언어 유형론에 기반한 다국어 공용 번역지식의 구축

최승권 김태완 박동인
시스템공학연구소 자연어정보처리연구부

A Construction of Multilingual Linguistic Translation Knowledge based on the Language Typology

Sung-Kwon Choi, Taewan Kim, Dong-In Park
Systems Engineering Research Institute, Machine Translation Lab.

요 약

본 논문은 다국어 자동번역시스템에서 다국어의 번역지식(사전,규칙,정보)구축을 위해 언어유형론을 도입하는 방법론을 제안한다.

다국어 번역지식의 구축과 관련하여 기존 다국어 자동번역 시스템들에서 항상 문제가 되고 있는 것은 번역지식의 구축, 관리, 재활용의 문제이다. 번역지식의 구축은 다국어를 위한 번역지식의 크기, 다국어의 수용정도와 관련되며, 번역지식의 관리는 번역지식의 단순화 정도와 관련되며, 번역지식의 재활용은 기존에 구축된 번역지식을 새로운 언어들에 재사용 정도와 관련된다. 이러한 문제점들을 해결하기 위해 본 논문에서는 한국어를 포함한 다국어의 언어 친족성에 따라 번역지식을 공유하도록 하는 언어유형론에 기반한 다국어 공용 번역지식 구축 방법론을 제안하고자 한다.

1. 서론

기존의 변환기반 다국어 자동번역 시스템들(SYSTRAN, EUROTRA, METAL, LOGOS, GETA 등)은 언어들간에 언어 유형론적인 공통성이 있음에도 불구하고 이를 수용하지 못하고 있다. 이때문에 기존의 변환기반 다국어 자동번역 시스템들은 단순히 양언어간 자동번역 시스템을 열거하여 놓은 모습을 가지게 되었으며 전체 시스템의 크기를 증가시키는 결과를 초래하게 되었다. 다국어 자동번역에서 변환 과정의 수를 줄이기 위해 피봇방식을 추구하는 다국어 자동번역 시스템이 있지만(CETA, SALAT, DLT, KANT 등) 언어학적 보편성 모델을 완성하기 어렵기 때문에 진정한 피봇방식의 자동번역 시스템 구현이 이루어지지 못하고 있다[Lewis 1992].

이런 관점에서 본 논문에서는 기존의 변환기반 다국어 자동번역 시스템들이 지녔던 번역지식의 크기와 피봇방식의 언어보편적인 번역지식의 구축 어려움을 극복하기 위해 언어유형론에 기반한 다국어 공용 번역지식 구축

방법을 제안하고자 한다. 여기서 언어유형론에 기반한 다국어 공용 번역지식의 구축이란 언어유형론에 따라 형성된 언어군의 공통적 언어현상을 그 언어군이 공유하는 번역지식으로 구성하여 번역지식을 모듈화한다는 것을 의미한다. 언어유형론에 기반한 다국어 공용 번역지식은 기존의 타 다국어 자동번역시스템들과 비교해 다음과 같은 장점을 가질 수 있다.

- *번역지식 크기의 축소*
  언어군에서 공유되는 번역지식이 단지 한번 호출(load)되어 해당 언어군에서 사용되기 때문에 다국어 번역시스템의 전체 메모리 크기를 줄일 수 있다.
- *번역지식 관리의 편리성*
  언어군에서 사용되는 번역규칙이나 번역정보를 작성, 수정하는데 일관성을 유지할 수 있고 관리가 편리하다.
- *번역지식의 재활용*
  새로운 언어를 첨가하여 기존의 언어들과 번역하고자 할 때에 언어친족성이 가장 가까운 언어가 사용한 번역지식을 새 언어에서도 전체 혹은 부분적으로 공유하게 함으로써 번역지식을 새로 작성하는 수고로움을 덜 수 있다.

본 논문은 다음과 같이 구성된다. 2 장에서는 언어유형론에 기반한 다국어 공용 번역지식에 의한 변환기반 다국어 자동번역 시스템의 전체적인 구성도를 소개하고자 한다. 3 장에서는 언어유형론에 기반한 다국어 공용 번역지식을 구성하기 위한 문법모델링을 기술하고자 하며 4 장에서는 언어유형론에 기반한 다국어 공용 번역지식을 구축하기 위한 언어유형론 파라미터 번역지식을 소개하고자 한다. 5 장에서는 다국어 공용 번역지식의 실험 결과를 기술할 것이다.

2. 시스템 구성

언어유형론에 기반한 다국어 공용 번역지식에 의한 변환기반 다국어 자동번역 시스템의 전체적인 구성은 그림 1 과 같다:

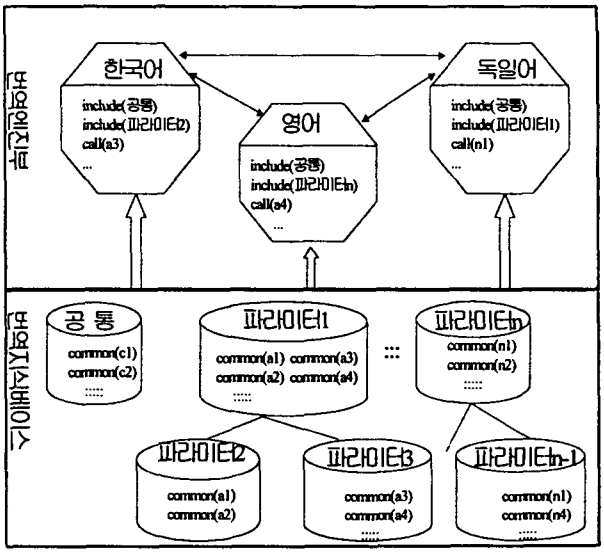

그림 1. 언어유형론기반 다국어자동번역 시스템 전체 구성도

그림 1 은 언어유형론기반 다국어자동번역시스템의 번역지식베이스와 번역엔진부를 보여주고 있다. 번역지식베이스에서 원기둥들은 언어유형에 따라 형성된 번역지식군을 나타내며 원기둥의 뚜껑부분에 있는 '공통'과 '파라미터 n'은 번역지식군의 파일이름을 나타낸다. 번역지식군간에는 계층적 상속관계를 형성할 수 있으며 이것은 그림 1 에서 사선으로 그려져 있다. 번역엔진부에 있는 각 언어는 자신의 번역지식을 번역지식베이스에 있는 번역지식군의 이름을 불러서 만드는데 이때 번역지식베이스와 번역엔진부를 연결해 주는 것이 common, call, include 와 같은 명령지시어이다. common, call, include 은 각각 다음의 의미를 가진다:

지시어	의미
common	공유번역지식 내용을 정의하는 명령지시어
call	common 명령지시어의 공유번역지식 내용을

	호출하는 명령지시어
include	공유번역지식군의 파일내용을 호출하는 명령지시어

도표 1: 명령지시어와 의미

이러한 명령지시어를 사용하여 각 언어는 번역지식베이스에 있는 번역지식군을 가지고 자신의 번역지식을 형성하는데 예를 들어 그림 1 에서 한국어의 번역지식은 다언어 공통번역지식군 'include(공통)'과 파라미터화된 번역지식군 'include(파라미터 2)' 그리고 파라미터화된 번역지식군 파라미터 3 으로부터 호출되는 개별 번역지식 'call(a3)'등으로 구성되어 영어나 독일어의 번역지식과는 다른 번역지식을 구성하는 것을 알 수 있다.

3. 문법 모델링

본 장에서는 다국어 공용 번역지식을 구성하기 위한 프레임으로써의 문법모델링을 소개하고자 한다.

3.1. 이분법 중심의 문법규칙

다국어 자동번역 시스템에서 다국어를 처리하기 위한 공용문법규칙은 무엇보다 각국 언어들의 개별언어현상에 대한 충분한 설명이 가능해야 하며 또한 다국어 언어현상의 기술이 간편하며 편리하게 문법가에 의해 작성될 수 있어야 한다. 즉 영어나 불어와 같은 구성적(configurational) 언어뿐만 아니라 어순이 비교적 자유로운 한국어나 일본어, 독일어와 같은 비구성적(non-configurational) 언어들의 언어 현상을 문법적으로 올바르게 설명할 수 있어야 한다. 이러한 새로운 문법형식을 위해 본 논문에서는 X-bar 통사 이론[Jackendoff 1977]의 구구조 형성 개념과 HPSG[Pollard 1994]의 정보전달 원칙을 혼합한 문법형식을 제안한다. 이 혼합형 문법형식은 삼분법인 접속어구 규칙이외에는 모두 이분법 구조의 규칙으로 이루어지며 정보는 머리어(head)나 기능어(functional word)에 의해 위로 상승된다. 이상의 문법형식에 근거한 문법규칙은 다음과 같다:

- 논항구조형성규칙
  {head:HEAD}.[{head:HEAD, frame:({arg1:ARG}; {arg2:ARG}; {arg3:ARG}; {arg4:ARG4})}, ARG]. [1]
- 수식어구조형성규칙
  {head:HEAD}.[{role:mod,head:{restr:RESTR}}, {head:HEAD}>>RESTR].
- 기능어구조형성규칙
  {head:HEAD}.[{role:funct,head:HEAD, frame:({arg1:ARG}; {arg2:ARG})}, ARG].

위의 문법규칙들은 분석시 언어현상의 두 구성요소만을 고려하기 때문에 기술하기가 편리하며 정보이동이 매우 단순하다는 장점을 가진다.

3.2. 정보구조

다국어 자동번역을 위한 정보구조를 좀더 일관되게 입력.관리.수정하기 위해 다국어용 정보구조를 작성하는 것이 필요하다. 사전 및 노드의 정보구조는 가능한 한 모든 언어학적 정보를 표현해야 하며 이러한 정보를 수월하게 이동시키기 위해 단층이 아닌 다층적인 구조를 형성하는 것이 바람직하다. 이런 이유에서 다국어용 정보구조로 자질구조(feature structure)를 채택하였으며 속성은 여러언어에서 동일하도록 정의하였다.
정보구조는 크게 6 개의 층위로 구분되는데 노드를 분류하기 위한 분류정보, 음운을 나타내는 음운정보, 형태소 정보를 위한 형태소 정보, 통사.의미관계정보를 등재하는 머리정보와 확장된 머리정보, 논항구조정보, 문맥에 관한 사항을 기입하는 문맥정보로 나누어진다. 이들의 정보구

[1] 규칙은 '규칙이름={상위교점}.[{하위교점 1},{하위교점 2}]'. 형식으로 구성되며 '>>'는 unification constraint 를 위한 operator 로써 이 operator 의 오른쪽 요소는 왼쪽 요소가 완전히 충족되었을 때만 성공한다는 것을 의미한다.

조와 정보종류는 그림 2 에서와 같다:

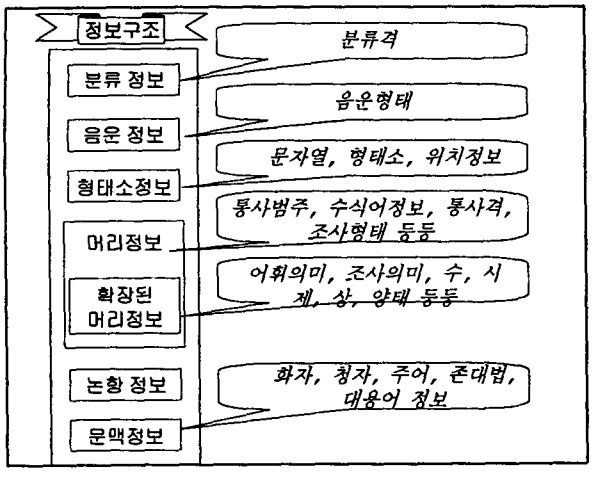

그림 2: 정보구조

위의 정보구조는 속성:값의 자질구조로 이루어지며 값은 다시금 자질구조를 형성하기 때문에 정보의 모듈화가 이루어지고 있다. 모든 정보가 채워질 필요는 없고 채워지지 않은 정보는 각 품사별로 존재하는 정보의 default 규칙에 의해 채워진다.

4. 파라미터화된 번역지식

언어보편성과 대별되는 개념이 언어유형론이다. 이 두 개념은 전자가 언어간 공통성을 연구하는 분야인 반면 후자는 언어간 차이를 연구하는 분야로써 서로 상반되는 개념이지만 상호보완관계를 이루고 있다. 왜냐하면 언어간 공통성을 찾기 위해서는 언어간 차이를 알아야 하기 때문이다. 언어유형론에서 언어간 친족성을 판별할 수 있는 언어학적 척도가 언어유형론 파라미터이며 이 파라미터를 이용하여 형성된 번역규칙이 파라미터화된 번역지식이다. 언어유형론 파라미터는 기존에는 형태론과 통사론에 근거한 두가지 유형으로 제시되었으나[Comrie 1981, Hawkins 1983, Dorr 1993] 본 논문에서는 통사론기반 파라미터를 통사의미론 파라미터로 확장하고 화용론기반 파라미터를 덧붙여 언어유형론 파라미터를 세가지 유형으로 확대하여 각각의 파라미터화된 번역지식을 기술하고자 한다. 이렇게 기술된 파라미터화된 번역지식은 언어유형에 따라 공유하게 된다.

4.1. 형태론기반 파라미터와 파라미터화된 지식

형태론기반 파라미터는 형태소간의 결합이나 굴절현상에 따라 친족언어군을 구별하는 요소로서 고립어(예:영어,베트남어 등), 첨가어(예:한국어,일본어,터어키어 등), 굴절어(러시아어 등)의 유형분류를 만들고 있다. 이러한 형태론에 근거한 언어유형분류는 다음과 같은 형태론기반 파라미터로 설명된다:

- 형태소분리결합: 기능형태소와 어휘형태소간의 형태론적 결합관계의 여부를 알려주는 파라미터.
- 형태소음운조합: 기능형태소와 어휘형태소간의 음운형태조합이 있을 경우를 나타내는 파라미터.

이와 같은 형태론기반 파라미터에 의한 파라미터화된 지식은 앞서 기술된 문법모델링을 이용한 문법지식으로 표현될 수 있다. 형태론기반 파라미터화된 문법지식은 다음과 같다:

- 형태소결합규칙={first:FI, last:LA}.[
  {first:FI, last:no},
  {role:funct, first:no, last:LA}].
  예)소년[first:yes, last:no]-은[role:funct, first:no, last:yes]
- 형태소분리규칙={first:yes, last:LA}.[
  {role:funct, first:FI, last:yes},
  {first:yes, last:LA}]
  예)the[role:funct,first:yes,last:yes] boy[first:yes,last:yes]
- 형태소음운좌조합={phon:PH2}.[{phon:PH1},
  {role:funct, phon:PH2,frame:{arg1:{phon:PH1}}}]
  예)소년[phon:con]-
  은[role:funct,phon:voc,frame:{arg1:{phon:con}}]
- 형태소음운우조합={phon:PH2}.[
  {role:funct,phon:PH2,frame:{arg1:{phon:PH1}}},
  {phon:PH1}].

예)a[role:funct, phon:voc, frame:{arg1:{phon:con}}]
boy[phon:con]

4.2. 통사의미론기반 파라미터와 파라미터화된 지식

통사의미론기반 파라미터는 언어들의 통사-의미론적 친족성을 설명해 주는 파라미터이다. 이 통사의미론기반 파라미터에는 다음과 같은 예들이 있다:

- 어순관계: 머리어와 논항, 머리어와 수식어, 기능어와 기능어 논항간의 위치관계를 나타내는 파라미터.
- 주어선택: 술어의 논항구조에서 주어가 필수적으로 요구되는지 혹은 선택적으로 요구되는 지를 나타내는 파라미터.
- 의미정보상승: 머리어의 위치에 따른 머리어 의미정보의 상위노드로의 상승 파라미터.

이밖에 기능어간의 위치, 서수의 제약범위 등등이 있다. 이와 같은 통사의미론기반 파라미터에 의한 파라미터화된 지식은 앞서 기술된 문법모델링을 이용한 문법지식으로 표현될 수 있다. 통사의미론기반 파라미터화된 문법지식은 다음과 같다:

- 좌술어-우인수구조규칙={head:HEAD}.[
  {head:HEAD, frame:({arg1:ARG}; {arg2:ARG};
  {arg3:ARG};{arg4:ARG4})}, ARG].
  예) 동사-목적어
- 좌인수-우술어구조규칙={head:HEAD}.[ARG ,
  {head:HEAD, frame:({arg1:ARG};
  {arg2:ARG};{arg3:ARG};{arg4:ARG4})}].
  예) 주어-동사(구)
- 좌머리어-우수식어구조규칙=
  {head:HEAD}.[{head:HEAD}>>RESTR,
  {role:mod,head:{restr:RESTR}}].
  예) 명사-관계절
- 좌수식어-우머리어구조규칙=
  {head:HEAD}.[{role:mod,head:{restr:RESTR}},
  {head:HEAD}>>RESTR].
  예) 관계절-명사(구)

어순관계 파라미터와 관련해 위에 열거된 파라미터화된 번역지식 이외에도 좌기능어-우머리어구조규칙, 좌머리어-우기능어구조규칙, 접속어구구조규칙등이 있다.

- 임의주어선택규칙={head:{cat:verb, voice:active}}>>
  {frame:{arg1:{oblig:no}}}}.[].
  예) 한국어, 일본어 등 주어생략 가능 언어군
- 필수주어선택규칙={head:{cat:verb, voice:active}}>>
  {frame:{arg1:{oblig:yes}}}}.[].
  예) 영어, 독일어 등 주어생략 불가능 언어군

4.3. 화용론기반 파라미터와 파라미터화된 지식

화용론기반 파라미터는 문장단위에서 발생되는 친족언어군간 화용론적 차이를 구별할 수 있는 파라미터들로써 이 파라미터에는 다음과 같은 예들이 있다:

- 존대법일치: 주어-술어에 나타나는 존대 현상의 일치관계를 기술해 주는 파라미터.
- 대용어생성: 언어에 따라 대용어의 형태를 결정하는 파라미터.
- 빈주어생성: 언어에 따라 주어의 생성을 요구하는지 하지 않는 지를 결정하는 파라미터.

이와 같은 화용론기반 파라미터에 의한 파라미터화된 지식은 앞서 기술된 문법모델링을 이용한 문법지식으로 표현될 수 있다. 화용론기반 파라미터화된 문법지식은 다음과 같다:

- 주어존대규칙=
  {head:HEAD}.[{head:{subj:yes}, context:{shon:yes}},
  {head:HEAD} >> {context:{shon:yes}}].
  예) 교수님께서(context:{shon:yes})
  오십니다(context:{shon:yes}).
- 대용어 생성규칙=

{head:{cat:pronoun, subj:yes},
context:{anaphor:{referent:REF, num:NUM}}}
>> {head:{cat:noun, lex:REF, num:NUM}}.[].

예) 한국어, 일본어와 같은 대용어를 명사로 받는 언어군.

이상의 파라미터화된 번역지식들은 언어유형론에 따른 친족언어군의 특성을 기술할 수 있기 때문에 다국어 자동번역시스템에 각 언어군의 언어적 특성을 제공할 수 있다.

5. 실 험

본 실험은 한국어, 영어, 독일어간의 다국어 자동번역에서 언어유형론에 기반한 다국어 공용 번역지식을 이용한 실험 결과[Choi 1995]를 요약한 것이다. 본 시스템은 SICstus-PROLOG 로 구현되었으며 Unix-Workstation 에서 작동하고 있고 분석-변환-생성으로 이루어지는 전형적인 변환기반 다국어 자동번역시스템이다. 각 언어의 분석과 생성 단계는 형태구조단계, 통사구조단계, 의존구조단계 3 단계로 이루어져있다. 언어유형론 기반 다국어공용번역지식을 이용한 다국어 자동번역에서 한국어 분석을 위해 사용한 문장은 국어 문법책에 나오는 언어현상별로 분류된 약 250 문장의 기본문이었으며 이 250 문장을 분석하기 위해 사용한 언어유형론 기반 다국어공용번역지식의 통계치는 다음과 같았다:

	통사분석	의존분석
총 규칙 수	104	55
파라미터 규칙 수	59	41
한국어 특수규칙 수	5	7

도표 2: 250 문에 대한 한국어 파라미터 규칙 수

이상의 한국어를 위한 파라미터화된 번역지식은 언어친족성이 있는 일본어 분석에도 재사용할 수 있을 것이다.

6. 결 론

본 논문에서는 언어유형론에 기반한 다국어 공용 번역지식의 구축에 대해 살펴 보았다. 다국어의 번역지식을 구축하는 방법으로써 언어유형론에 입각한 다국어 공용 파라미터화된 번역지식의 구축방안을 소개하였다.

본 논문에서 제안한 언어유형론에 기반한 다국어 공용번역지식 구축의 방법으로 얻을 수 있는 다국어 자동번역 시스템의 장점은 (1) 번역지식의 공유로 인한 번역지식크기의 축소 (2) 번역지식 공유로 인한 번역지식 작성 및 관리의 편리와 일관성 (3) 언어친족성을 가지는 새로운 언어에도 기존의 번역지식 리소스를 재활용할 수 있는 번역지식의 재활용으로 결론지을 수 있다.

참고문헌

[Choi 1997] Sung-Kwon Choi (1997). Unifikationsbasierte Maschinelle Uebersetzung mit Koreanisch als Quellsprache. IAI Working Papers N.34. Saarbruecken, Germany.

[Comrie 1981] Comrie, B (1981). Language Universals and Linguistic Typology. Basil Blackwell, Oxford.

[Dorr 1993] Dorr, B.J. (1993). Machine Translation: A View from the Lexicon. MIT Press, Cambridge, Massachusetts. London, England.

[Jackendoff 1977] Jackendoff, R.S. (1977). X-bar Syntax: A Study of Phrase Structure. Cambridge: MIT Press.

[Hawkins 1983] Hawkins, J.A.(1983). Word order universals. Academic Press.

[Lewis 1992] Lewis, D. (1992). Computers and Translation. In: Christopher Butler(ed.) Computers and written Texts. Blackwell, 75-114.

[Pollard 1994] Pollard, C. and I. Sag (1994). Head-Driven Phrase Structure Grammar. Studies in Contemporary Linguistics. The University of Chicago Press, Chicago & London.