기능
  일반적인 윈도우즈 프로그램에서 지원하는 기본 기능을 포함한다. 문서나 그림 파일 열기, 문서 저장, 인쇄, 인쇄기 설정, 미리 보기 등의 파일 메뉴와 잘라 내기, 붙여 넣기 등의 편집 기능을 제공한다.

- 번역 기능
  번역에 관련된 기능을 제공한다.

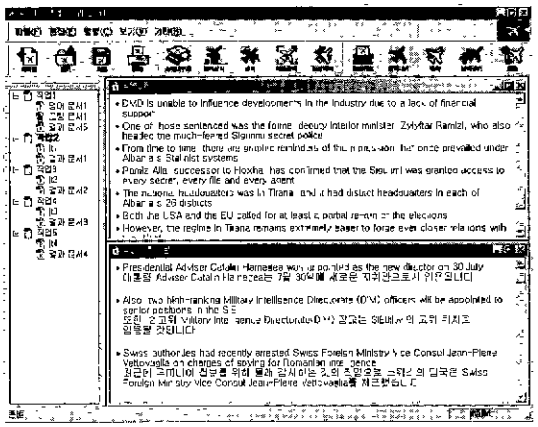

그림 9: ALKOL의 사용자 인터페이스

- 문자 인식 기능
인쇄된 영어 문서를 스캐너를 이용하여 그림 파일으로 받아들이고 문자 인식, 문장 단위 인식을 통하여 영어 텍스트 파일로 만든다.

- 번역 실행 기능
문장 단위로 인식된 영어 문장들에 대한 번역을 수행한다. 번역 중단과 문장 건너 뛰기가 가능하다.

- 일괄 처리 기능
그림 파일로 받아들여진 내용들에 대해서 문자인식과 번역 작업을 일괄적으로 실행시킨다.

- 역문 선택 기능
ALKOL에서는 번역 결과로 하나 이상의 구나 문장을 출력한다. 사용자 인터페이스에서는 주어진 영어 문장에 대해서 가장 높은 선호도를 얻은 한글 번역만을 번역창에 보여준다 영어 문장에 대한 번역 결과가 하나일 때는 번역 결과 문장이 검정색으로 표시되고, 복수 후보가 있으면 결과 문장이 파란색으로 표시된다. 파란색으로 표시된 문장을 더블 클릭하면 그림 10와 같은 번역 후보 선택창이 떠서 이외의 번역 결과를 선택할 수 있다. 그림 10은 "I saw the boy with a telescope."에 대한 문장 단위 후보를 선택하기 위한 창을 보인다. 망원경을 가진의 부분만 선택하면, '망원경과 함께' 등의 'with'에 대한 구단위 복수 번역을 선택할 수 있다.

- 사전 선택 기능
번역에 사용할 사용자 사전을 선택할 수 있다. 하나 이상의 사용자 사전을 선택할 수 있고, 선택된 사용자 사전은 번역 과정에서 시스템 사전보다 우선적으로 이용된다.

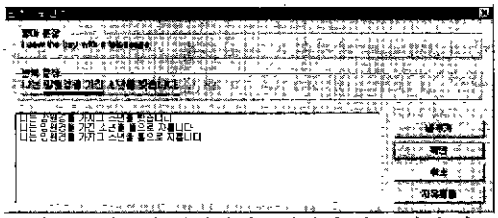

그림 10: 사용자 인터페이스에서의 역문 선택 기능

- 사용자 편의 지원
사용자의 사용 편의를 고려하여 제공되는 기능이다.

  - 전편집-후편집 기능
  번역 입력 문장에 대한 전편집과 번역 결과 문장에 대한 후편집을 지원한다.

  - 문장 단위 인식 기능
  번역은 문장을 단위로 이루어지므로, 번역 대상 문서는 한 문장이 한 줄을 구성해야 한다. 만일 시스템에서 입력 문서에 대한 문장 단위 인식을 디폴트로 수행시키면, 정확하게 문장 단위로 정렬된 문서도 문장 단위 인식기의 오류로 잘못된 문서로 바뀔 수 있다. ALKOL에서는 사용자가 원하는 경우에 문장 단위 인식을 수행할 수 있도록 메뉴를 제공하여, 사용자가 일일이 문장 단위로 문서를 구성하는 번거로움을 없애고 문장 단위 인식에 의한 오류는 전편집을 통해 해결할 수 있게 한다. 문장 단위 인식을 위해서는 [10]의 방법을 이용한다.

  - 편집기 연결 기능
  번역 입력 문서와 출력 문서를 편집기나 워드 프로세서와 같은 외부 프로그램으로 보내는 기능을 제공한다. 편집기를 이용하여 글자체, 인쇄 형태 등을 사용자가 원하는 수준으로 변경하거나 문자 인식 결과나 번역 결과에 나타나는 철자 오류 등을 수정할 수 있다.

  - 설정 기능
  번역 서버와 사전 서버 주소를 설정하거나, 번역 결과 보기의 형태를 변경하거나, 원하는 편집기나 워드 프로세서를 선택할 수 있다.

6 평가 및 토론

ALKOL의 사전은 현재 약 6만 엔트리의 어휘 사전, 약 5천 엔트리의 분석 사전, 약 9천 엔트리의 변환 사전과 약 7만 엔트리의 생성 사전으로 구성되어 있다. 일반 분석-변환 규칙은 약 4백개로 구성된다 ALKOL의 사전은 대상 영역인 공군 정보 분야에 맞도록 특정 영역화되어 있다.

본 연구에서는 ALKOL의 번역 결과를 영한 번역 전문가 수준의 3명에게 제시하여 이해도를 기준으로 번역 결과에 대한 평가를 수행하였다. 이해도 평가 기준은 아래 표 1와 같다.

표 1: 이해도 평가 기준

단계	수준
1	문장의 뜻이 명확하여, 어떤 수정도 할 필요가 없다
2	문장의 의미는 파악되지만, 다소 수정이 필요하다
3	전체적인 문장의 뜻은 파악이 되나, 세부적인 이해에는 전문가의 도움이 필요하다
4	문법,용언 용법 등에 문제가 많아, 문장을 겨우 이해하거나 거의 이해할 수 없다
5	문장을 전혀 이해할 수 없다

평가에 사용된 문장은 ALKOL이 대상 영역으로 하고 있는 공군 정보 분야의 전문 잡지 Jane's Intelligence Review의 1,000문장이며, 사용자 인터페이스의 복수 번역 선택 기능을 이용하여 가장 좋은 번역 결과를 선택한 뒤 평가자에게 제시했다. 문장은 평균 15단어로 구성된다.

평가 결과에서는 1단계로 평가된 번역 결과, 즉 번역결과에 대한 수정이 필요없는 정확한 번역 결과가 48.16%, 사용자가 번역 결과를 이해할 수 있는 수준인 3단계 이상의 문장이 72.81%였다.

7 결론

본 논문에서는 공군 정보 분야를 대상으로 구축된 영한 기계 번역 시스템 ALKOL에 대해서 소개하였다. ALKOL은 어휘화된 규칙에 기반한 번역 시스템으로 어휘 특성을 반영할 수 있으므로 정확하고 자연스러운 번역 결과를 낼 수 있다. 어휘화된 규칙은 어휘-분석-변환-생성 네 단계의 연결된 구조로 사전에 저장되므로 사전 검색에 드는 시간을 줄여 번역 과정의 효율성을 높인다. 형태소 분석, 태깅, 구문 분석, 변환, 생성의 각 번역 단계에서는 대상 영역의 특성을 처리할 수 있게 전/후처리를 강화하여 정확도를 높였으며, 구문 분석 전처리 단계를 도입하여 분석 단계의 복잡도를 낮춰 번역 과정의 효율성과 정확성을 높일 수 있었다. 각 단계에서는 하나 이상의 결과를 다음 단계로 전달하여 문장 단위와 구 단위의 하나 이상의 번역 결과를 사용자에게 제시하여 번역 결과를 유연하게 이용할 수 있게 했다. 또한 사용자의 편의를 고려한 사용자 인터페이스를 구축하여 사용자의 번역 작업의 능률성을 높일 수 있었다.

사전에 기반한 구문 변환 방식 번역 시스템에서는 사전 개발에 드는 시간과 비용이 개발 과정에서 발생하는 가장 큰 문제점 중의 하나다. 이를 해결하기 위해 번역 사전의 자동-반자동 구축에 대한 연구가 진행 중이다. 향후 연구로는 시스템의 속도 문제 개선, 한영 번역시스템의 개발 등이 있다.

참고문헌

[1] 임철수, 이현아, 최명석, 장병규, 이공주, 김길창, "어휘화된 규칙에 기반한 영한 기계번역 시스템," 한국정보과학회 추계학술대회 발표 논문집, 1997.

[2] 이현아, 이공주, 김길창, "자연스러운 번역을 위한 두단계 영한 변환 시스템," 한국정보과학회 추계학술대회 발표 논문집, 1997

[3] 이성욱, 이공주, 서정연, "영한 기계 번역 품사 집합과 펜트리뱅크 코퍼스 품사 집합간의 품사 대응," 한국정보과학회 추계학술대회 발표 논문집, 1999.

[4] 노윤형, "대상 영역 코퍼스를 이용한 번역 사전의 특정 영역화를 위한 워크벤치," 한국과학기술원 전산학과 석사학위논문, 2000

[5] 이현아, 장병규, 강인호, 이신목, 김길창, "분산 환경에서의 번역 시스템 개발 : 사전 개발과 테스트 환경을 중심으로," 제 12회 한글 및 한국어 정보처리 학술대회, 2000.

[6] 한국과학기술원, "군사정보 기계번역 시스템 개발 4차년도 최종보고서," 2000.

[7] 이현아, 임철수, 김길창, "영한 기계 번역 사전 작성 요령," 한국과학기술원 전산학과 *Technical Report, CS/TR-98-126*, 1998.

[8] 합산컴퓨터, "아르미 사용 설명서," 2000.

[9] W J Hutchins, H L Somers, "An introduction to Machine Translation," *Academic Press*, 1992

[10] Reynar, J.C. and Ratnaparkhi, "A Maximum Entropy Approach to Identifying Sentence Boundaries," In *Proceeding of the Fifth Conference on Applied Natural Language Processing*, 1997.

[11] German Rigau, Jordi Atserias and Eneko Agirre, "Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation." In *Proceeding of the 35'th Annual Meeting of the Association for Computational Linguistics*, 1997