문틀기반 영한 자동번역 시스템

최승권, 서광준, 김영길, 서영애, 노윤형, 이현근
한국전자통신연구원, 지식처리연구팀
{choisk, seokj, kimyk, yaseo, yhnoh, lhk62698}@etri.re.kr

Sentence-Frame based English-to-Korean Machine Translation

Sung-Kwon Choi, Kwang-Jun Seo, Young-Kil Kim, Young-Ae Seo, Yoon-Hyung Roh, HyunKeun Lee
Knowledge Processing Research Team. ETRI

요 약

국내에서 영한 자동번역 시스템을 1985 년부터 개발한 지 벌써 15 년이 흐르고 있다. 15 년의 영한 자동번역 기술개발에도 불구하고 아직도 영한 자동번역 시스템의 번역품질은 40%를 넘지 못하고 있다. 이렇게 번역품질이 낮은 이유는 다음과 같이 요약할 수 있을 것이다.

- 입력문에 대해 파싱할 때 오른쪽 경계를 잘못 인식함으로써 구조적 모호성의 발생문제: 예를 들어 등위 접속절에서 오른쪽 등위절이 등위 접속절에 포함되는 지의 모호성.
- 번역 단위로써 전체 문장을 대상으로 한 번역패턴이 아닌 구나 절과 같은 부분적인 번역패턴으로 인한 문장 전체의 잘못된 번역 결과 발생.
- 점차 증가하는 대용량 번역지식의 구축과 관련해 새로 구축되는 번역 지식과 기구축된 대용량 번역지식들 간의 상호 충돌로 인한 번역 품질의 저하.

이러한 심각한 원인들을 극복하기 위해 본 논문에서는 문틀에 기반한 새로운 영한 자동번역 방법론을 소개하고자 한다. 이 문틀에 기반한 영한 자동번역 방법론은 현재 CNN 뉴스 방송 자막을 대상으로 한 영한 자동번역 시스템에서 실제 활용되고 있다. 이 방법론은 기본적으로 data-driven 방법론에 속한다. 문틀기반 자동번역 방법론은 규칙기반 자동번역 방법론보다는 낮은 단계에서 예제기반 자동번역 방법론 보다는 높은 단계에서 번역을 하는 번역방법론이다. 이 방법론은 영한 자동번역에 뿐만 아니라 다른 언어쌍에서의 번역에도 적용할 수 있을 것이다

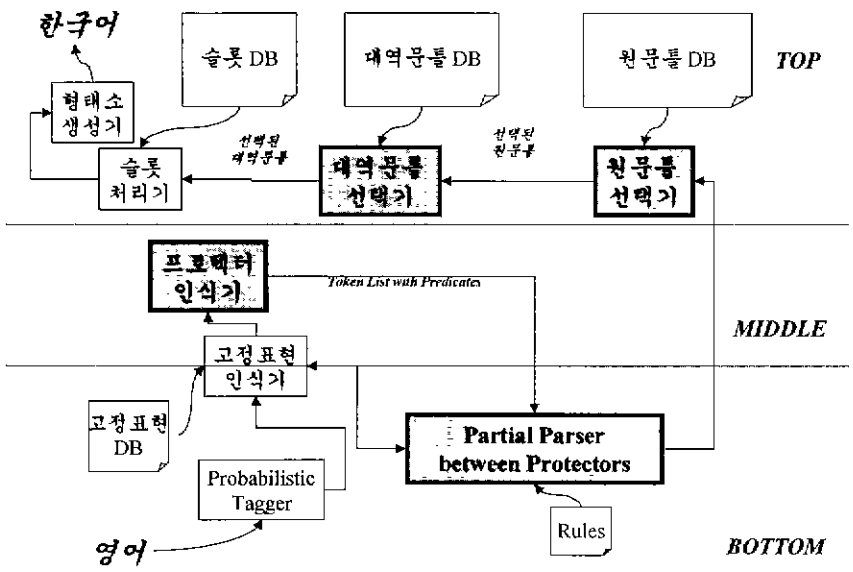
1. 서 론

언어장벽을 해결하기 위해 영한, 일한과 같은 자동번역 시스템이 1985 년부터 국내에서 개발되어 왔었다. 이중에서 영한 자동번역 시스템들은 대부분 문법적으로 정형적인 텍스트만을 대상으로 번역을 시도하여 왔는데 이들 모두는 언어학적 규칙기반 자동번역 방법론들이었다.

이와 같은 언어학적 규칙기반 영한 자동번역 시스템들은 가까운 장래에 쉽게 해결하기 어려운 다음과 같은 문제점들을 가지고 있었다: [3] :

- 비연속적 관용표현의 처리 문제.
- 형태소 모호성이나 구조적 모호성에 대해 처리 문제.
- 장문이나 비문법적 문장에 대한 robust 처리 문제.
- 대역어 선정 문제

이러한 문제점들은 결국에는 번역품질을 저하시키는 원인이 된다[3]. 이런 문제점들 중 몇몇은 최근에 자연언어처리기술이 더욱 발



[그림 1] 시스템 구성도

전하면서 해결되었었다[4].

- 비연속적 관용표현을 처리하기 위한 "복합단위 인식기" 개발[2]
- 형태소 모호성 해결을 위한 "후처리 규칙을 가진 신경망 태거" 개발[6]
- 장문이나 비문법적 문장의 robust 한 처리를 위한 "장문분절기와 실패완화기" 개발.[2]
- 대역어 선정을 위한 "대용량 공기사전과 어휘규칙" 개발.[1]

이러한 그 동안의 기술적 개발에도 불구하고 영한 자동번역 시스템의 번역 품질은 아직도 40%를 넘지 못하고 있다. 이러한 근본적인 원인을 요약하면 다음과 같다:

- 입력문에 대해 파싱할 때 오른쪽 경계를 잘못 인식함으로써 구조적 모호성의 발생문제: 예를 들어 등위 접속절에서 오른쪽 등위절이 등위 접속절에 포함되는 지의 모호성.
- 번역 단위로써 전체 문장을 대상으로 한 번역패턴이 아닌 구나 절과 같은 부분적인 번역패턴으로 인한 문장 전체의 잘못된 번역 결과 발생.
- 점차 증가하는 대용량 번역지식의 구축과 관련해 새로 구축되는 번역 지식과 기구축된 대용량 번역지식들 간의 상호 충돌로 인한 번역 품질의 저하.

이러한 심각한 원인들을 극복하기 위해 본 논문에서는 문틀에 기반한 새로운 영한 자동번역 방법론을 소개하고자 한다. 이 문틀에 기반한 영한 자동번역 방법론은 현재 CNN 뉴스 방송 자막을 대상으로 한 영한 자동번역 시스템에서 실제 활용되고 있다. 이 방법론은 영한 자동번역에 뿐만 아니라 다른 언어쌍에서의 번역에도 적용할 수 있을 것이다..

2. 시스템 구성

문틀기반 자동번역 방법론은 기본적으로 data-driven 방법론에 속한다. 문틀기반 자동번역 방법론은 규칙기반 자동번역 방법론보다는 낮은 단계에서, 예제기반 자동번역 방법론 보다는 높은 단계에서 번역을 하는 번역 방법론이다. 그림 1 이 문틀기반 자동번역 시스템 "CaptionEye/EK"의 시스템 구성도이다.

문틀기반 영한 자동번역 시스템에서 영어 입력문은 우선 형태소 분석과 태깅을 Partial Parser between Protectors 에 의해 구단위의 분석이 행해진다. 이 구단위의 분석 결과가 영어 원문에 대한 원문틀이 되며 이 영어 원문틀은 한국어의 어순이 반영되어 있는 해당 한국어 대역문틀로 전환되며 전환된 대역문틀은 다시 대역문틀을 구성하는 각 영어 슬롯이 Slot Pattern Processor 에 의해 한국어 슬롯으로 전화되어 형태소 생성기를 거치면 한국어가 생성된다.

이러한 문틀기반 영한 번역 절차는 다음과 같은 수동 번역 시뮬레이션에 의해 더욱 자

세히 설명될 것이다:

[입력문]
The White House said the president decided to grant duty-free status for 18 categories, but turned down such treatment for other types.

[고정표현인식]
(det:determiner, prep:preposition, num:number, adj:adjective,punct punctuation,conj:conjunction)

The(det) White House(noun) said(verb) the(det) president(noun) decided to grant(verb) duty-free(noun) status(noun) for(prep) 18(num) categories(noun) ,(punct) but(conj) turned down(verb) such(adj) treatment(noun) for(prep) other(adj) types(noun)

[프로텍터인식]
det noun verb det noun verb noun noun prep noun noun punct conj verb adj noun prep adj noun

[Partial Parsing between Protectors]
(det noun => NP) verb (det noun => NP) verb (noun noun prep num noun => NP) punct conj verb (adj noun prep adj noun => NP PP)

[원문틀 선택]
nVnVnPCVnp(= NP1 verb1 NP2 verb2 NP3 punct conj verb3 NP4 PP)

[대역문틀 선택]
NP1 verb1 NP2 verb2 NP3 punct conj verb3 NP4 PP => NP1은 NP2이 NP3을 verb2다고 verb1다 punct conj PP NP4는 verb3

[슬롯 처리]
(det noun => noun)은 (det noun => noun)이 (noun1 noun2 prep num noun3 => num의 noun3 prep noun1 noun2)을 verb2다고 verb1다 punct conj (prep adj noun => adj noun prep) (adj noun => adj noun)는 verb3

[출력문]
백악관은 대통령이 18개의 항목을 위한 세금 면재 상태를 인정하기로 결정했다고 말했다, 그러나 다른 종류에 대하여 그러한 처리는 거절했다.

그림 1 에서 검게 칠해진 부분이 다음장에서 상세히 설명될 것이다

3. 프로텍터 인식기

규칙기반 영한자동번역에서 번역실패의 주된 원인 중에 하나는 파싱할 때 오른쪽 경계의 잘못된 인식으로 구조적 모호성이 대량으로 발생하기 때문이다. 이러한 예는 등위접속절에서 등위접속사에 묶인 오른쪽 등위절이 실제로 등위접속절의 성분인지 아니면 완전히 독립된 성분인지가 잘 인식이 되지 않는 에이다 파싱에서의 이러한 구조적 모호성을 줄일 수 있기 위해 문틀기반 자동번역에서 프로텍터(Protector)라는 개념이 만들어졌다. 프로텍터는 구조분석에서 구조적 모호성을 야기시키는 언어학적 품사들을 말하는 것으로 영어에서는 동사, 접속사, 기호가 프로텍터에 속한다.

4. PARTIAL PARSER BETWEEN PROTECTORS(=PBP)

Partial parser between protectors 라는 것은 말 그대로 프로텍터 사이에 있는 형태소 단위의 연속된 품사들을 대상으로 하여 구로 reduce 시키는 partial parser 를 말한다. 이 파서의 목적은 문장 전체에 대해 구조적 모호성을 줄이고 문장 성분들간의 기능을 문장 단위에서 파악하기 위함이다.
Partial parser between protectors 는 확장된 문맥자유문법(augmented context-free grammar)의 형식을 가지며 규칙기술, 조건부, 실행부로 이루어진다. 이 Partial parser between protectors 의 BNF 를 기술하면 다음과 같다:
(lpro: left protector, rpro:right protector, epos: English part-of-speech)

```
<pbps> := (<pbp>)+ #
<pbp> ::= <morph_nodes> (=> <phrase_nodes>
          <condition> <action>)+
<morph_nodes> ::= ( <morph_node> )+
<phrase_nodes> .:= ( <phrase_node> )+
<morph_node> ::= (noun | ... |prep)
<phrase_node> ::= (np | ... | pp )
<condition> ::= { ( <cond> ; )* }
<cond> ::= ( <condexpr> ) |
                _AND ( <cond> , <cond> ) |
                _OR ( <cond> , <cond> )
<condexpr> ::= <condfeature><ops>
                ( <condfeature> | <value> )
<condfeature> ::= ( lpro | rpro )([<num>]) :
                    epos == [
                    (<verb>|<conj>|<punct>) ]
<action> ::= { (<assign> ; )* }
<assign> ::= <rhsfeature> := ( <lhsfeature> |
            | <value> )
<rhsfeature>::= rhs[<num>]: <value>
<lhsfeature>::=lhs[<num>](: <value>)
```

다음은 2 절에서 제시된 예문 'The White House said the president decided to grant duty-free status for 18 categories, but turned down such treatment for other types'에서 'such treatment for other types'라는 구가 NP PP 로 reduce 되는 partial parser between protectors 의 규칙이다.

```
adj noun prep adj noun => NP
{ }
{ rhs[1]:start := lhs[1];
  rhs[1]:end := lhs[5];
  rhs[1]:etype := lhs[2]:etype; }
                        -> NP PP
{_AND(lpro:epos==[verb],rpro:epos==[illegal]);}
{ rhs[1]:start := lhs[1];
  rhs[1]:end := lhs[2];
  rhs[2]:start := lhs[3];
  rhs[2]:end := lhs[5];
  rhs[1]:etype := lhs[2]:etype,
  rhs[2]:etype := lhs[5] etype;}
```

위의 규칙은 adj noun prep adj noun.의 형태소 품사들은 좌프록텍터가 verb 이고 우프로텍터가 문장의 종결기호일 때 NP PP 로 reduce 된다는 context-sensitive 규칙을 말한다.

5. 원문틀 선택기

원문틀이란 영어 원문에 대한 partial parser between protectors 의 분석결과를 의미하는 것으로서 프로텍터와 슬롯으로 구성된다.

5.1. 원문틀 DB

원문틀 DB 의 key-word 는 DB 메모리를 줄이고 readability 를 증가시키기 위해 각 프로텍터와 슬롯에 대해 encoded form 를 부여하여 만들었다. 이들의 목록을 보이면 다음과 같다:

[도표 1] 프로텍터와 슬롯의 Encoded name

품사	Code	
CONJ(CONJunction)	C	프로텍터
VERB(VERB)	V	
PUNCT(PUNCTuation)	T	
ADP(ADverbialPhrase)	v	슬롯
AP(Adjective Phrase)	a	
DETP(DETerminer Phrase)	d	
NP(Noun Phrase)	n	
PP(Prepositional Phrase)	p	
IPREP(IsolatedPREPpositon)	i	
SP(SentencialPhrase)	s	

원문틀 DB 에 입력되어 있는 임의의 엔트리에 대한 예를 들면 다음과 같다:

Example: I love you
Key-Word: nVn
Content: S-nVn

5.2. 원문틀 선택

Partial Parser between protectors 에 의한 분석결과는 문장단위로 구축된 원문틀 DB 에서 일치하는 원문틀과 일치할 때만 선택된다.

6. 대역문틀 선택기

대역문틀 선택기는 크게 두가지로 문틀 선택을 순차적으로 행하고 있다. 우선 원문틀에 대한 제약조건에 의한 문틀선택이 이루어지고 난 후 대역문틀에 대한 가중치 부여에 의한 선택으로 이루어진다. 이들은 대역문틀 선택 절에서 상세히 기술될 것이다.

6.1. 대역문틀 DB

6.1.1. 제약조건 포함 원문틀 DB

대역문틀 DB 에 있는 원문틀은 제약조건이 자질로써 부여되어 있다. 이들은 모두 프로텍터에 주어져 있으며 그들의 유형은 각 프로텍터에 대해 eform (English morphological form information) 과 etype (English syntactic type information) 로 구성된다. 예를 들어 I love you 라는 문장에 대한 제약조건이 부여된 원문틀의 예를 보이면 다음과 같다:

Example: I love you
Key-Word: S-nVn
Content: { NP verb:[t1,vb] NP2} T-nnV3
{ NP verb:[t1,vg] NP2} T-nVn1
{ NP verb:[t1,vn] NP2} T-nVn2

위의 예는 원문틀의 key-word 인 S-nVn 이 프로텍터인 동사에 etype:t1(타동사)과 eform:vb (서술문형)를 부여하면 대역문틀은 T-nnV3 가 되어야 한다는 것이다.

6.1.2. *대역문틀 DB*

대역문틀 DB 는 제약조건이 프로텍터 뿐만 아니라 슬롯에도 부여되어 있는 대역문틀 원문부와 그것의 한국어 대역부로 구성된다. 한국어 대역부는 한국어 어순과 한국어의

언어학적 정보가 자질로써 부여되어 있는 형태이다. 'I love you'라는 문장에 대한 대역 문틀 DB 는 다음과 같다:

Example: I love you
Key-Word: T-nnV3
Content: {NP1:[etype ** [demo]] VERB1!:[etype == [t1]] NP2:[etype ** [demo]] } -> { NP1:[kcase := [topic]] NP2:[kcase := [obj]] VERB1! }

위의 대역문틀은 VERB1 이 타동사이며 주어와 목적어가 지시대명사일 때 영어 어순 NP1(주어) VERB1 NP2(목적어)이 NP1(주어) NP2(목적어) VERB1 로 변형된다는 것을 말한다.

6.2. 대역문틀 선택

하나의 대역문틀 영어 원문부에 대해 여러 개의 한국어 대역부가 일치할 수가 있다. 이 때 더욱 올바른 것으로 일치시키기 위해서 우리는 다음과 같은 대역문틀 선택 원칙을 세웠다

- 단계별로 prunning하지 않고 가능한 정보를 최대한 수집한 후 가중치로서 선택함으로써 정확도를 향상함.
- 엔진에서 부여하는 Systemic한 가중치를 이용함으로써 점증성 손상을 최소화함.
- 단순하고 일관성 있는 선택 메커니즘 제공.

그리고 위와 같은 원칙을 바탕으로 다음과 같은 휴리스틱 선택 선호도를 정하였다.(아래로 내려갈수록 더욱 높은 선호도를 가진다)

- 보다 많은 고정표현, 비연속 숙어를 포함한 문틀을 우선한다.
- 구문 슬롯의 개수가 적은 문틀을 우선한다. 즉, 구문 트리의 깊이가 얕은 경우를 우선한다.
- 보다 많은 제약조건을 만족한 문틀을 우선한다.
- 보다 많은 어휘 제약을 만족한 문틀을 우선한다.
- 대역어가 자연스러운 문장을 우선한다.
- 대역어 선정 과정에서 계산된 가중치 반영

7. 예측되는 번역율

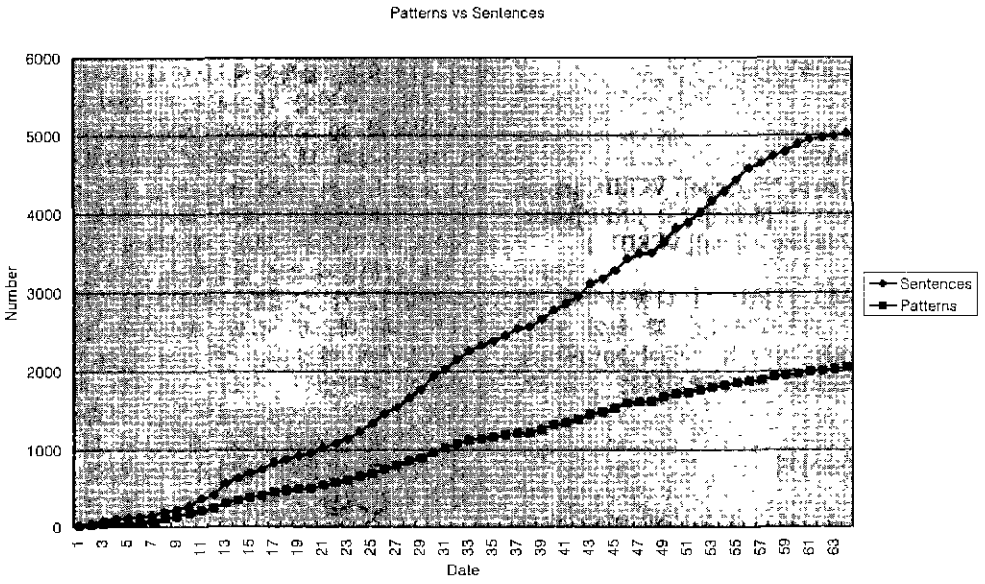
우리는 우리의 문틀기반 자동번역 방법론의 타당성을 조사하기 위해 실제 CNN 뉴스 방송자막을 수집한 후에 이들을 가지고 향후의 번역율을 예측하여 보았다. 우선 수집된 데이터는 모집합 데이터로써 1999 년 3 월 25 일부터 30 일까지의 CNN 뉴스 방송자막 3,230 문장과 테스트 집합 데이타로써 1999 년 4 월 1 일의 CNN 뉴스 방송자막 194 문장을 수집하였다. 그리고 나서 이들 모집합 데이터와 테스트 집합 데이터에 대해 각각 3,718 원문틀과 275 원문틀을 반자동으로 구축하였다. 각 모집합 데이터의 3,718 개의 원문틀과 테스트 집합 데이터의 275 개의 원문틀간의 매칭율은 73 개로써 26.5%의 매칭율이 나왔었다. 우리는 이러한 매칭율로부터 앞으로 대량의 원문틀을 구축하게 되면 임의의 문장에 대해 매칭율이 더욱 높아질 것이며 더불어 번역율도 높아질 것으로 예측하고 있다.

8. 실 험

우리는 우리의 방법론에 의해 2 달 동안에 직접 구축된 자료를 토대로 번역실험을 하여 보았다. 이 기간동안 8 단어 이상으로 된 CNN 뉴스 방송자막 문장 5,024 문장에 대해 2,045 문틀을 구축하였었다. 여기서 영어 문장 대비 구축된 문틀간에 차이가 나는 2,979 문장은 기구축된 문틀에 일치함으로써 완전 자동으로 번역된 문장을 의미한다. 이 차이가 커지면 커질수록 매칭율이 높아질 것이다. 그림 2 가 두달동안 구축된 문틀에 의한 번역결과를 나타낸다.
그림 2 에서 상위에 있는 선이 입력된 문장수를 나타내며 아래의 선이 입력 문장에 대한 구축된 문틀수를 나타낸다. 그림 2 에 따르면 문틀기반 영한 자동번역기의 번역율은 약 41%에 달하고 있으며 이것은 문틀 수가 증가할수록 점증적으로 올라갈 것으로 예측된다.

9. 결론

본 논문에서는 CNN 뉴스 방송자막을 대상으로 한 영한 자동번역 시스템에서의 번역 방법론으로써 문틀기반 자동번역 방법론이 기술되었다. 여기서 문틀기반 자동번역의 핵심적인 역할을 하는 프로택터는 영어 입력문의 구조 분석에서 모호성을 해결하기 위해 도입되었으며 영어에서 프로택터는 동사, 접속사, 기호라는 것이 설명되었다. 문틀은 입력문장 전체를 대상으로 구축되며 이것은 원문틀과 대역문틀로 구성된다는 것이 설명되었다.



[그림 2] 8 단어 이상의 입력문 대비 문틀간의 매칭 관계

문틀기반 자동번역 방법론은 기존의 영한 자동번역 시스템 뿐만 아니라 대부분의 자동번역 시스템들이 극복하여야만 하였던 여러가지 문제점들, 예를 들어 오른쪽 경계 모호성으로 야기되는 구조적 모호성 문제, 부분 번역지식 구축으로 인해 전체 번역이 잘못되는 문제, 대용량의 번역규칙 구축시에 나타나는 번역규칙들 간의 충돌문제를 상당히 해결하였다.

현재 문틀기반 자동번역 시스템인 CaptionEye/EK 는 계속해서 진행중에 있다. 앞으로 문틀기반 자동번역 시스템이 더 좋은 품질의 번역을 이루기 위해 하여야 할 것은 번역사전의 엔트리를 증가시키고, 대용량의 문틀을 구축하여야 한다. 또한 영어 장문에 대한 처리가 보완되어야 한다

REFERENCES

[1] 심철민, 최승권, 여상화(1998) "기계 번역을 위한 문법 기술 언어의 확장" 한국정보과학회 추계학술대회.

[2] 정한민, 최승권, 김영길, 심철민(1998) " 두 단계 구문 규칙을 이용한 후-실패 완화 기법" 한국정보과학회 추계학술대회.

[3] Choi K.S., Lee S.M., Kim H.G., and Kim D.B. (1994) *An English-to-Korean Machine Translator: MATES/EK.* COLING94, pp. 129-133.

[4] Sung-Kwon Choi, Taewan Kim, Sang-Hwa Yuh, Han-min Jung, Chul-Min Sim, Sangkyu Park(1999) "English-to-Korean Web Translator: "FromTo/Web-EK" ", MTSUMMIT99, Singapore.

[5] Hutchins W.J. and Somers H.L. (1992) *An Introduction to Machine Translation.* Academic Press.

[6] Sanghwa Yuh, Hanmin Jung(1999) *NeuTag*: A Hybrid Neural Network English Tagger with Pre-Fail Softener. ICCPOL.