영한자동번역에서의 프로택터와 문틀

최승권

한국전자통신연구원
지식처리연구팀

E-mail: choisk@etri.re.kr
URL: http://wwwkle.etri.re.kr/~choisk/index.html

목 차

❑ 개요
- ✒ 자동 번역이란 무엇인가?
- ✒ 영한자동번역 수준은 어느 정도인가?
- ✒ 지난 20년간 자동 번역 방법론은 어떻게 변화했는가?
- ✒ 자동 번역의 문제점은 무엇인가?
- ✒ 문제점을 어떻게 해결할 수 있을까?

❑ 프로택터
- ✒ 기존의 영어 자동구문분석의 문제점?
- ✒ 프로택터란 무엇인가?
- ✒ 영어의 프로택터 종류?
- ✒ 프로택터에 기반한 영어 구문분석(PBP)
- ✒ 프로택터에 기반한 영어구문분석 예

❑ 문틀
- ✒ 문틀이란 무엇인가?
- ✒ 영한의 문틀 종류?

❑ 문틀에 의한 번역
- ✒ 번역의 예

❑ 문틀의 구축 방법
- ✒ 문틀 구축 방법은 어떻게 할 것인가?

❑ 문틀의 번역률
- ✒ 예상되는 번역률은 어느 정도인가?
- ✒ 문틀의 양은 어느 정도여야 하는가?

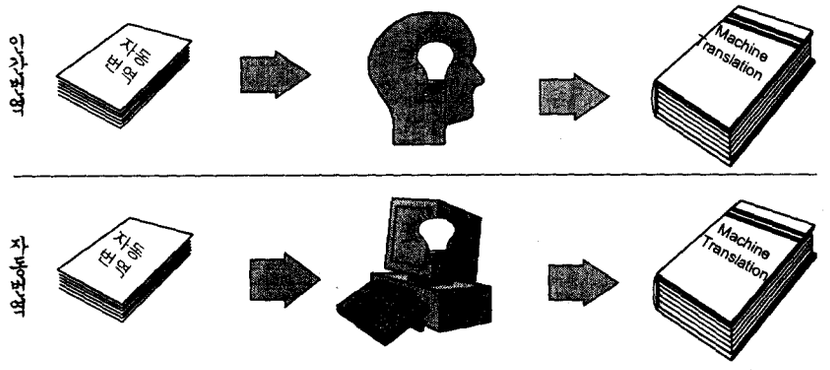
❑ 번역 실험
- ✒ 문틀에 의한 번역 실험 결과

❑ 참고 문헌

개요

- 자동 번역이란 무엇인가? -

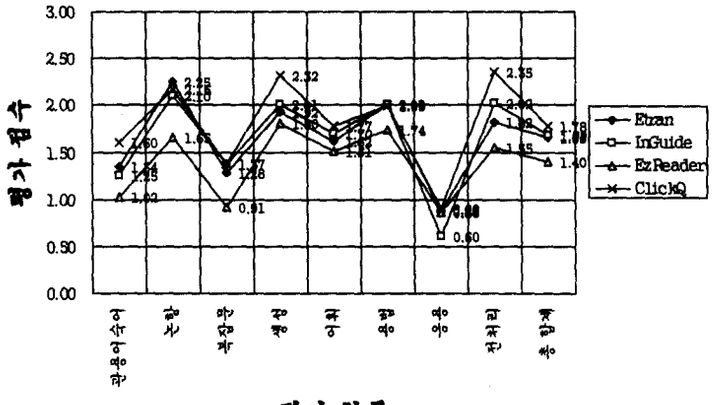
정의 : 컴퓨터가 한 언어의 텍스트를 다른 언어의 텍스트로 변환하는 것



개요

- 영한 자동 번역 수준은 어느 정도인가? -

❑ 국내 영한 자동번역기의 성능 평가 [???]



개요

- 지난 20년간 자동 번역 방법론은 어떻게 변화했는가? -

년도	1978 - 1988	1989 - 1998
규칙기반 MT	• 통사론 지향적 • 제한적 트리 변형 지향 • 분석/변환 지향 • 이해와 모호성 해결 지향 • 단방향 지향 • 언어학적 정보 • 사전 컴파일	• 어휘 지향적 • 단순적 제약/통합 지향 • 생성 지향 • 스타일/품질 지향 • 양방향 지향 • 어휘 개념적 데이타뱅크 • 사전 획득
코퍼스기반 MT		• 통계적 기반 (코퍼스의 직/간접 사용, Alignment 의 사용) • 예제 기반 (대역 코퍼스의 사용)
기타 MT		• 자동통역 (*British Telecom, ATR, C-STAR, Verbmobil*) • *Controlled input* MT system (Systran, Carnegie-Mellon for Caterpillar) • Domain-specific and sublanguage MT (Meteo, Pangloss, Carnegie-Mellon for Caterpillar, ATR speech translation, Verbmobil) • User-specific and custom-built MT systems (Winger, TRADEX)

개요

- 자동 번역의 문제점은 무엇인가? -

❑ 언어(Language)의 문제점들 [Hovy, 1999]

✒ Bar-Hillel's Argument (1960):

- The text must be understood, at some minimal level, for true translation to occur ("*the box is in the pen*")
- Text is too difficult for computers to understand
- Ergo, MT is impossible!

✒ Martin Kay's Observation (1993):

- Text contains just the skeleton; the reader provides the flesh him/herself
- => system needs tons of world knowledge

개요

- 자동 번역의 문제점은 무엇인가? -

❑ 언어(Language)의 문제점들

- ✒ 통사적 모호성
  I saw the man on the hill with the telescope.
  => a. 나는 망원경으로 언덕위에 있는 사람을 보았다.
  b. 나는 망원경을 들고 언덕에 서 있는 그 사람을 보았다.
  c. 나는 망원경이 서 있는 언덕에서 그 사람을 보았다.
- ✒ 의미적 모호성 [Homography, Polysemy]
  ball => a. 공 (pelota), b. 사교댄스 (baile)
- ✒ 문맥 모호성:
  The computer outputs the data; it is fast. / The computer outputs the data; it is stored in ascii.
- ✒ 어휘 선택:
  찾다 => find/seek, be => 있다 / 이다
- ✒ 시제/상의 차이:
  He has been in LA. => 그는 LA에 가 보았다. *He is loving the girl.=> 그는 그 소녀를 사랑하고 있다.

개요

- 자동 번역의 문제점은 무엇인가? -

❑ 기술(Technology)의 문제점들 [Choi et.al., 1994]

- ✒ 불연속 고정표현(Frozen expression)의 처리 미숙
- ✒ 형태소 태깅 모호성 처리 미숙
- ✒ 구조적 모호성의 처리 미숙
- ✒ 장문 처리 및 비문법적 문장 처리 미숙
- ✒ 대역어 선정 처리 미숙

❑ 그 밖의 기술(Technology)의 문제점들

- ✒ 번역 속도의 개선 문제
- ✒ 번역지식 리소스의 재활용성 문제
- ✒ 대용량 번역지식의 획득 문제
- ✒ 번역지식 구축의 충돌 방지 및 일관성 유지

개요

- 문제점을 어떻게 해결할 수 있을까? -

❑ 언어의 문제점들
- tons of world knowledge
- 통사/의미/문맥 모호성
- 어휘 선택
- 시제/상의 차이:

❑ 기술의 문제점들
- 불연속 고정표현의 처리 미숙
- 형태소태깅/구조적 모호성 처리 미숙
- 장문 처리 및 비문법적 문장 처리 미숙
- 대역어 선정 처리 미숙
- 번역 속도의 개선 문제
- 번역지식 리소스의 재활용성 문제
- 대용량 번역지식의 획득 문제
- 번역지식 구축의 충돌 방지 및 일관성 유지

❑ 언어의 문제점 해결책
- 현재로써는 불가능
- 구조적 모호성 해소 Formalism 개발
- Collocation 정보의 활용
- 언어간 차이 극복 표현 기술 개발

❑ 기술의 문제점 해결책
- 불연속 고정표현 장치 개발
- 태깅후처리기 및 모호성 해소장치 개발
- 장문분절기 개발
- Collocation 정보의 활용
- 복잡하지 않은 번역엔진 개발
- 재활용 가공도구의 개발
- 번역지식 획득도구의 개발
- 번역지식의 일관되는 점증적 구축 방법 개발

프로젝터

- 기존의 영어 자동구문분석(Parsing)의 문제점? -

❑ 기존의 영어 자동구문분석(Parsing)의 문제점
- 구조의 우측 경계의 인식 문제
  - 구단위의 좌.우측 경계선에 대한 컴퓨터언어학적 인식의 부정확성으로 인해 분석결과가 증폭하여 구조적 모호성을 대량 발생시켰음.
- 국부적 구조 기술의 문제
  - 하위 품사들이 상위 품사로 묶여가는(reduce) 것이 국부적으로(local) 기술되었기 때문에 구와 구간의 기능이 정확하게 인식되기가 어려웠다.
- 근본적인 문제
  - 때문에 대량의 구조적으로 모호한 분석결과에서 올바른 분석 결과(Best first parse tree)를 찾는 것은 매우 어려웠음.
  - 이러한 분석 결과의 모호성은 결국에는 번역결과에 영향을 미치어 좋은 품질의 번역을 내지 못하는 결과를 초래하였음.

프로텍터

- 프로텍터란 무엇인가? -

- 프로텍터의 정의
  - 프로텍터 (Protector)는 영어 문장에서 문장의 구조를 결정하는 핵심적인 역할을 하는 구성성분을 의미한다.
- 프로텍터의 기능
  - 구단위의 좌.우측 경계선을 인식함으로써 구조적 모호성을 해소하는 기능을 가진다.
  - 구와 구간의 기능을 인식하도록 하는 기능을 가진다.
  - 구단위의 Shallow parsing 유도

프로텍터

- 영어 프로텍터의 종류? -

- 동사(Verb)
  - 동사 단어들, 동사포함 고정표현들
  - 조동사를 포함할 때에는 조동사는 자질(Feature)로 표시
- 접속사(Conjunction)
  - Subordinate(although, because, except, if, lest, unless, that, how, when, whenever, where, whereas, whereat, wherever, whether, while, why)
  - Neutral (after, as, before, for, like ,since, than, though,till ,until ,without)
  - Coordinate (and, but, or, nor, plus)
  - Relative pronoun (that, what, whatever, which, whichever, who, whoever, whom, whomever, whomsoever, whose, whoseever)
- 기호(Punct)
  - ',', --, "", '' ,;,: (기타의 기호는 프로텍터가 아님)
- 조동사(Auxiliary verb, 본동사와 분리된 조동사)
  - be, become, can, could, dare, do, get, have, may, might, must, need, shall, should, will, would 등

프로텍터

- 프로텍터에 기반한 영어구문분석 (PBP)-

❑ 프로텍터에 의한 영어 구문분석(PBP: Parsing between Protectors)

- ✒ 프로텍터들 사이에 오는 모든 형태소 품사들을 문맥에 맞게 알맞은 구문품사로 reduce시키는 작업.
- ✒ 규칙은 자질정보를 포함하는 확장된 구구조 규칙으로 규칙기술부, 규칙조건부, 규칙실행부로 구성된다.
- ✒ 규칙기술의 원칙
  - 규칙기술부의 기술 원칙
    - 프로텍터 사이의 모든 형태소 품사들을 모두 나열하고 해당 구문 품사로 reduce 되도록 한다.
  - 규칙조건부의 기술 원칙
    - 형태소 노드들이 하나의 구문노드로 reduce될 때에는 제약조건을 기술하지 않는다. (문맥자유문법의 구축)
    - 형태소 노드들이 두개이상의 구문노드로 reduce될 때에는 좌우의 프로텍터를 기술하고 더 자세한 기술이 요구될 시에는 프로텍터의 유형을 기술한다. (문맥의존문법의 구축)
  - 규칙실행부의 기술 원칙
    - Head-driven Phrase Structure Grammar의 Head Feature Principle을 따른다.

13

프로텍터

- 프로텍터에 기반한 영어구문분석 예 -

❑ 원문

- ✒ the navy brought it to the surface this morning using the remotely operated vehicle "deep drone."

❑ PBP

- ✒ (the navy)NP1 (brought)VERB (it/PRON to/PREP the/DET surface/NOUN this_morning/ADV)NP2 PP1 (using)VERB2 (the remotely_operated_vehicle)NP3 (")LDQUOT (deep_drone)NP4 (.)PERIOD (")RDQUOT

❑ 전문번역가번역문

- ✒ 해군은 오늘 아침 원격으로 조정되는 잠수정 "deep drone"을 사용하여 그것을 해면으로 가져왔습니다.

14

문 틀

- 문틀이란 무엇인가? -

❑ 기존의 번역 패턴의 정의

✒ [이정민&배영남,1987]

- 여러가지 언어분석의 층위에서 각 언어요소 간에 일정하고 정연한 규칙성이나 상대적 관계가 있다는 것과 따라서 언어요소의 규칙적인 분류가 가능함을 나타내는 데 사용되는 술어.

✒ [서병락, 1996]

- 번역 패턴은 분석 단계에서 숙어나 패턴으로 인식된 원시 언어 표현을 번역하여 얻은 목표 언어 표현에 문장 생성에 도움을 주는 목표 언어 구조와 필수 정보를 부가하여 표현한 것으로 원시 언어에서의 숙어와 패턴을 생성의 관점에서 본 대응 지식 표현 구조이다.

✒ [Takeda, 1996]

- A pattern is a pair of CFG rules, and zero or more syntactic head and link constraints for non-terminal symbols.
- 예) NP:1 miss:V:2 NP:3 => S:2 S:2 <= NP3 manquer:V:2 à NP:1

문 틀

- 문틀이란 무엇인가? -

❑ 기존의 번역 패턴의 장단점

✒ 장점

- 목표언어의 생성을 고려하여 패턴이 기술되므로 출발언어분석 / 목표언어로의 변환 / 목표언어의 생성과 같이 분리되었을 때보다 목표언어의 표현이 더욱 자연스럽게 나온다.
- 구나 절 단위로 패턴을 기술하여 상위구나 절로 축약시킴으로 무한한 패턴의 량을 줄일 수 있다.

✒ 단점

- 구나 절 단위의 패턴구축으로 인해 통사적 모호성이 줄지 않는다.
- 구나 절 단위의 패턴 구축으로 패턴들 간의 구축 충돌이 발생할 수 있다.
- 패턴 구축에 충돌이 발생하므로 패턴에 일관성 유지가 어렵다.

문 틀
- 문틀이란 무엇인가? -

❑ 새로운 번역 패턴의 정의

✒ 문 틀(Sentence frame)
- 입력문장의 전체 또는 부분에 해당하는 완전한 문장을 포용하는 패턴을 문틀로 정의한다.
- 문틀 ⊇ 부분문틀 = {N|N은 완전한 문장이다}
- 문틀은 프로텍터를 반드시 포함한다

✒ 부분문틀(Partial Sentence frame)
- 문틀구축 시에 입력 문장의 전체에 해당하지는 않지만 완전한 문장을 포용하는 패턴을 부분문틀로 정의
- 기본적으로 부분문틀은 하나의 완전한 문틀이다
- 부분문틀은 프로텍터를 반드시 포함한다

(완전한 문장이란 접 레벨의 의미적 독립성을 가지는 문장)*

문 틀
- 문틀이란 무엇인가? -

❑ 새로운 번역 패턴의 장단점

✒ 장점
- 목표언어의 생성을 고려하여 패턴이 기술되므로 출발언어분석 / 목표언어로의 변환 / 목표언어의 생성과 같이 분리되었을 때보다 목표언어의 표현이 더욱 자연스럽게 나온다.
- 프로텍터의 설정으로 통사적 모호성이 줄어든다.
- 문장단위의 패턴구축으로 패턴들 간의 구축 충돌을 피할 수 있다.
- 패턴 구축에 충돌이 발생하지 않으므로 일관성 유지가 쉽다.

✒ 단점
- 문장 위주로 패턴을 기술하므로 대용량의 패턴을 구축하여야 한다.
- 원문 문장에 대한 대량의 문틀이 구축되었을 때 정확한(Korrekt) 문틀 선택의 기술 개발

문 틀

- 영한의 문틀 종류? -

- ❑ 원문틀
  - ✒ 프로텍터에 의해 구단위 분석된 영어 구문구조로써 제약조건자질이 명기된 (부분)문틀을 말한다.
- ❑ 대역문틀
  - ✒ (부분)문틀에서 실행자질이 명기된 한국어 출력구조를 말한다.
- ❑ 슬롯구조
  - ✒ (부분)문틀에서 프로텍터를 제외한 모든 구단위의 영한 대역구조

문틀에 의한 번역

- 번역의 예 -

입력문:
The man loves a woman

문틀구분	Key	구축할 내용
형태소분석		DET/the/DT NOUN/man/NN VERB/love/VBZ DET/a/DT NOUN/woman/NN
고정표현		
PBP	DET NOUN	det noun -> NP { } {/* SYNTACTIC SCOPE */ rhs[1]:start := lhs[1]: rhs[1]:end := lhs[2]: /* SYNTACTIC HEAD */ rhs[1]:eroot := lhs[2]:eroot: rhs[1]:etype := lhs[2]:etype: rhs[1]:detfex := lhs[1]:eroot: /* SEMANTIC HEAD */ rhs[1]:sem := lhs[2]:sem:} {}
원문틀	nVn	KKY-S-nVn@SP
제약조건	KKY-S-nVn@SP	{NP VERB [t1,vb] NP}
대역문틀	MJY-T-nVn@SP	{NP1:[etype ** pron demo cent comm] VERB1![{etype==t1) _AND (eform == vb)} NP2:[etype ** pron demo cent comm]} -> {NP1:[kcase:=subj] NP2:[kcase:=obj] VERB1[}
슬롯구조	DET NOUN	{ DET:[etype ** [a the]] NOUN:[etype == [comm]] } -> { NOUN[}
조동사생성태이블	punctual nul nul nul nul pres clex	(KROOT) (GCODE)

출력문:
남자는 여자를 사랑합니다.

문틀에 의한 번역

- 번역의 예 -

입력문:
NATO has pledged to the Serbs here that their rights will be respected.

분틀구분	Key	구축틀 내용
형태소분석		NOUN/nato/NNP AUX/have/MD VERB/pledge/VBN PREP/to/TO DET/the/DT NOUN/serb/NN ADV/here/RB CONJ/that/CONJ DET/their/PRP$ NOUN/right/NNS AUX/will/MD VERB/be/VB VERB/respect/VBN PUNCT/PERIOD/.
고정표현	pledge_to	pledge_to -> 약속하다

PBP	DET NOUN	det noun -> NP
		{ }
		{/* SYNTACTIC SCOPE */
		rhs[1].start := lhs[1];
		rhs[1].end := lhs[2];
		/* SYNTACTIC HEAD */
		rhs[1].eroot := lhs[2].eroot;
		rhs[1].etype := lhs[2].etype;
		rhs[1].detlex := lhs[1].eroot;
		/* SEMANTIC HEAD */
		rhs[1].sem := lhs[2].sem;}
		{}.

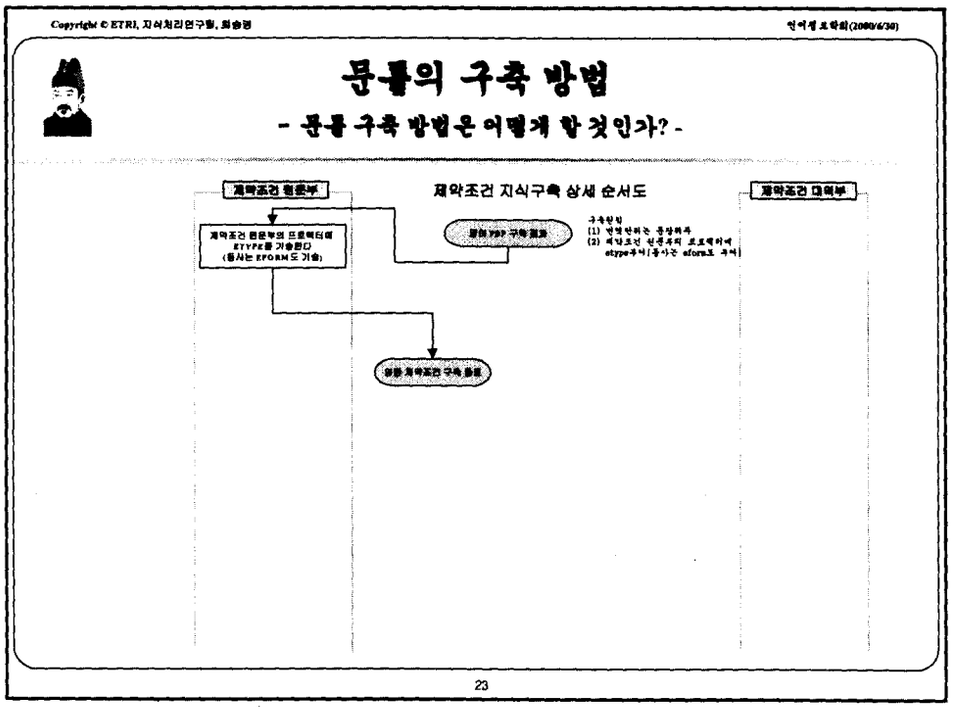
원문틀	nVnCnV	MJV-S-nVnCnV@SP
형학조건	MJV-S-nVnCnV @SP	(NP VERB:(vb,v5] NP CONJ:[subd] NP VERB:(vb,t1]) MJV-T-nnVCnV?@SP
대역문틀	MJV-T-nnVCnV? @SP	{NP1:[(etype ** [comm demo cent prop t1])] VERB1:[(etype *= [v5]) _AND (eform == [vb])] NP2:[(etype ** [comm demo cent prop t1])] CONJ1:[(etype *= [subd])] NP3:[(etype ** [comm demo cent prop t1])] VERB2 [(etype *= [t1]) _AND (eform == [vb])] } <- {NP1:[kcase := [topic]] NP3:[kcase := [sub]] VERB1 CONJ1:[kroot :: VERB1.kflex3, kcode :: VERB1.kfcode3, kcase := [obj]] NP2:[kcase :: VERB3.kflex2, kjcode :: VERB3.kfcode2, kcase :: [iobj]] VERB1 }
술분구조	DET NOUN	{ DET1:[etype ** [a the]] NOUN1:[etype *= [comm]] } -> { NOUN1 }
조동사생성태어틀	durative will null null passive pres et es	(KROOT = _것이) (GCODE El00323)

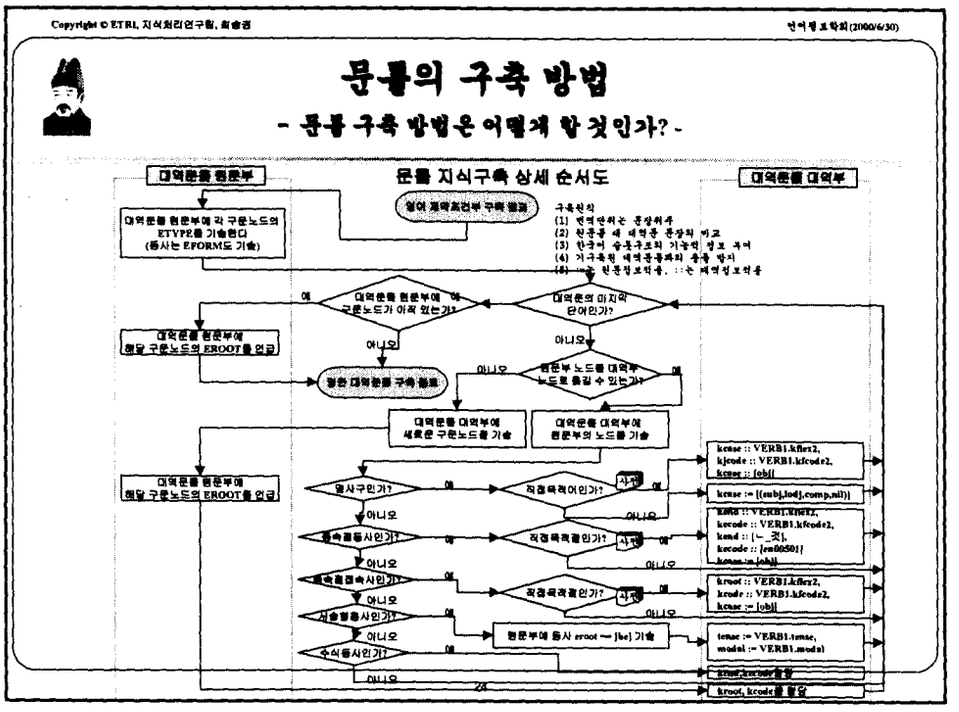
출력문 :
NATO는 자신들의 권리가 존중될 것이라는 것을 세르비아인들에게 약속하였습니다.

문틀의 구축 방법

- 문틀 구축 방법은 어떻게 할 것인가? -

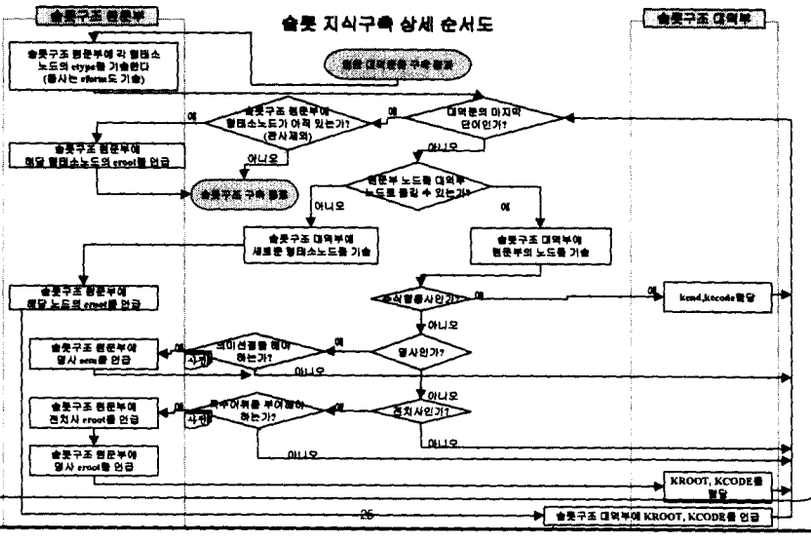
- ❑ 동일한 원문 문장에 대해 다수의 번역 패턴 구축자가 동일한 문틀을 구축할 수 있는 원칙 및 방법을 정한다.
- ❑ 이를 위해 다수 구축자가 대용량 번역 패턴을 구축할 수 있도록 다수 사용자용 번역 패턴 구축 도구를 사용한다.
- ❑ 또한 다수 구축자가 동일한 문틀을 구축할 수 있도록 번역 패턴에 기술되야 할 언어학적 지식을 (반)자동으로 부여할 수 있는 알고리즘을 사용한다.
- ❑ 대역어를 위한 어휘선택은 대용량 언어사전에 의해 해결한다.


Copyright © ETRI, 지식처리연구팀, 최승권
언어정보학회(2000/6/30)
문틀의 구축 방법
- 문틀 구축 방법은 어떻게 할 것인가? -
제약조건 원문부
제약조건 지식구축 상세 순서도
제약조건 대역부
23


Copyright © ETRI, 지식처리연구팀, 최승권
언어정보학회(2000/6/30)
문틀의 구축 방법
- 문틀 구축 방법은 어떻게 할 것인가? -
대역문틀 원문부
문틀 지식구축 상세 순서도
대역문틀 대역부
24

문틀의 구축 방법

- 문틀 구축 방법은 어떻게 할 것인가? -



문틀의 번역률

- 예상되는 번역률은 어느 정도인가? -

- 번역률 계산을 위한 모집합 코퍼스와 테스트 코퍼스의 양
  - 모집합 코퍼스: 1999.3.25-30일자 CNN뉴스로부터 임의 추출한 3230문장
  - 테스트 코퍼스: 1999.4.1일자 CNN뉴스로부터 임의 추출한 194문장
- 코퍼스로부터 번역패턴의 추출 양
  - 모집합 코퍼스: 원문틀과 부분원문틀의 합은 3718개
  - 테스트 코퍼스: 원문틀과 부분원문틀의 합은 275개
- 모집합 대 테스트 코퍼스의 매칭률
  - 모집합 원문틀에 매칭된 테스트 코퍼스의 패턴은 71개(25.82%)
  - 모집합 부분원문틀에 매칭된 테스트 코퍼스의 패턴은 2개(0.73%)
  - 모집합 패턴에 매칭된 테스트 코퍼스의 패턴은 73개(26.55%)
- 비록 같은 CNN뉴스의 방송자막문으로 실험한 결과지만 수천개의 원문틀만으로도 26.55%(73/275)의 전체 또는 부분적인 매칭 성공을 보여주어 수만 혹은 수십만의 문틀을 구축하면 성공적인 번역률을 예측할 수 있다는 긍정적인 예측을 하게 되었다.

문틀의 번역률

- 문틀의 양은 어느 정도여야 하는가? -

❑ 예상되는 문틀 수

✒ 1년 분량의 1개 방송사의 뉴스 자료를 모은 경우, 방송 당 약 40분, 508문장, 모든 문장이 서로 다른 문틀이라고 가정할 경우, 508문장 * 365일 = 185,420 문틀 (전체)

✒ 각 문틀이 평균 2개의 부분 문틀을 포함한다고 가정할 경우, 58,400 문형 * 2 문틀 = 370,840 문형 패턴 (전체 또는 부분)

♥ 부분문틀은 문틀 구축 도구에 의해 자동으로 구축됨

♥ 위의 가정은 겹치는 문장이 하나도 존재하지 않는다고 가정하는 최악의 상황

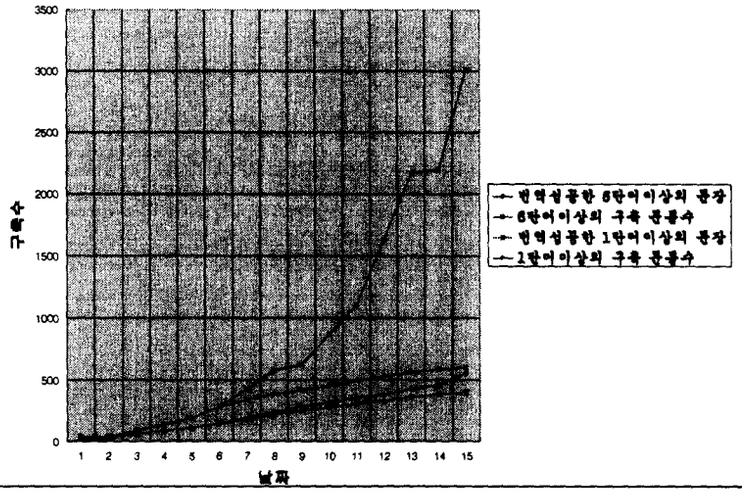
❑ 예상되는 문틀 구축 시간

✒ 자동화된 문틀 패턴 구축 도구에 의하면, 하나의 입력 문장에 대해 사람의 노력이 들어가는 부분은, 태깅 결과 수정, 파싱 결과 검증, 부분적 자질 수정, 부분 문틀 체크, 대역어 생성으로, 최대 10분, 하루에 8시간 작업하는 것으로 가정하는 경우, 1명이 6개월을 작업하는 경우, 6문형 * 8시간 * 365일 = 17,520 문형 (전체)

♥ 최악의 경우에도 10.5명을 투입하면 1년 내에 1년 분량의 방송 뉴스의 문형 패턴 구축 가능

번역 실험

- 문틀에 의한 번역 실험 결과? -



참고 문헌

- [Choi et.al., 1994] Choi K.S., Lee S.M., Kim H.G., and Kim D.B. (1994) *An English-to-Korean Machine Translator: MATES/EK.* In: COLING94, pp. 129-133.
- [Hovy, 1994] Eduard Hovy (1994) Machine Translation. In: MT Summit99.
- [Hutchins, 1994] J. Hutchins (1994) Research methods and system designs in machine translation – a ten-year review, 1984-1994. In *Machine Translation Ten Years On*. Cranfield University.
- [Mason & Rinsche, 1995] J. Mason and A.Rinsche (1995) Ovum Evaluates – Translation Technology Products. Ovum Ltd.
- [Takeda, 1996] Koichi Takeda (1996) Pattern-Based Context-Free Grammars for Machine Translation. In: ACL96.
- A Japanese View of Machine Translation in light of the Considerations and Recommendations Reported by ALPAC, U.S.A.
- [서병락, 1996] 한영 기계 번역을 위한 번역 패턴에 기반한 영어 문장 생성기. In: 정보과학회논문지, 제 23권 제5호, 520-529쪽.
- [시스템공학연구소, 1998] 시스템공학연구소 (1998) 1998년도 기계번역 연구실 연구 백서.
- [이정민&배영남, 1987] 이정민, 배영남 공저(1987) 언어학 사전, 박영사.