

한영 기계 번역을 위한 조사 사전 구성에 관한 연구

최 재혁\*, 김 권양\*, 박 상규\*, 이 상조\*

부산여자대학 전자계산학과\*, 경북대학교 전자계산기공학과\*

The Study of Josa Dictionary Construction for Korean-English Machine Translation

Choi Jae-hyuk\*, Kim Kweon-yang\*, Park Sang-gyu\*, Lee Sang-jo\*

Pusan Women's University\*, Kyungpook National University\*

요 약

본 연구는 한영 기계 번역을 위한 사전 중에서 한국어 조사 사전에 대한 모델을 제시하였다. 특히 정확한 억어 선택을 위한 제약 정보를 수집하는데 중점을 두었다. 지금까지의 한국어 조사에 대한 억어 선택 방법은 제언의 속성 정보에 의한 억어 선택과 default 억어의 선택이었다. 그러나 한국어의 한 조사에 대응하는 영어의 전치사의 수가 너무 많음으로 인하여 이러한 기존의 방법을 사용할 경우 고질의 번역은 기대할 수 없다. 따라서 본 논문에서는 정확한 억어의 선택을 위하여 조사의 격 분류와 제언의 속성 정보를 더욱 세분화 시키고, 이를 이용한 용언의 기본 구문 패턴을 재정립하였다. 또한 한 두개의 default 억어로 인한 번역의 부정확성을 탈피하기 위하여 default 억어 및 제언의 속성 정보에 의한 억어를 용언의 의미적 분류에 의해 더욱 세분화 시킴으로써 정확한 억어를 선택하는 방법을 제시하였다.

I. 서론

한국의 국제적 지위의 향상과 국제간의 정보 교류의 확대에 따라 번역에 대한 수요가 증대되고 있으며, 각종 정보를 자국어 혹은 타국어로 신속하게 번역하여야 하는 요구가 증대되고 있다. 이러한 의미에서 번역 과정을 기계화하려는 시도가 외국에서는 이미 어느정도 정착되어가고 있으며, 특수 분야에서 다수의 자동 번역 시스템이 부분적으로 상용화되어 사용되고 있다. 국내의 경우, 기계 번역에 대한 연구가 최근 활발히 진행되고 있으나 기계 번역시 가장 중요한 사전에 대한 연구는 상당히 미비한 실정이다. 특히 국어의 특징 가운데 큰 비중을 차지하고 있는 조사에 대한 사전 연구는 거의 전무한 실정이다. 한영 기계 번역시 조사를 영어로 번역시키기 위해서는 한국어 조사의 문법적인 기능과 의미를 분석하고 난 후 그에 대응하는 억어를 선택해야 한다. 한영 기계 번역시 지금까지의 한국어 조사에 대한 억어 선택 방법은 거의가 default로서, 한 조사에 대해 한 두개의 억어를 대응시켰다. 그러나 실질 번역에 있어서는 한개의 조사에 대해 수십개의 억어인 영어의

전치사가 대응되기 때문에 이러한 억어 중 가장 최적한 억어의 선택이 상당한 난관에 부딪혀 있었다.

이의 해결을 위하여 본 연구에서는 조사에 대한 격 표시 정보를 기존의 것보다 더욱 세분화시키고, 이를 이용한 용언의 구문 정보와 속성 정보, 그리고 용언의 의미적 분류에 의하여 최적한 억어를 선택하는 메커니즘을 제시하고 이를 사전 정보에 기술하였다. 그러나 한국어 조사의 수가 너무 많음으로 인해 가장 널리 사용되는 기본 조사 15개를 선정하고, 이 15개의 조사에 대해서만 사전을 구성하였으나, 지면 관계상 이들 중 4개만을 부록에 수록하고, 그 이외의 특수 조사(도, 간, 뿐...)를 비롯한 나머지 조사에 대해서는 추후 연구하기로 하였다.[12]

II 본론

1. 용언의 구문 정보

본 논문에서 사용하는 용언의 구문 정보는 천기석이

제시한 의미적 사용에 따른 용언 패턴 분류를 기반으로 하여 하위 범주화 정보를 설정하되 한글, 한영 사진의 예문을 통해 구체적인 검증 과정을 행하였으며, 본류된 패턴을 결합과 유형에 따라 19개로 설정하였다.[9,11]

이와같이 설정된 구문 정보를 사용함으로써 파본 역어 선택을 가능하게 하는데 이의 장점은 다음과 같다.

첫째, 동사와 형용사의 품사 구분이 가능하다.

(예; 맛이 쓰다, 모자를 쓰다)

둘째, 자동사와 타동사를 구분함으로써 역어 선택시 고려해야 할 역어의 수를 어느 정도 줄여준다. (예; 답이 맞다, 때를 맞다)

셋째, 용언의 모호성을 해결하기 위하여 제안의 속성 정보를 이해해야 할 경우, 비교해야 하는 제언을 미리 구문 정보에서 지정함으로써 비교해야 하는 제안의 수를 줄여 준다.

넷째, 용언의 구문 패턴에 따라 선택된 영어의 동사에 속어적으로 사용되는 조사의 역어인 전치사를 미리 제공할 수도 있으며, 또한 제공할 수 없는 경우도 조사의 격을 미리 결정하여 조사에 대응되는 전치사의 수를 줄여 줄 수가 있어 조사의 역어 선택을 아주 용이하게 한다.

본 논문에서 사용하고 있는 용언의 기본 구문 패턴은 그림 1과 같다.

vpk1) SBJ ;상승하다, 녹다, 어다, ...

vpk2) SBJ, OBJ ; 먹다, 부정하다, 파괴하다, 입다, 쓰다

vpk3) SBJ, OBJ, DAT ;

- 팔다, 매술하다, 임대하다, 주다, 나누다,
- 할당하다, 본배하다, 신청하다, 가르치다, 구걸하다, 축하다.
- 첨가하다, 더하다, ...

vpk4) SBJ, OBJ, ORI ;

- 사다, 얻다, 수입하다, 차용하다, 배우다,
- 감하다, 추출하다, ...

vpk5) SBJ, SFR, STO ; 가다, 오다, 움직이다, 돌다, 이동하다, ...

vpk6) SBJ, OBJ, SFR, STO ; 이동하다, 옮기다, ...

vpk7) SBJ, OBJ, OPP ;

- 나누다 (교환), 바꾸다, 뺏다 (상점)
- 합성하다, 찍다 (형성)

vpk8) SBJ, COM ; 낫다, 뛰어난다, 못하다, ...

vpk9) SBJ, CPL ; 되다, ...

vpk10) SBJ, OPP ; 약수하다, 헤어지다, 싸우다,

연애하다, 언쟁하다, 동맹하다 (상점)

vpk11) SBJ, OBJ, CPN ; 나누다 (분할), 분해하다, 분산하다, ...

vpk12) SBJ, OBJ, MAT ; 만듦다, 종합하다 (형성)

vpk13) SBJ, SPA ; 사망하다, 살다, 놀다, 비치다 (가변), 있다, ...

vpk14) SBJ, OBJ, CAU ; 중지하다, 휴고하다, 휴식하다, 걸식하다, 쉬다 (중단)

vpk15) SBJ, OBJ, ROL ; 고송하다, 임명하다, ...

vpk16) SBJ, OBJ, PUR ; 소비하다, 낭비하다, 쓰다, ...

vpk17) SBJ, OBJ, SPA ; 싣다, 적재하다, ...

vpk18) SBJ, OBJ, MAN ; 구분하다, 나누다, ...

vpk19) SBJ, CAU ; 놀라다, ...

그림 1. 한국어 용언의 기본 구문 패턴

## 2. 제언의 속성 정보

제언은 실질 형태소인 품사 정보로서 명사, 대명사, 수사로 나누어지며, 제언 자체의 역어 선정시의 모호성 뿐만 아니라 조사와 용언의 다의성 해결에 필요한 정보를 제공한다. 이러한 제언의 속성 정보의 양과 질에 따라 번역 시스템의 질이 좌우된다해도 과언이 아니다.

본 논문에서 분류한 제언의 속성 정보는 크게 두 가지로 나눌 수 있는데 하나는 제언 자체의 의미 분류이고, 다른 하나는 용언과 조사에 필요한 실행용 위주의 분류이다. 따라서 조사의 필요성에 의하여 분류한 제언 속성 정보의 대표적인 예로 지명과 시간을 들수 있는데 영어로의 번역시 지명일 경우는 일반적으로 대지명이면 in, 소지명이면 at으로 번역하며, 시간일 경우 현재의 시각을 나타낼 경우는 at, 요일을 나타낼 경우는 on, 아침과 같은 시점을 나타낼 경우는 in 등으로 번역을 해야 한다. 이와같이 조사에서의 정확한 역어 선택을 위해서는 제언의 속성 정보를 보다 세분화해야 하는 작업이 필요하다.

따라서 본 논문에서 사용하는 제언의 속성 정보는 제언, 용언, 그리고 조사의 역어 선정시의 모호성 해결에 중점을 두고 약 200여개로 분류하였으며 이의 특성은 다음과 같다.[12]

o 계층적인 나무 구조 : is - a 관계표시

o 용언, 조사 위주 분류와 제언 위주 분류의 겹목

o 하위 부분 분류의 세분화

o 의미 속성에 대한 의미 네트워크 구성

: part - of 관계, 집합관계, 연칭, 성, 학문분야, ...

그림 2는 본 논문에서 분류한 제언 속성 정보를 지면 관계상 그 일부만을 나타낸 것이다.

## 3. 한국어 조사의 격 표지 정보

본 연구에서 설정한 격 표지 정보는 격 본분 이론을 바탕으로 이미 국내 학자들이 제시한 각 조사의 분류와 일명 기계 번역 시스템인 Mu-system에서 설정한 일본어의 격 표지 정보 및 기계 번역시 필요로 하는 사항



- 33) 화제격(TOPic) : 은/는
- 34) 관점격(VIEWpoint) : 에서, 로부터, 인 점에서
- \* 35) 비교격(ComPaRison) : 처럼, 같이, 보다, 만큼, 와/과, 에, 에서, 하고, 만, 읊/촬
- 36) 수반격(ACComPany) : 와 함께, 와 동반해서
- 37) 정도격(DEGree) : 씩, 만큼, 가
- 38) 서술격(PREdicative) : 이다

그림 3 한국어의 각 표지 정보 분류

그림 4는 각 표지 정보를 15 개의 조사에 대해서 제 본배한 것이다.

- o 이/가 : SUB, CPL, ADJ
- o 은/는 : SUB, TOP
- o 을/를 : OBJ, SFR, STO, STH, TOO, CPR, DUR
- o 로/으로 : OBJ, GOA, CAU, TOO, SPA, STO, CAU, PUR, CPR, ENU, GOA
- o 와/과 : PAR, OPP, CPR, ACC, ENU
- o 고 : QUA, ENU
- o 에 : DAT, ORI, PAR, TIM, TFR, TTO, SPA, STO, CAU, PUR, CPR, ENU, GOA
- o 에서 : SUB, ORI, TFR, SPA, SOR, RAN, VIE, CPR
- o 까지 : STO, DEG, TTO
- o 에서는 : SPA, RAN
- o 예게 : DAT, ORI
- o 예게서 : DAT, CPN
- o 로서 : BOL, MAN
- o 로서 : TOO, CDT
- o 의 : ADJ

그림 4 한국어의 각 표지 정보 분류표

4. 조사 사전에 필요한 정보 및 조법

지금까지의 조사 사전은 제언의 속성 정보 및 용언의 구문 정보에만 의존하여 억어인 전치사를 선택하는 방법을 채택하고 있다. 이러한 기존 방법의 문제점으로 다음과 같은 예를 들 수 있다.

- (1) 비스토 서울을 가다.
- (2) 비스토 길을 막다.

위 예제에서 조사 '로'의 번역은 제언의 속성 정보가 vehicle 일 경우 수단을 나타내는 방식적으로 'by'로 번역되는 것이 일반적이거나, (2)의 예제에서 보면 버스가 라는 목적이 아니라 하나의 도구로서 사용되었기 때문에 'with'로 번역하는 것이 옳을 것이다.

위의 예제에서 처럼 제언의 속성 정보만으로는 정확한 억어 선택을 보장 받을 수 없기 때문에 제언의 속성 정보와 더불어 이의 구분을 위하여 제언이 가지는 용언의 의미적 분류 혹은 용언의 구문 정보를 동시에 고려하여야만 한다는 것을 알 수 있다. 이를 위하여는 용언

의 의미적 분류를 행하여야만 하는데 본 연구에서는 전 기석이 제시한 한국어 용언의 의미적 분류에 따르도록 하였다.[9,12]

또한 추가적으로 억어 선택시 고려해야할 정보로는 제언과 제언 간의 관계(relationship)이다. 일반적으로 제언들 간의 속성정보는 IS-A 계층 관계를 나타내는 데 이 외에도 제언들 간의 'Part-of' 관계, 'Membership' 관계, 성분 관계, 재료 관계, 속성 관계 등의 관계를 고려하여야만 보다 정확한 억어를 선택할 수 있다. 특히 조사 '의' 사전에서 이러한 제언들 간의 관계가 억어 선택을 위해 사용되어진다.[12]

다음은 조사 사전의 구성 요소 및 데이터 항목에 대한 특성과 의미를 설명한 것이다.

(1) 제언의 속성 정보

제언 사전에서 미리 제시된 제언의 속성 정보를 이용한다.

(2) 각표지 정보

그림 3에 나타난 각 표지 분류에 의해 사용된 조사의 각 표지 정보를 나타낸다.

(3) 부가 구문 정보

부가 구문 정보는 단일화를 취할때 필요한 정보를 나타낸다.

o null : 용언 사전에 있는 구문 부가 정보와 단일화

o v : 동사의 부가어

o n : 명사와 단일화

(4) 억어 변환 의미 정보

의미 정보는 용언의 의미적 분류 정보와 제언 간의 관계 정보로 크게 나눌 수 있다. 용언의 기본 구문 패턴에서 제시하지 못하는 것에 한하여 용언의 의미적 분류에 의하여 default 억어를 취함으로써 좀더 정확한 억어를 선택할 수가 있다.

o null : v(\*) 에서 \*를 제외한 나머지 용언에 대한 default.

o v(\*) : 용언의 의미적 분류

o R(\*) : 제언간의 관계 정보

(5) 억어 및 억어의 구문정보

o null : 억어가 용언의 기본 구문 패턴에 따라 용언 사전에서 제시

o 각 표지(\*) : 억어 및 억어의 부분적인 어순을 나타냄

5. 억어 선택 메카니즘

구성된 조사 사전에서 억어를 선택하는 알고리즘은 다음과 같다.

[알고리즘] 억어 선택 수행 순서

1. 제언의 속성 정보와 일치하는 사전 속성 정보가

있는지 조사

2. 만약 있다면, 일치된 tuple 발췌
3. 만약 없다면, 사전의 속성 정보가 N1 인 모든 tuple 발췌
4. 발췌된 tuple 의 부가 구문 정보가  
Null 이면 용언과 단일화  
N 이면 명사와 단일화  
V 이면 용언의 부가어
5. 만약 단일화를 취할때 tuple 내에 용언 의미 정보가 있을 경우, 용언 사전에서 의미 정보와 일치하는 tuple 과 단일화.

위의 알고리즘에 대한 억어 선택 수행 순서의 예로 '나는 버스로 학교에 간다' 라는 문장에서 버스의 제언 속성 정보가 vehicle 이므로 '로' 사전에서 다음과 같은 tuple 을 발췌한다.

N1(vehicle)	Manner	V	V(유생개체, 전제소유 대상물 제공)	MAN(by+N1)
	TOOL	V		TOO(with+N1)

다음은 '가다' 라는 동사에 대한 용언의 의미적 분류가 유생 개체에 속하므로

N1(vehicle)	Manner	V	V(유생 개체)	MAN(by+N1)
-------------	--------	---	----------	------------

위의 tuple 이 최종적으로 선택되고 억어는 by bus로 번역된다.

### III. 결론

본 연구는 한영 기계 번역 시스템에서 주변 역할을 하는 사전중에서 조사 사전에 대한 모델을 제시하였다. 특히 정확한 억어 변환을 위한 제약 정보를 수집하는데 중점을 두었다.

그리하여 제언의 속성 정보와 용언의 기본 구분 패턴만으로 조사의 억어를 선택하는 기존 방법보다 좀더 정확한 억어를 선택하기 위하여 조사의 각 분류와 제언의 속성 정보를 더욱 세분화시키고 이를 이용한 용언의 기본 구분 패턴을 제정립하였다.

추가적으로 default 억어의 부정확성을 탈피하기 위하여 default 억어를 용언의 의미적 분류에 의하여 더욱 세분화시켜 정확한 억어를 선택하는 방법을 제시하였다.

이렇게 구성된 사전 시스템의 타당성 및 질적 평가를 위해 국정 교과서의 예문을 통해 영역한 결과 대부분의 경우에 적절한 억어로 변환할 수 있었다.

그러나 본 조사 사전에서 채택한 용언의 의미적 분류가 단순한 한국어 용언의 의미적 분류임으로 인하여 좀더 번역의 질을 높이기 위해서는 한영 번역을 위한

용언의 의미적 분류 작업이 행하여져야만 할 것이다.

### 참고 문헌

- [1] 김 민수, 국어 의미론, 일조각, 1980
- [2] 김 제훈외 2인, "한영 양국어 간의 위상 구조 변환을 위한 Transfer 모델", 한국정보과학회 학술 발표 논문집, Vol.15, No.2, 1988
- [3] 김 종현, "국어의 내포론 통제와 무제한 의존 관계", 서울대 언어학과 석사 학위 논문, 1988
- [4] 성 광수, 국어 조사에 관한 연구, 형설 출판사, 1979
- [5] 이 상조 외 2인, "한영 기계 번역을 위한 사전 구성에 관한 연구", 제1회 기계 번역 workshop 발표 논문집, 1989.
- [6] 이 상조 외 3인, "한영 기계 번역을 위한 사전 구성에 관한 연구", 제2회 기계 번역 workshop 발표 논문집, 1989.
- [7] 이 희승, 국어 대사전, 민중서림, 1961
- [8] 정 의성, "한글 구조조 분석", 제1회 자연 언어 처리 워크숍 발표 논문집, 1989
- [9] 천 기석, "국어의 동작 동사와 상태 동사의 체계연구", 경북대 박사 학위 논문, 1983
- [10] 최 현배, 우리 말본 제판, 정음사, 1946
- [11] 뉴월드 한영 대사전, 1979
- [12] 한영 번역을 위한 전자 사전 구성에 관한 연구, 경북대학교, 한국전자통신연구소, 1989.
- [13] H.Tsurumaru, T.Hitaka, S.Yoshida, "An attempt to automatic thesaurus construction from an ordinary japanese dictionary", COLING 86, 1986
- [14] J.Tsujii, J.Nakamura and M.Nagao, "Analysis grammar of japanese in the Mu Project- A procedural approach to analysis grammar", COLING 84, 1984
- [15] T.Gunji, "JPSG : Japanese Phrase Structure Grammar", ICOT-TR-199,1986
- [16] Y.Sakamoto, T.Ishikawa and M.Satoh, "Concept and Structure of Semantic Markers for Machine Translation in Mu-Project", COLING 86, 1986

부류 : 조사 사전의 예

1. '이/가' 조사

채언 속성 정보	격표지정보	부가 구문 정보	역어 변환 의미 정보	역어 및 역어의 구문 정보
N1 (동물, 식물)	Adject	N+주격조사	R(part-of)	ADJ (N*of*N1)
	Complement	N+주격조사	not R(part-of) and V (상대동사)	CPL(N1)
	Subject			SUB(N1)
N1 (인명, 호칭, 인간지칭, 직위직업)	Adject	N+주격조사	R(part-of)	ADJ(N1*s*N)
	Complement	N+주격조사	not R(part-of) and V (상대동사)	
	Subject			SUB(N1)
N1	Complement	N+주격조사	V(상대동사)	CPL(N1)
	Complement		V('되다', '아니다')	
	Subject		V(동작동사)	SBJ(N1)
PN1	Adject	N+주격조사	R(part-of)	ADJ(PN1:poss*N1)
	Subject			SUB(PN1:subj)

2. '을/를' 사전

채언 속성 정보	격표지정보	부가 구문 정보	역어 변환 의미 정보	역어 및 역어의 구문 정보
N1(장소)	Space-from		V(유생개체 out)	
	Space-to		V(유생개체 in)	
N2(기간)	Duration	N1(기수)		DUR(for*N1)
N1	Object		V(대상분할)	
			V(타동사)	OBJ(N1)
	Verb-object	V(하다, 지다, 자다.....)		VOB(N)
PN1	Object		V(타동사)	OBJ(PN1:obj)

3. '의' 사전

채언 속성 정보	격표지정보	부가구문정보	역어 변환 의미 정보	역어 및 역어의 부문 구조
N1(동물, 식물)	Adject	N2		ADJ(N2*of*N1)
N1(인명, 호칭, 인간지칭, 직위직업)	Adject	N2	R(자차)	ADJ(N2*by*N1)
				ADJ(N1*s*N2)
				ADJ(N2*by*N1)
N1(단위)	Adject	N2		ADJ(N1*of*N2)
N1(시점)	Adject	N2		ADJ(N1*s*N2)
N1(학문)	Adject	N2(속성)		ADJ(N2*on*N1)
		N2		ADJ(N2*of*N1)
N1(병리현상)	Adject	N2(의약품)		ADJ(N2*for*N1)
N1(직위직업)	Adject	N2(조직)		ADJ(N2*with*N1)
N1(소재명)	Adject	N2(동물, 식물, 인조물)	R(특산물)	ADJ(N2*from*N1)
				ADJ(N2*at*N1)
		N2(장소개념)	R(위치)	ADJ(N2*of*N1)

N1(소재명)	Adject	N2(동물, 식물, 인조물)	R(특산물)	ADJ(N2*from*N1)
		N2(장소개념)	R(위치)	ADJ(N2*in*N1)
				ADJ(N2*of*N1)
N1	Adject	N2		ADJ(N2*of*N1)
PN1	Adject	N2		ADJ(PN1:poss*N2)

4. '로/으로' 사전

채언 속성 정보	격표지정보	부가 구문 정보	역어 변환 의미 정보	역어 및 역어의 구문 정보
N1(장소)	Space-to		V(유생개체, 무생개체)	
		V		STO(to*N1)
N1(방향)	Space-to		V(유생개체, 무생개체)	
		V		SPA(to*N1)
N1(vehicle)	Manner	V	V(유생개체, 전계소유, 대상물제공)	MAN(by*N1)
	Tool	V		TOO(with*N1)
N1(방식)	Manner		V(대상분할, 형성, 중합)	
		V		MAN(in*N1)
N1(병리현상)	Cause	V	V(대상감소, 대상증대, 행위자시연)	AU(on account of*N1)
N1(인공물)	Material	V	V(중합)	MAT(from*N1)
	Tool	V		TOO(with*N1)
N1(직위직업)	Role	V		ROL(as*N1)
N1(언어)	Goal	V	(도구조건, 발화전달)	GOA(in*N1)
				GOA(by*N1)
N1(기간)	Duration	V		DUR(for*N1)
N1(시점)	Time	V		TIM(at*N1)
N1(나이)	Goal			GOA(for*N1)
N1	Manner		V(발화전달, 요구질정)	
		V		MAN(by*N1)
	Condition	V	V(수혜자 격하)	MAN(in*N1)
		V	V(도구조건)	
	Goal	V		CON(by*N1)
		V	V(대상변경)	
	Component		V(형성)	
		V		CPR(of*N1)
	Type		V(중합)	
		V		TYP(in*N1)
Default				
	V		DEF(by*N1)	