중간언어 기계번역방식을 위한 어휘지식 표현체계에 관한 연구

이휘봉, 송성대, 이종혁, 이근배

포항공과대학교 전자계산학과

A Study on Lexical Knowledge Representation for Interlingua Machine Translation

Hui-Feng Li, Seong-Dae Song, Jong-Hyeok Lee, Geunbae Lee

Dept. of Computer Science & Engineering, POSTECH

요 약

본 논문에서는 중간언어 설계의 일부분으로, 중간의미 표현을 위한 어휘지식 표현 방안에 관하여 논한다. 기존 중간언어들은 단어의 의미 구별법이 단순한 선택적 제한을 기반으로 하고 있으며, 시소러스 체계도 단일하게 유지하고 있다. 따라서, 단어의 의미간 중첩성이 반영되지 못하고 단어의 창조적 사용(creative use)에 대한 대처능력도 떨어진다. 또한 단일 시소러스체계를 통해서는 단어들의 명확한 분류기준을 파악할 수가 없다. 이러한 어휘지식 표현체계의 문제점들을 극복하기 위한 해결책으로서 생성사전(Generative Lexicon)을 도입하고, 중간표현의 관계기호를 효과적으로 파악하기 위한 관점에서의 시소러스 분류체계를 제안한다. 또한 이 같은 어휘지식 표현체계를 이용하여 문장의 구문구조로부터 중간표현을 나타내는 과정을 제시한다.

1. 서론

일반적으로 기계번역은 원시언어의 분석, 목표언어 생성을 위한 구조로의 변환, 그리고 목표언어의 생성이라는 세 단계를 거쳐서 이루어 진다.

중간언어방식은 원시언어와 목표언어의 언어구조는 독립적인 의미구조를 매개로 번역이 이루어지며 다음과 같은 장점으로 인해 꾸준히 연구되어 오고 있다[CICC93].

- 기계번역시스템의 개발을 대상 언어에 지역화(localize) 시킬 수 있다.
- 개발과정에서 얻어지는 지식을 중간언어를 통하여 공유할 수 있다.

본 논문은 기존 중간언어들의 어휘지식표현의 문제점들을 파악하고 그러한 문제점을 극복한 멀티 시소러스에 기반한 어휘지식 표현체계를 제안한다. 특히 중간표현의 관계기호를 파악하는데 유용한 관점에서의 어휘지식 표현체계를 제안한다.

* 본 연구는 한국과학재단 국제공동연구과제 "한-중 기계번역 시스템" 연구('93.5 ~ '95.5)의 일부분임.

2. 기존의 연구

기존의 기계번역 시스템 개발 현황에서, 중간언어 방식의 연구가 많이 이루어져 왔다.

예를 들면, 영국 캠브리지 대학의 TRANSLATOR, 프랑스 그레노블대학의 CETA, 일본 NEC 의 PIVOT, CICC 의 Interlingua, 그리고 미국 CMU 의 KBMT, 메릴랜드 대학의 UNITRAN, 국내의 시스템공학연구소의 중간언어에 관한 연구 등 연구들이 있었다[CICC93, Dorr93, Hut92, Nir87, Nir89, 龜井 88. 김상 93].

일반적으로 중간언어의 어휘 부분은 가장 기본적인 의미단위로서의 개념기호(concept symbol) 부분과 개념기호간의 관계를 명시하는 관계기호(relation symbol) 부분, 그리고 개념기호의 의미에 제한을 가하는 속성 (attribute) 부분으로 구성된다[Nir87, 龜井 88, CICC93, Yao90, Dorr93, Dorr94]. 그리고 표현체계에 있어서는 대체적으로 서술어 노드가 중심역할을 한다.

3. 기존 어휘지식 표현체계의 문제점 및 그 해결방안으로서의 Generative Lexicon

본 논문이 다루는 초점은 중간언어 어휘 구성 요소들 중에서 개념기호(concept symbol)의 어휘지식 표현체계이다.

3.1. 기존 어휘지식 표현체계의 문제점

기존의 어휘지식 표현체계에서 여러가지 문제점들을 야기시키는 두 가지 주된 원인은 다음과 같다.

- 단어의 의미 구별법(word sense enumeration)이 단순한 선택적 제한(selectional restriction)을 기반으로 이루어진다 [Puts93].
- 시소러스 체계가 단일하다[Puts93].

이와 같은 두 가지 원인은, 단어의 의미를 추상화 (abstraction) 하는 관점이 자연스럽지 못하고 비효율적이다 [Puts91, Puts93].

(1) 선택적 제한에 기반한 단어 의미 구별법

예) John baked potatoes. (bake1)

Mary baked a cake. (bake2)

위에 언급된 예문에서 "bake"라는 단어의 "상태의 변화"(bake1)와 "제조"(bake2)라는 제한을 기반으로 의미가 구별되고 있다. 이러한 구별이 가지는 문제는, 단어 의미간의 중첩성(overlap)이 반영되지 못하다는 것이다[Puts93]. 위에 제시된 예문의 경우에도 "열을 가하여 물질의 상태를 변화시킴"이라는 의미적 중첩성은 반영되지 못하고 있다. 이러한 중첩성 무시는 단어 선택(word selection)에 있어서, 기존의 어휘지식 표현체계가 제공하는 정보가 매우 미약함을 의미한다. 정확한 번역을 위해 올바른 단어의 선택이 이루어져야 한다는 점에서, 중첩성의 무시는 심각하지 않을 수 없다.

예) Mary began a book.

또한 단어의 창조적 사용(creative use)에 대한 대처능력도 없다[Puts93]. 위의 예문에서 "begin"은 "동작성 체언"을 요구한다. 그러나 "book"은 "동작성 체언"이 아니기 때문에 분석에 실패하게 된다. 그러나 위의 문장은 의미해석이 가능하다. 문장의 올바른 해석을 위하여 기존의 어휘지식 표현체계에서의 "책을 목적어로 취함"을 "begin2"를 새롭게 정의해야 한다. 이처럼 예기치 않은 경우에 대한 대처능력이 없어, 시스템의 견고성(robustness) 관점에서 큰 문제가 된다.

(2) 단일 시소러스 체계

일본의 대표적인 시소러스인 大野(81)의 "類語新辭典"에서 동물에 대한 분류는 다음과 같다.

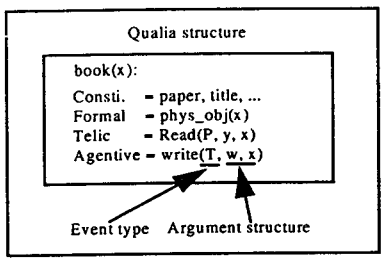
동물->(생물, 동물, 어류, 충류, 기관, 각미, 근골, 내장, 알, 성)

위와 같이 기존의 대다수 시소러스는 단일 시소러스로서, 분류체계의 명확한 기준이 없다. 즉 상위/하위 관계, 부분/전체 관계, 연상관계 등이 복합적으로 적용되어 각 노드를 연결하는 링크들의 의미가 때로는 불명확하고 일관성이 없는데, 이는 분류기준이 불확실함을 나타낸다. 분류기준이 불명확하고 다양한 관점이 복합적으로 적용이 되어 있으면, 실제 분류에 있어서 일관성을 유지하기가 어렵게 된다. 즉 사전구축의 과정에 많은 애로가 따르고, 또한 저장된 정보자체가 체계적이지 못하기 때문에 정보의 이용도 효율적으로 이루어 지기 힘들다.

3.2. 해결책으로서의 Generative Lexicon(GL)

앞에서 언급된 문제점들을 해결하기위한 어휘지식 표현체계를 위하여 본 논문에서는 Pustejovsky가 제안한 Generative Lexicon(GL)을 도입한다. GL을 도입하는 이유는 기존의 어휘지식 표현체계들 중에서 GL이 단어의 의미를 보는 관점이 가장 효율적이고 자연스럽기 때문이다. 즉 어휘 의미 구별이 단순한 선택적 제한(selectional restriction)을 통하여 이루어지는 것이 아니라 효율적인 메카니즘 하에서 이루어지느데, 이는 멀티 시소러스 체계에 근거하여 어휘 의미를 보다 체계적으로 표현하고 있기 때문이다.

3.3. GL의 어휘 의미 표현을 위한 네 가지 구조(4 levels of word sense representation)



[그림 1] Generative Lexicon의 단어 의미 표현 구조

[그림 1]은 GL에서의 어휘 의미의 표현 구조를 나타낸다. GL은 선택적 제한이 아닌 다음과 같은 네 가지 구조를 가지고 단어의 의미를 표현한다.

- 매개변수 구조(Argument structure): 서술어가 요구하는 매개변수들의 특성을 표현한다. 일종의 서술어의 격틀(case frame)형태로 볼수 있다.
- 서술어 구조(Event structure): 서술어 타입이 가지는 의미를 계층구조로 표현한 것으로서, 기초 사건 유형은 상태(state, S), 과정(process, P), 변환(transition, T) 등으로 분류된다.
- 의미특성 구조(Qualia structure): 체언성 단어의 의미특성을 표현한다.
- 어휘특성 상속 구조(Lexical inheritance structure): 단어의 의미들은 상호간의 관계를 기반으로 특성을 주고 받는다.

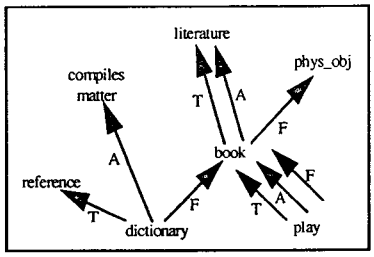
3.4. 멀티 시소러스 체계(multi-lattice for inheritance)

GL에서 멀티 시소러스 체계가 구체적으로 적용되는 부분은 의미특성 구조(Qualia structure) 이다. 이것은 역할이라는 측면에 주안점을 두고 각 의미의 특성을 표현하고 있다.

주로 체언성 단어의 의미 특성을 표현하는 구조로서 기존의 어휘지식 표현체계보다 체언성 단어의 의미 기능을 강화하게 된다. 의미특성 구조가 표현하는 역할들은 다음과 같다.

- 구성요소 관점의 역할(constitutive role): 어휘 의미가 어떤 의미들로 구성되었는가를 나타낸다.
- 정형적인 관점의 역할(Formal role): 다른 단어의 의미들과 구별되는 특성을 나타낸다.
- 주요기능 관점의 역할(Telic role): 단어 의미의 주요 기능이나 그 목적을 나타낸다.
- 생성 근원 관점의 역할(Agentive role): 단어 의미가 생성되는 관점이나 그 근원을 표현한다.

Pustejovsky 는 효과적인 관점들의 채택을 위하여, 기존에 존재하던 시소러스 체계를 연구하여 위와 같은 관점의 시소러스 체계를 채택하고, 그 유용성을 보이고 있다[Pust93]. 본 논문에서는 이러한 관점을 그대로 수용하기로 한다.



[그림 2] GL의 어휘특성 상속 구조

[그림 2]에는 의미특성 구조와 연관되어 어휘특성 상속 구조가 표현되어 있다. 각각의 영문 대문자는 위에 언급된 역할들의 영문 표현의 첫 자들을 나타낸다.

4. 관계기호의 추출에 효과적인 어휘지식 표현체계의 설계

본 연구에서 제안하는 어휘지식 표현체계는 언어 독립적인 중간언어의 개념어휘 표현을 위한 체계이다. 따라서 GL을 도입함에 있어서 중간표현이라는 목적을 고려하여야 한다. 본 절에서는 이러한 목적의 반영을 위하여 중간표현에 효과적인, 특히 각 어휘 의미들의 문장에서의 의미적 역할(관계기호)을 파악하는데 유용한 시소러스 체계를 설계하고, 이 시소러스 체계를 기반으로 GL의 어휘지식 표현체계에 따라 개념기호의 지식을 표현하는 방안에 대하여 논한다.

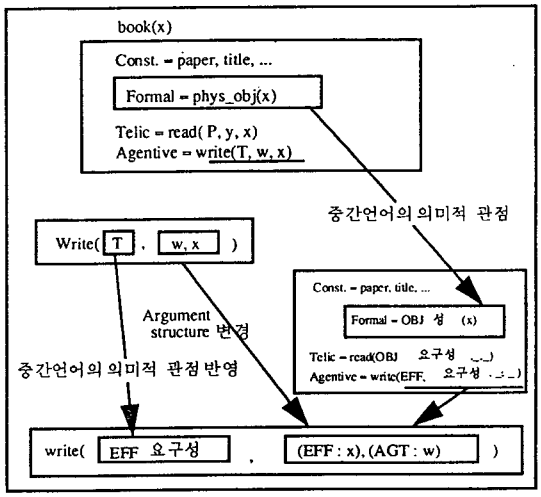
4.1. 어휘지식 표현체계의 설계를 위한 고찰

4.1.1. Generative Lexicon의 도입

GL은 어휘지식 표현체계에 있어 기존의 어휘지식 표현체계가 가지는 문제점을 효율적으로 극복할 수 있으나 그 표현체계에 담겨 있는 정보는 단순히 구문적 관점만이 반영되고 있어, 이러한 구문적 관점을 중간언어의 개념간 의미적 관계를 위한 의미적 관점으로 변환할 필요가 있다.

- 서술어 구조 및 의미특성 구조의 수정

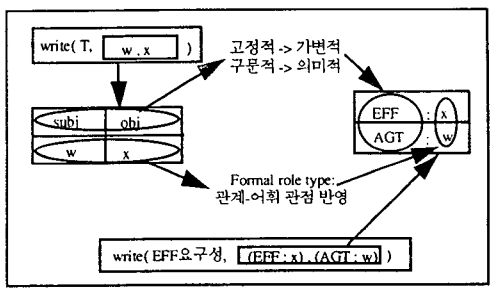
[그림 3]에는 "write"의 이벤트 타입인 "T"가 "EFF" 요구성으로 바뀌고 있다. 기존 GL의 이벤트 타입은 서술어 구조에서의 의미를 나타내고 있는데, 이러한 서술어 구조가 구문적 관점을 반영하고 있다. 따라서 이러한 구문적 관점의 서술어 구조를 중간표현에 적합하게 재설계하고 그 결과를 이벤트 타입에 반영해야 한다. 또한 [그림 3] 에는 의미특성 구조(Qualia structure)의 정형적 관점의 역할(Formal role)인 "phys_obj"이 "OBJ 성"으로 바뀌고 있다.



[그림 3] GL의 적용 방안

이 또한 서술어 구조의 경우처럼 중간표현의 관점을 반영한다는 점에서 그 변경이 요구되는 것이다.

- 매개변수 구조의 수정



[그림 4] 매개변수 구조의 수정

[그림 3]에 표현된 매개변수 구조의 변경에 대하여 세부적으로 표현하면 [그림 4]과 같다. 구문 분석의 측면에서는 매개변수들에게 부여될 역할(syntactic role)만을 표현하는 것으로 매개변수 구조는 충분하다. 그러나 중간표현의 경우, 일단 부여될 역할(semantic role)이 40가지 이상이라, 필수격을 기존 GL에서처럼 위치에 내포시키는 것이 불합리하다. 그리고 매개변수의 특성도 중간 표현의 관점을 고려하여 나

타내야 하는 것이다.

4.1.2. 관계기호 관점의 시소러스 체계

기계번역에서의 어휘지식표현체계는 원시언어의 문장으로부터 중간언어의 표현을 추출함에 있어 다음과 같은 2가지 작업이 용이하도록 설계되어야 한다.

- 정확한 개념기호의 선택
- 선택된 개념기호의 정확한 역할 파악, 즉 개념간 올바른 의미관계의 설정

그러나 기존의 GL은 중간언어에 대한 고려가 없었으므로 다음과 같은 단계를 거쳐, 새로운 어휘지식 표현체계를 설계한다.

- **관계기호 집합의 확정:** 기존 중간언어들의 연구결과를 토대로 하여 개념간 의미관계를 나타내기 위한 관계기호 집합을 확정한다.
- **관계기호 집합에 대한 분석:** 확정된 관계기호들을 분석하여, 충분한 이해를 거친다.
- **관계기호 관점의 시소러스 체계 설계:** 관계기호들에 대한 충분한 이해를 기반으로 관계기호 관점의 시소러스 체계를 설계한다. 이 때, 체언성 개념기호와 용언성 개념기호를 구별하여 시소러스 체계를 설계하는데, 이는 Generative Lexicon의 서술어 구조와 의미특성 구조를 대응시키기 위함이다.
- **새로운 시소러스 체계를 Generative Lexicon에 적용:** 새롭게 설계된 시소러스 체계를 GL에 적용하여, 중간언어의 개념기호가 중간표현에 유용한 정보를 체계적으로 제공할 수 있도록 한다.

4.2. 관계기호 관점의 시소러스 체계 설계

본 절에서는 기존의 중간언어들을 비교, 분석하여 확정된 관계기호의 집합으로부터 관계기호 관점의 새로운 시소러스 체계를 설계하는 과정과 결과를 제시한다.

4.2.1. 관계기호 집합의 분석

용언성-용언성:

AND(AND), BUT(BUT), CMP(CoMParison), CON(CONdition), GOA(GOAl), MEA(MEAn), OBJ(OBJect), PAR(PARtner), REA(REAson), SUP(SUPposition), TIM(TIMe)

용언성-체언성:

AFF(AFFected), AFO(Artificial FOrce), AGT(AGent), BEN(BENefactive), CHA(CHAracterized), CRI(CRIterion), CSE(CauSE), CTE(ConTEnt), DAT(DATive), DEG(DEGree), DIR(DIRection), DIS(DIStance), DUR(Duration), EFF(EFFected), ELM(ELeMent) EXI(EXIstent), EXP(EXPeriment), FCS(FoCuS), GOA(GOAl), HRR(HeaReR), INS(INStrument), LOC(LOCation), MAN(MANner), MAT(MATerial), MEA(MEAn), MTE(MaTching Entity), MTG(MaTching Goal), MGL(Mental GoaL), MOD(MODifier), NAM(NAMe), OBJ(OBJect), PAR(PARtner), QUA(QUAntity), REA(REAson), RUT(RoUTe), SCN(SCeNe), SOR(SOuRce), STP(STandPoint), STA(STAtus), TIM(TIMe), TAR(TARget)

체언성-체언성:

AND(AND), CMP(CoMParision), ELM(ELeMent), MAT(MATerial), POS(POSsesor), QUA(QUAntity), STA(STAtus)

위 결과는 본 논문의 관계기호 집합을 분류한 것이다. 의미관계를 맺어주는 두 개념기호간의 체언성, 용언성 특성을 고려하여 분류하였다. 관계기호 관점의 시소러스 체계는 체언성 개념기호와 용언성 개념기호로 구별하여 설계한다. 이것은 체언성 개념기호의 의미 역할(semantic role) 관점에서 새로운 시소러스 체계를 구축함을 의미한다.

이러한 작업을 위하여 "용언성-체언성"으로 분류된 관계기호들을 다시 분류할 필요가 있다.

누가(Who):

주체: AFO, AGT, CHA, EXI, EXP

비주체: BEN, CSE, DAT, HRR, MTG, STA

무엇을(What):AFF, CTE, EFF, MGL, NAM, OBJ

어디서(Where): DIR, DIS, LOC, RUT, SCN, SOR, TAR

언제(When): DUR, SOR, TIM, TAR

왜(Why): REA, GOA

어떻게(How): CRI, DEG, ELM, FCS, INS, MAN, MAT, MTE, MEA, PAR, QUA, STP

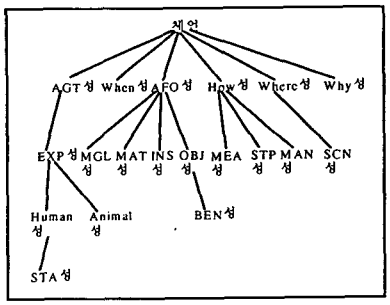
위의 용언성-체언성 분류는 인간이 문장을 분석하는 방법인 6하 원칙에 의거하여 분류한 것이다. 분류에 나타난 결과를 기반으로 본 연구에서는 관계기호 관점의 시소러스 체계를 설계하였다.

4.2.2. 체언성 개념기호의 시소러스 체계 설계

본 절에서는 관계기호 분류를 기반으로 체언성 개념기호의 시소러스 체계를 설계하였는데, 설계의 기준은 다음과 같다.

- 문장에서 다른 어휘들과의 관계보다는 어휘 자체의 기능 및 특성을 참고하여 설계하였다.
- 관계기호 간의 관계를 고려하여 설계하였다.
- 타 기능어에 의하여 구별이 가능한 역할들은 하나의 단위로 취급하였다.
- 일반적인 역할은 설계 과정에 반영하지 않았다.

위에 언급된 기준으로 분류한 체언성 개념기호의 시소러스 구조는 [그림 5]에 제시된다. 각 노드의 이름으로 관계기호가 사용되고 있는데 이는 역할의 관점에서 설계된 시소러스체계에 대한 이해도를 높이기 위한 목적을 반영한 결과이다.



[그림 5] 체언성 개념의 시소러스 체계

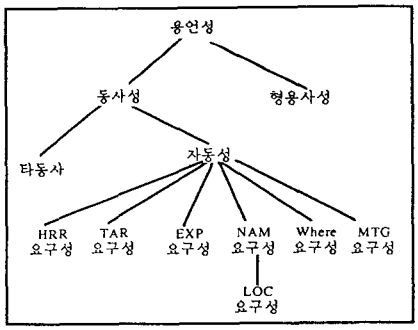
시소러스 체계의 노드에 해당하는 의미에 대하여 일부분을 언급하면 다음과 같다.

- 체언성 : 체언성 개념기호
- AGT 성: 체언성, AGT 역할을 할 수 있는 개념기호
- EXP 성: 체언성, AGT 성, EXP 역할을 할 수 있는 개념기호

4.2.3. 용언성 개념기호의 시소러스 체계 설계

관계기호의 분류를 기반으로 용언성 개념기호의 시소러스 설계기준은 다음과 같다.

- 의미 역할의 관점에서 요구되는 필수격을 기반으로 한다.
- 체계적인 한국어 문형을 바탕으로 한다[강은 93].



[그림 6] 용언성 개념의 시소러스 체계 일부

이같은 기준을 기반으로하여 설계된 용언성 개념기호의 시소러스 체계의 일부가 [그림 6]에 제시되었다.

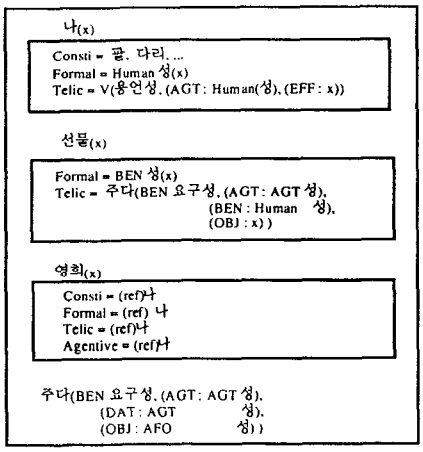
요구되는 필수격에 대한 표현 기호는 앞에서 설계된 체언성 개념기호의 시소러스 체계의 표현 기호와 동일하다. 용언성 분류는 실제 [그림 6]보다 더 상세하게 나누어 진다.

4.3. 새로운 시소러스 체계의 적용 및 그 결과의 이용

본 절에서는 Generative Lexicon에 관계기호 관점의 시소러스 체계를 적용한 예와 이같은 GL 지식을 이용하여 입력 예문을 실제 중간표현으로 변환하는 과정을 설명하기로 한다.

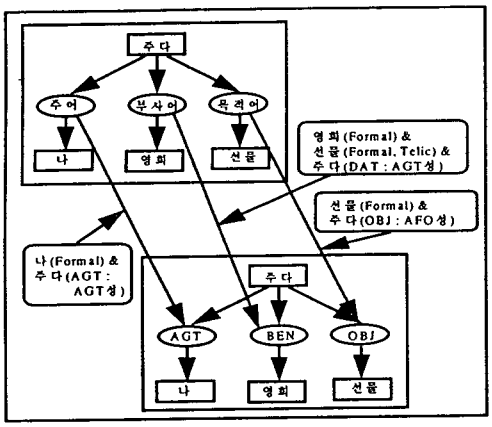
예) 나는 영희에게 선물을 주었다.

이 문장을 분석하기 위하여 지식베이스로부터 필요한 정보를 제공받아야 한다. 제공되는 정보는 관계기호 관점의 시소러스 체계가 적용된 Generative Lexicon의 표현체계에 따라 지식베이스에 저장된 정보이다. 실제로 주어지는 정보의 예가 [그림 7]에 표현되었다. 그림에 표현된 정보를 보면, 일단 체언성 개념기호와 용언성 개념기호가 다른 구조를 보이고 있다.



[그림 7] "나는 영희에게 선물을 주었다"의 중간표현을 위한 정보

그리고 "선물"이라는 개념기호는 구성요소 관점(Constitutive role)과 생성근원 관점(Agentive role)이 표현되어 있지 않은데, 이것은 개념기호를 네 가지 관점에서 표현 가능한 관점이나 필요한 관점에 한하여 표현하면 된다는 것을 의미한다. 그리고 "영희"의 경우, 네 가지 관점의 특성이나 기능을 모두 "나"로 부터 상속받고 있다. "나"의 경우를 살펴보면, 기존의 Generative Lexicon의 표현 내용에 관계기호 관점의 시소러스 체계가 적용되어 있다. 위에 표현된 정보가 구문분석 결과로부터 중간표현을 유도하는데 어떤 역할을 하는지에 대한 표현은 [그림 8]에 나타나 있다.



[그림 8] "나는 영희에게 선물을 주었다"의 중간표현 과정

본 연구의 관계기호 관점의 분류 체계는 문장분석시 의미적 역할을 효율적으로 파악하려는 목적이다. 이런 관점에서 [그림 8]에 제시된 중간표현은 의미적 역할의 파악에 초점이 맞추어져 있다. 아울러 구문적인 역할(주어, 부사어, 목적어)를 의미적 역할(AGT, BEN, OBJ)로 변환하는데 어떤 정보가 이용되는지 명확하게 표현된다.

●"주어"를 "AGT"로 변환

"나"의 정형적 관점의 역할 정보와 "주다"의 매개변수 구조 중, "AGT : AGT성" 정보가 이용되어 "주어"역할이 "AGT" 역할로 파악되었다.

●"목적어"를 "OBJ"로 변환

"선물"의 정형적 관점의 역할 정보와 "주다"의 매개변수 구조 중, "OBJ : AFO성"정보가 이용되어 "목적어" 역할이 "OBJ"역할로 파악되었다.

●"부사어"를 "BEN"으로 변환

"나"의 정형적 관점의 역할 정보로부터 상속된 "영희"의 정형적 관점의 역할 정보와 "주다"의 매개변수 구조 중, "DAT : AGT" 정보가 이용되어, "부사어" 역할이 "DAT" 역할로 파악되었다가, "선물"의 정형적 관점의 역할 정보 및 주요 기능 관점의 역할 정보가 이용되어 "BEN" 역할로 정정이 되었다.

이용된 정보를 살펴볼 때, 주목이 되는 부분이 "BEN" 역할을 파악할 때 사용되고 있는 "선물"에 관한 정보이다. 체언성 개념기호의 기능이 강화되어 올바른 역할이 파악될 수 있도록 하고 있다.

[그림 8]에 표현된 과정으로 중간표현이 이루어지려면, 설계된 정보를 효율적으로 이용하는 부분이 첨가되어야 한다. 즉, 본 논문의 개념기호 체계를 효율적으로 이용할 수 있도록 규칙(rule)이나 제어 시스템이 개발되어야 한다. 이에 대한 지속적인 연구가 요구된다.

5. 새로운 어휘지식 표현체계에 대한 평가

본 절에서는 관계기호 관점의 시소러스 체계에 대한 평가로서 설계된 어휘지식 표현체계에 대한 평가를 대신한다.

5.1. 정보 부여의 용의성

[강은 93]에서 추출한 100 개의 문장에 나타난 각 어휘들에 대하여 정보를 부여하는 작업을 하였다. 대부분의 어휘들에 있어서, 정보의 부여는 비교적 용이하였다. 이러한 용이성은 GL의 어휘지식 표현체계가 가지는 장점으로, 관계기호 관점의 시소러스 체계 하에서도 그대로 유지되고 있는 것이다. 부여된 정보들을 이용하여 중간표현한 예제로서 일부분은 아래와 같다.

- 책장(LOC)에 책(EXI)이 많이 있다[자동성, (EXI:체언성)].
- 철수(AFO)가 직장장(STA)으로 임명되었다[자동성, (AFO:Human 성), (STA:STA)].

그러나 다음에 언급되는 경우에 있어서는 정보 부여 방안에 대한 연구가 요구됨이 파악되기도 하였다.

● 언어 종속적인 어휘

"그분은 나와 구촌간이다"라는 문장의 "구촌"이라는 한국어 종속적 어휘는 정보부여에 있어서 주의(attention)가 요구된다.

● 복합어휘

"감방안은 세상과 격리되었다"라는 문장의 "감방안"은 "감방"과 "안"이 복합된 어휘로서 이에 대한 정보의 부여는 주의해야 한다.

● "명사+이다"형의 어휘

"여기는 가을이다"라는 문장의 "가을이다"에 대한 정보 부여는 현재의 어휘지식 표현체계 하에서 거의 불가능하다.

● 동사의 사동형이나 피동형 어휘

동사의 사동형이나 피동형은 본동사의 의미와 이벤트 타입 및 매개변수 구조에 있어서 차이를 보인다. 이러한 차이를 고려하여 주의깊게 정보를 부여해야 한다.

위에 언급된 사항 중에서 언어 종속적인 어휘의 경우나 복합적 어휘의 경우는 언어 종속적인 어휘에 대한 대처 방안을 연구하여 정보를 부여하였다. 동사의 사동형이나 피동형의 경우도 "명사+이다"형의 경우처럼 연구해야 할 부분이다. 일반적으로 동사의 사동형이나 피동형에 대한 정보는 본동사에 표현하는 것이 타당하기 때문이다. 경우에 따라서는 독립적으로 사전에 등록하는 것이 바람직하기도 하지만, 대부분의 동사는 그렇지가 않다.

5.2. 정보자체의 효율성

본 절에서는 설계된 표현체계를 기반으로 해당 문장을

중간표현하여 본 결과에 대하여 논한다.

구분		횟수	비율
중간표현성공		83회	83%
중간표현실패	실패유형1	2회	2%
	실패유형2	7회	7%
	실패유형3	8회	8%
	합계	17회	17%

[표 1] 100개 문장에 대한 실험 결과[표 1]에는 [강은93]의 100개 문장을 실험한 결과가 제시되고 있다. 17개의 문장이 중간표현에 실패를 하여, 83%의 성공률을 보이고 있다. 실패의 경우, 그 원인들은 다음과 같은 유형들로 구분될 수 있다.

실패유형 1:

부여된 정보가 부족한 경우이다. "저는 용수라고 합니다"는 문장에서 "용수"는 "NAM"역할을 하고 있는데, 이를 파악하기 위한 정보가 충분히 부여되지 못하고 있는 경우이다.

실패유형 2:

"그는 그림에 호감을 가진다"라는 문장에서 올바른 중간표현이 가능하려면, "호감을 가진다"가 하나의 단위로 작용을 해야 한다. 이런 유형의 오류는 "CTE+하다"형에서 빈번하게 발생할 것으로 생각된다. 의미적으로 하나의 단위를 이루는 어휘를 합성개념으로 다룰 수 있는 방안이 연구되어야 해결이 가능한 오류이다.

실패유형 3:

5.1절에서 언급한 것처럼, "명사+이다"형의 어휘에 대한 정보 부여에 오류가 생겨서 실패한 경우이다. "철수는 동무들과 싸움질이다"라는 문장을 정확하게 중간표현하려면, "싸움질이다"라는 어휘의 정보가 정확하게 부여되어 있어야 한다. 그러나 현재의 개념기호 체계에서는 이 부분에 대해서 고려하지 않고 있기 때문에, 차후의 연구가 요구되는 부분이다.

세가지 실패유형 중에서 실패유형3이 가장 큰 비율을 차지하고 있는데, 이는 검증을 위한 문장 100개가 "조선어 문형 연구"에서 추출된 것과 많은 관계가 있을 것이다. 실제로 가장 많은 비율을 차지하리라 여겨지는 실패유형은 실패유형2일 것으로 추정된다.

더욱 정밀한 중간표현을 위해서는 관계기호 관점의 시소러스 체계를 실제 상황에 맞게 지속적으로 수정, 보완하여야 할 것이다.

6. 결론 및 향후 연구 방향

본 논문에서는 중간언어의 개념기호의 의미를 위한 어휘지식 표현체계를 제안하였다. 특히 관계기호 관점의 시소러스 체계를 설계하고, 이를 기반으로 한 정보를 Generative Lexicon의 어휘지식 표현체계로 표현하는 방안을 제시하였다. 검증과정에서 5.2절에서 언급한바로 여러가지 부족한 점들을 발견하였다. 또한 관계기호 관점의 시소러스 체계에 대하여 더욱 세밀한 분류작업, 각종 기능어들에 대한 어휘지식 표현방안에 대한 연구와 관계기호를 추출하는 과정에서 사용될 규칙들에 대한 연구가 요구되었다. 이러한 문제점들에 대하여 향후 많은 연구가 필요된다.

7. 참고문헌

[CICC93] CICC(Center of the International Cooperation for Computation), Interlingua(Final Edition), Japan, 1993.

[Dorr93] Bonnie Lean Dorr, Machine Translation: A View from the Lexicon, MIT Press, 1993.

[Dorr94] Bonnie Lean Dorr, Jye-hoon Lee, Sunki Suh, "Development of Cross-Linguistic Syntactic and Semantic Parameters for Parsing and Generation," MD/CS-TR-3233/UMIACS-TR-94-26, 1994.

[Hut92] W. J. Hutchins, Harold L. Somers, An Introduction to Machine Translation, Academic Press, 1992.

[Nir87] Sergei Nirenburg, Machine Translation: theoretical and methodological issue, Cambridge University Press, 1987.

[Nir89] Sergei Nirenburg, "Knowledge-Based Machine Translation," Machine Translation, Vol 4, No.1, pp.5-24, 1989.

[Pust91] James Pustejovsky, "The Generative Lexicon," Computational Linguistics, Vol. 17, No.4, pp.409-441, 1991.

[Pust93] James Pustejovsky, "Lexical knowledge representation and natural language processing," Artificial Intelligence, Vol. 63, No.1-2, pp.193-223, 1993.

[Yao90] Zhang Gui-Ping, Yao Tain-Shun, "The Chinese Analyzer for Multilingua Machine Translation," Computer Processing of Chinese & Oriental Languages, Vol. 5, No.1, pp. 1-8, 1990.

[龜井 88] 龜井, 中間表現, NEC 技術報告書, Japan, 1988.

[강은93] 강은국, 조선어 문형연구, 도서출판 박이정, pp.97-308, 1993.

[김상93] 김상국, 박찬호,"중간언어에 기반한 기계번역 시스템의 설계," 제5회 한글 및 한국어 정보처리 학술발표논문집, pp.521-525, 1993.

[박찬95]박찬모, 이종혁, 이근배, 한-중 기계번역 시스템에 관한 연 구, 한국과학재단최종보고서, 1995.5.

[大野 81]大野普, 浜西正人, 類語新辭典, 角川書店, 東京, 1981.