한-독 기계번역의 대역어 선택 문제에 대하여

홍문표 (한국전자통신연구원)

I. 서론

20세기말부터 시작된 인터넷/인트라넷 등과 같은 정보통신 분야의 눈부신 발전은 21세기에 들어 더욱 더 그 발전 속도가 가속화되고 있다. 이러한 정보/통신 분야의 발전은 세계를 시간과 공간적 제약이 없는 하나의 커뮤니티로 재구성하고 있다. 이러한 과정에서 기계번역 등과 같은 언어처리 기술은 큰 몫을 차지하고 있다고 본다. 이러한 시대적 흐름에 발맞추어 이미 유럽, 미국, 일본 등에서는 물론이고 한국에서도 한국전자통신연구원 ETRI 등과 같은 정부출연연구소는 물론이고 대기업 연구소, 벤쳐기업 연구소 등에서도 기계번역에 관한 연구가 활발히 이루어지고 있다. 현재는 그 대상 언어가 한국어와 영어, 일본어, 중국어 등이지만 향후 수년 내에 독일어, 스페인어, 프랑스어 등과 같은 유럽어로 확대될 것으로 보인다. 한-독 기계번역은 1990년대 중반부터 기본적인 연구가 시작되어 현재까지는 프로토타입 수준의 시스템들만이 개발된 상태이다.[1] 본 논문은 향후 한-독 기계번역 시스템이 상용화될 단계에 이르렀을 때 부딪히게 될 가장 큰 문제점의 하나인 대역어 선택 Zielwortselektion 문제에 대해 다룬다.

영어권 연구자들에 의해서는 'target word selection'으로 불리기도 하는 대역어 선택의 문제는 많은 경우 번역의 대상이 되는 단어가 다의적이라서, 번역될 때 여러 가지로 번역이 가능함으로 인해 발생한다. 예를 들어 독일어 단어 'Tor'는 크게 축구에서 '공이 골라인을 넘어가는 것'이라는 의미와 '사람 등이 지나다니는 통로'라는 두 가지 의미를 지니고 있다. 이에 따라 이 단어가 한국어로 번역이 될 경우 각각 '골인', '문' 등으로 상이하게 번역된다. 그러나 실제 기계번역

1 한-독 기계번역 시스템에 대해서는 Choi (1995), Hong (2001) 참조.

뿐만 아니라 일반 번역에서도 발견할 수 있는 현상은 어떠한 단어가 출발언어 Quellsprache의 측면에서 볼 때에는 다의적이지 않지만, 목표언어 Zielsprache의 측면에서 보면 여러 가지 문맥에 따라 상이하게 번역되어야만 하는 경우이다.

많은 경우에 대역어 선택의 문제와 의미모호성 해소 문제가 혼동되는 경우가 있다. 이는 대역어 선택을 위하여 대부분의 경우 의미모호성 해소가 전제되어야 하기 때문이다. 그러나 의미모호성 해소가 올바른 대역어 선택으로 곧바로 연결되는 것은 아니다.

본 논문에서는 이에 따라 의미모호성 해소와 대역어 선택 단계를 분리한 방법론을 제시하고자 한다. 즉, 분석 단계에서 번역 대상 단어의 의미를 결정하는 모듈과 생성단계에서 대역어를 선택하는 모듈을 분리하였다. 의미분석을 위해서는 일반적으로 많이 사용되는 의미제약 기반 방법론을 확대한 Streiter (1998)의 SDL 이론이 적용되었으며, 대역어 선택을 위하여는 코퍼스로부터 추출한 경험적 지식인 단어사용 빈도 정보가 사용되었다. 본 논문에서는 그 연구대상을 동사로만 제한하며, 본 논문에서 제시하는 방법론들은 부분적으로 CAT2 시스템을 기반으로 구현되었다.

본 논문의 구성은 다음과 같다. II장에서는 서론에서 잠시 언급한 단일어 모호성과 번역 모호성 개념을 좀 더 자세히 다룬다. III장에서는 의미 모호성 해소 단계와 대역어 선택 단계를 분리한 방법론을 제안한다. IV장에서는 III장에서 기술한 방법론을 실제 구현한 결과에 대해 소개한다. 끝으로 V장에서는 본 논문을 정리하고 향후 연구 방향을 제시한다.

II. 단일어 모호성/번역 모호성

아이러니컬하게도 많은 경우 '모호성 Ambiguität'이라는 단어는 그 자체가 모호하게 사용된다. 예를 들어 우리는 'Absatz'라는 단어가 두개의 의미, 즉 i) 'ein unter der Ferse befindlicher Teil des Schuhs', ii) 'mit einer neuen Zeile beginnende Unterbrechung in einem sonst fortlaufenden Text'를 가지고 있을 때, 이 단어는 '모호하다'라고 말한다. 또한 우리는 '만들다'라는 한국어 단어가 독일어에서 경우에 따라 'herstellen', 'produzieren', 'zubereiten', 'zimmern', 'brennen',

'drehen', 'bauen' 등으로 번역될 때 이 단어가 '모호하다'라는 말을 한다. 이는 '만들다'의 개념이 위 독일어 단어들의 의미영역과 부분적으로 일치하기 때문일 것이다.

여기서 우리는 과연 '만들다'라는 단어가 정말로 '모호한' 단어인가라는 문제제기를 할 수 있다. 왜냐하면 대다수의 우리는 한국어 모국어 화자로서 '만들다'라는 단어를 '모호하다'라고 생각하지 않을 것이기 때문이다. 만약 이 단어가 의미적으로 모호한 단어라면 위 독일어 단어들 간의 의미를 구별해주는 의미체계는 과연 무엇일까 하는 의문이 생길 수 있다.

이와 같은 이유로 본 논문에서는 단일어 모호성 Monolinguale Ambiguität과 번역 모호성 Übersetzungsambiguität 개념을 구분하고자 한다. 단일어 모호성 현상은 어떠한 단어가 출발언어에서 그 자체적으로 의미적 모호성을 나타내어서, 각각의 의미에 대해 다른 대역어가 연결되는 것을 말한다. 물론 각각의 의미에 대해 반드시 하나의 대역어만이 대응되는 것은 아니다.

이에 반해 번역 모호성 현상은 하나의 의미에 대해 여러 가지 대역어가 올 수 있는 현상을 뜻한다. 다음의 예들은 한-독 번역 상에서 나타나는 단일어 모호성과 번역 모호성의 예들이다.

단일어 모호성

'가깝다'
의미1: '거리'
가까운 길 : der *kürzere* Weg
가장 *가까운* 정류장 : die *nächste* Station

의미2: '시간'
세시가 *가까웠다* : Es ist *bald* drei
시험 때가 *가까웠다* : Die Prüfung ist *nahe*

의미3: '관계'
가까운 친구 : der *vertraute* Freund
가까운 친척 : der *nahe* Verwandte

번역 모호성

'만들다'

상자를 *만들다* : eine Kiste *zimmern*
나무로 책상을 *만들다* : aus Holz einen Tisch *herstellen*
자동차를 대량으로 *만들다* : Autos in Massen *produzieren*
쌀로 술을 *만들다* : aus Reis Reisbrannwein *brennen*
음식을 *만들다* : ein Essen *zubereiten*
영화를 *만들다* : einen Film *drehen*[2]
배를 *만들다* : ein Schiff *bauen*
계약서를 *만들다* : einen Vertrag *entwerfen*
초고를 *만들다* : einen Entwurft *anfertigen*
길을 *만들다* : eine Straße *anlegen*
회사를 *만들다* : eine Firma *aufbauen*
만든 *이야기* : eine *erfundene* Geschichte
하느님이 하늘과 땅을 *만드셨다* : Der Gott *schuff* Himmel und Erde
음악을 *만들다* : Musik *komponieren*

단일어 모호성의 경우는 일반적으로 기계번역 상에서 의미 모호성 해소가 출발언어 분석 단계에서 이루어진다. 그러나 번역 모호성의 경우에는 출발언어 분석 단계에서 모호성을 해소하는 방안도 있으며, 분석단계에서는 모호성을 해소하지 않고 그대로 유지하다가 변환 Transfer단계에서 모호성을 해소하는 방안도 있다.[3]

본 논문에서는 그러나 번역 모호성의 경우도 변환단계에서 모호성 해소 절차를 거치지 않고, 여러 개의 후보 단어가 대역어로 추천되는 방법론을 채택한다. 문맥에 가장 알맞은 대역어는 코퍼스로부터 추출된 경험적인 지식에 의해 선택된다.

---

2 '영화를 만들다'에서 '만들다'의 대역어로는 'drehen'은 물론 'machen', 'produzieren' 등도 가능하다. 위 예에서는 상위 개념으로 분류할 수 있는 'machen', 'produzieren' 등보다는 좀 더 세부적인 하위 개념어인 'drehen'을 사용하였음을 밝혀둔다.

3 Hauenschild (1986:178): "Alle Mehrdeutigkeiten, die schon innerhalb der Ausgangssprache (unabhängig von einer Zielsprache) auftreten, sollen im Rahmen der Analyse behandelt werden, während Ambiguitäten, die sich erst im Hinblick auf die jeweilige Zielsprache ergeben, im Transfer zu bearbeiten sind"

다음 장에서는 본 논문에서 제시하는 단어 의미 모호성 해소 방안과 코퍼스 기반 대역어 선택 방법론에 대하여 자세히 언급한다. 또한 단어 의미 모호성 해소 방안에 대한 여러 가지 기존 연구에 대해 소개하며, 왜 본 논문에서는 Hauenschild (1986)에서와 같이 번역 모호성을 변환단계에서 다루지 않고 생성단계에서 다루는지를 설명할 것이다.

III. 대역어 선택 방법론 제안

이 장에서는 본 논문에서 제시하는 대역어 선택 방안이 소개된다. 대역어 선택 방안은 '단어 의미 모호성 해소', '변환', '대역어 후보로부터의 선택'과 같이 3단계로 구성된다. 첫 번째 단계에서는 단일어 모호성을 나타내는 단어들의 경우 의미 파악이 이루어진다. 두 번째 단계에서는 번역 모호성을 나타내는 단어들이 대역어 후보그룹으로 변환된다. 세 번째 단계에서는 대역어 후보그룹에서 해당 문맥에 가장 적합한 하나의 대역어가 선택된다.

1. 단어 의미 모호성 해소에 대한 기존 연구

본 논문에서 소개할 단어 의미 모호성 해소 방안에 대해 언급하기 전에, 기계번역 분야에서 일반적으로 사용하고 있는 언어학 기반 의미 모호성 해소 방안을 소개하고, 각각의 장단점에 대해 논의한다. 언어학에 기반한 의미 모호성 해소 방안은 첫째, 선택제약기반 방법론, 둘째, 지식기반 방법론 등이다. 먼저 선택제약기반 방법론에 대해 살펴본다.

1) 선택제약기반 방법론

세부적으로 어떠한 방법론을 도입하고 있건 선택제약기반 방법론은 Katz & Fodor (1963)의 방법론과 크게 다르지 않다고 볼 수 있다. 이 방법론은 각 단어의 의미 특성 혹은 의미 자질을 의미 모호성 해소를 위한 기준으로 삼고 있다. 이 방법론의 기본 아이디어는 모든 단어는 의미적으로 적합한 문장을 이루기 위해서는 특정한 의미 자질을 가지는 단어들과만 공기할 수 있다는 데 있다. 이러

한 의미 자질들은 일반적으로 동사의 하위범주 사전에서 논항 위치에 제약 Constraints으로서 명기된다. 이러한 의미자질들은 분석과정에서 여러 개의 의미 후보로부터 적합한 의미를 찾아내기 위하여 사용된다. 다음 예문을 보자.

(1) Durch eine vorausschauende und koordinierende Raumordnungspolitik sind die dafür notwendigen *Gebiete* in der freien Natur zu sichern und *auszubauen*[4]

위 예문에서 'Gebiet'라는 단어는 그 의미가 모호하다고 할 수 있다. 'Gebiet'가 의미할 수 있는 것은 첫째로 'eine geographische Region'이고, 둘째로는 'ein Sachbereich'이다. 그러나 'ausbauen'이라는 동사의 의미를 생각하면 첫 번째 의미, 즉 'eine geographische Region'이 이 문장에서 사용되는 의미임을 알 수 있다. 그 이유는 'ausbauen' 동사가 그 목적어로서 [+raümlich]라는 자질을 가진 명사만을 취하기 때문이다. 이러한 의미제약 기반 방법론은 매우 많은 시스템에서 사용되었으며, 현존하는 상용 기계번역 시스템의 거의 대부분에서 사용되고 있다. 그럼에도 불구하고 순수 의미제약 방식은 다음과 같은 약점을 지니고 있다.

첫째, 선택제약기반 방법론만으로는 만약 의미 모호성 해소 단서가 문장단위를 넘어서 위치할 경우, 의미파악이 어렵다.

둘째, 의미제약을 일차원적으로 표현하는 기존의 방법론으로는 단어 의미의 변화무쌍하고도 생성적인 성질을 파악하기 어렵다. 다음의 예문이 그러한 경우를 나타낸다.

(2) a. Mein *BMW* verbraucht zu viel Benzin
b. *BMW* hat am Freitag ein neues Angebot für seine hochverschuldete britische Konzerntochter Rover erhalten

위 예문에서 a의 BMW는 BMW 회사에서 생산한 자동차를 의미하며, b의 BMW는 BMW 회사의 대표나 대변인을 의미한다.

셋째, 이러한 방법론의 가장 큰 문제점은 의미자질이 종종 언어학적인 근거를 상실한 채 단지 대역어 선택만을 위하여 임의로 만들어진다는 점이다. 다음 예문

4 이 예문은 Laffling (1991:37)로부터 차용하였음.

을 보자.

(3) a. ein Paket erhalten - 소포를 받다/*수신하다
b. ein Telegram erhalten - 전보를 받다/수신하다

동사 'erhalten'은 한국어에서는 '받다', 혹은 '수신하다' 등으로 번역된다. 위 예문에서 a와 b에서의 'erhalten'의 의미차이는 첫 번째 경우에는 목적어가 [+Waren] 의미자질을 갖는다는 점이며, 두 번째 경우에는 [+Nachricht]라는 자질을 갖는다는 점이다. 이러한 의미자질의 차이로 a 와 b에서의 'erhalten'의 의미차이가 드러나며, 각각에 해당하는 올바른 대역어가 선택된다. 그러나 이때의 문제점은 독일어 분석을 할 때 이러한 의미자질이 과연 언어학적 타당성을 지니고 있느냐하는 문제이다. 예를 들어 독-영 기계번역의 경우에는 'erhalten'이 모두 'receive'로 번역이 될 것이므로 이러한 자질이 불필요하게 된다.

2) 지식기반 방법론

Carbonell et al. (1981), Schank (1982) 등에 의해 소개된 지식기반 의미분석 방법론은 스크립트 script 혹은 프레임 frame이라고 부르는 개념에 기반한다. 스크립트는 개념간의 의존관계를 표현하는 일종의 형식언어이다. 이러한 스크립트는 매우 한정된 분야 (예를 들어 '레스토랑에서의 식사 상황', '자동차 사고' 등)에 대한 세계지식을 담고 있다. 이 방법론의 기본 아이디어는 인간은 의미 해석을 위해 자신 주위의 세계에 대한 지식에 의존한다는 것이다. 예를 들어 '레스토랑에서의 식사 상황'이라는 스크립트가 활성화되어 있을 경우 'note'라는 영어 단어는 'Bankscheck'의 의미로 해석될 것이다.

이러한 방법론의 문제점은 어떠한 스크립트가 활성화되어 있을 경우, 그 스크립트에 적합하지 않은 의미는 처음부터 아예 고려대상에서 제외된다는 점이다.[5] 다음의 예문은 이러한 문제점을 명확하게 보여준다.

(4) Erst wenn sich die Erkenntnis durchsetzt, dass mit Bildung$_i$ mehr gemeint ist -

5 Weber (1981:327): "Lesarten, die nicht offensichtlich in den vorgegebenen Textrahmen passen, werden von Anfang an unterdrückt oder gar nicht vorgesehen"

> nämlich die $Bildung_{ii}$, Formung und Entwicklung der gesamten menschlichen Persönlichkeit mit allen sozialen, emotionalen und geistigen Beziehungen zur Umwelt - kann die berufliche Qualifikation wieder zu ihrem Recht kommen.[6]

이러한 문장에 대해 현재 '교육'이라는 스크립트가 활성화되어 있다면, 이 예문에서 $Bildung_{ii}$의 의미, 즉 'Schaffung' 혹은 'Entwicklung'은 제대로 그 의미가 분석되지 못할 것이다.

2. SDL 기반 단일어 의미 모호성 해소

본 절에서 소개하는 SDL 기반 의미 모호성 해소 방안도 선택제약기반 방법론 범주에 속한다. 앞서 살펴본 선택제약기반 방법론의 단점에도 불구하고 이 방법론을 택한 이유는 다음과 같다.

첫째, 선택제약기반 방법론은 실제 시스템을 구현하는데 매우 용이하다. 의미제약들은 통사규칙에 제약으로서 쉽게 접목될 수 있다.

둘째, SDL 기반 방법론에서는 대역어 선택이 의미 모호성 해소 단계에서 이루어지지 않는다. 단일어 모호성을 갖는 단어들에 대해서만 의미 모호성 해소가 이루어지므로, 앞서 살펴본 [+Nachricht] 등과 같은 언어학적 동기가 결여된 임의적인 의미제약을 가정할 필요가 없다.

셋째, 단어의 의미 결정을 위해 국소 문맥만을 고려함으로써 생기는 정보의 불충분함은 통계 정보를 이용한 경험적 지식을 사용함으로써 상쇄될 수 있다.

본 논문에서 의미 모호성 해소를 위해 도입하는 SDL (Semantic Description Language) 이론은 Streiter (1998)에 의해 소개되었다. SDL은 단어의 의미를 여러 가지 측면에서 사전에 기술할 수 있는 장점을 지니고 있다. 즉, 어떠한 개념을 그 개념의 물질적 속성, 추상적 속성, 사태 구조 Ereignisstruktur 뿐만 아니라, 'Agentivity', 'Thematicity', 'Directivity'라는 측면과, 주제 영역, 스타일 속성 등에 관해서도 기술할 수 있는 장치를 제공한다. 이를 통해 (2)의 예에서 나타난 것과 같은 단어 의미의 가변성 등을 기술할 수 있다. 'Buch'에 대한 사전 엔트리의 예를 통해 이러한 특징들을 살펴보기로 한다.[7]

---

6 이 예문은 Laffling (1991:50)에서 차용하였음.

(5)
'Buch' = {lex=buch, head={cat=n, ehead={sem={ccr={t=instr}, abs=text, evt=nil}}}}.[].

'Buch' 같은 단어들은 의미적으로 모호성을 나타낸다.[8] 이러한 단어들은 한편으로는 독서를 위한 도구로서의 구체적인 의미를 나타내며, 다른 한편으로는 책에 담겨있는 내용 등과 같은 추상적인 의미를 나타내기도 한다. 이러한 본질적인 모호성은 사전 구조에 ccr(concrete)과 abs(abstract)라는 자질을 통해 나타나 있다.

의미제약은 동사의 하위범주 사전에 명시된다. 모든 동사들은 이와 관련하여 다음과 같은 프로세스 타입에 따라 분류된다: 'Perception', 'Mental', 'Communication', 'Exchange', 'Transformation', 'Aspectual', 'Activity', 'Movement', 'Relation'.

각각의 프로세스 타입은 자신의 논항들에 대한 의미역을 한정하고 있으며, 이를 통하여 논항 명사의 의미를 간접적으로 제한한다. 즉, 예를 들어 'Activity'타입을 갖는 동사에 의해 하위범주되는 명사들은 의미적으로 'Activity'타입과 모순되지 말아야 한다. 어떠한 명사가 어떤 프로세스 타입과 모순되느냐 아니냐에 대한 정보는 사전에 바로 기록될 수 있다. 앞서 본 'Buch'는 의미적으로 CAT2 시스템 상에서 'TEXT'라는 의미 매크로에 할당되는데, 이 매크로의 내부 구조는 다음과 같다.

(6) define(TEXT, 'gl(=(info;act;emot;comm;exc;asp), evt=nil, ag=(nil;comm;asp;info;emot;perc;trans), abs=trans, ccr={t=instr, sex=nil, state=s}').

이 의미 매크로에 따르면 이에 속하는 단어들은 comm, asp, info, emot, perc, trans의 프로세스 타입을 갖는 동사들의 ag(Agens)로서 사용될 수 있다. 이 제약은 CAT2 시스템 상에서 다음과 같은 규칙을 통해 표현된다.

(7)
process_type=
{head={proc=PROC}}>>

7 기술의 편의를 위해 논의와 상관이 있는 자질들만 표기하였음.
8 Pustejovsky (1995)는 이러한 모호성을 'complementary polysemy'라고 부르고 있다.

{sc={a=({r=a, head={ehead={sem={ag={PROC}}}}
;{r=t, head={ehead={sem={th={PROC}}}}
;{r=g, head={ehead={sem={gl={PROC}}}}},
b=({r=a, head={ehead={sem={ag={PROC}}}}
;{r=t, head={ehead={sem={th={PROC}}}}
;{r=g, head={ehead={sem={gl={PROC}}}}},
c=({r=a, head={ehead={sem={ag={PROC}}}}
;{r=t, head={ehead={sem={th={PROC}}}}
;{r=g, head={ehead={sem={gl={PROC}}}}})}}.[].

이 규칙이 의미하는 것은 모든 동사는 자신의 프로세스 타입을 자신의 논항 위치에 통합 Unifikation 연산을 통하여 복사한다는 것이다.

이상에서 논의한 내용들을 모두 고려하여 '만들다'의 사전 구조를 정의하면 다음과 같다.

(8)
'만들다'=
{lex=mantulta, slex=mantulta, h={cat=v, proc=trans},
{sc={a={AGENT,h={eh={sem={ag=trans}}}},
b={THEME, h={eh={sem={th=trans}}}}}} &
{trans={de=({t=(anfertigen; anlegen; aufbauen; bauen; brennen; drehen; entwerfen; erfinden; herstellen; komponieren; produzieren; schaffen; zimmern; zubereiten)})}} &
{}.[].

(8)의 사전구조에 따르면 한국어 동사 '만들다'는 'agentivity'와 'thematicity' 자질이 모두 Transformation 프로세스 타입과 모순되지 않는 명사들만을 하위범주한다. 이를 통하여 다음 예문과 같은 경우 '만들다' 동사의 단일어 의미모호성이 해소될 수 있다.

(9)
a. 그 음악이 나를 기분 좋게 만든다 : Die Musik macht mich glücklich
b. 그 회사가 그 자동차를 만든다 : Die Firma produziert das Auto

(9)의 예들은 '만들다' 동사의 단일어 모호성을 보여준다. (9.a)에서는 목적어의 감정의 변화를 의미하며, (9.b)에서는 '만들다'가 '생산'의 의미를 나타낸다. 이 경우 (9.a)에서는 '생산'의 의미는 제외된다. 왜냐하면 주어 '음악'은 다음의 사전구조에 나타난 것과 같이 Transformation 프로세스 타입의 주어로서 사용될 수 없기 때문이다.

(10) {ag=(asp;info;nil), gl(=(info;emot;comm;exc;act;asp), th(=move, ccr={t=instr, sex=nil}, abs=(text; period), evt=(nil; {aspect=dur})}

이 절에서는 단일어 모호성을 SDL이라는 장치에 기반하여 해소할 수 있음을 보였다.

3. 통계정보 기반 대역어 선택

III장 2절에서 소개한 방식으로 단일어 의미 모호성이 해소되면 분석된 의미에 대응하는 대역어가 생성 과정을 위해 선택된다. 본 논문에서는 번역 모호성을 지니는 단어들의 대역어 선택을 위하여 변환단계에서 입력 단어에 대해 하나 이상의 대역어 후보가 넘어가게 된다. Palmer et al. (1999)에서도 이와 유사하게 입력 단어에 대해 복수개의 대역어 후보가 생성 단계로 넘어간다. 다음 영-한 예문을 보자.

(11)
a. *receive* the supply <=> 공급물을 *받다*
b. *receive* the telegram <=> 전보를 *수신하다*

Palmer et al. (1999)는 위 예에서 'receive'의 올바른 대역어 선택을 위하여 분석 단계에서 [+Nachricht], [+Waren] 등과 같은 영어에 대해 언어학적 동기가 결여된 자질을 설정하는 것은 옳지 못하다고 주장한다. 이에 따라 이들은 입력 단어를 하나의 특정 대역어로 변환하는 것이 아니라, 변환단계에서는 여러 가지의 대역어 후보군으로 변환한다.

(12) receive <=> {받다, 수신하다}

한국어에서 '받다'와 '수신하다' 동사는 목적어로 취하는 명사에 대해 '받다'의 경우에는 의미적 제약이 없으나, '수신하다'의 경우에는 [+Nachricht] 와 같은 자질을 요구한다.[9] 따라서 생성단계에서 비로소 [+Nachricht], [+Waren]가 한국어 단어에 대한 의미제약으로서 사용된다. 그러나 이러한 방법론도 많은 문제점을 가지고 있다. II장에서 소개되었던 '만들다'의 예를 다시 보도록 하자.

(13)
a. 배를 만들다 : ein Schiff bauen
b. 자동차를 대량으로 만들다 : Autos in Massen produzieren
c. 만든 이야기 : eine erfundene Geschichte
d. 영화를 만들다 : einen Film drehen
e. 음악을 만들다 : Musik komponieren

Palmer et al. (1999)의 방법론을 따를 경우 위의 예에서 예를 들어 '이야기'와 '영화'를 의미적으로 구분지어 독일어에서 상이한 대역어로 번역되게끔 하는 변별자질이 과연 무엇인지 분명하지 않다. 물론 이러한 자질을 어떠한 방식으로도 정의할 수는 있겠지만 이는 언어학적 동기가 결여된 임시방편적인 해결책이 아닐 수 없다. 결국 우리는 대역어 선택의 문제에 있어 의미 자질 이외에 다른 요소를 고려해야만 한다는 결론에 다다른다.

대역어 선택에 있어서 주제 영역 Subjektdomäne과 단일어 코퍼스 monolinguale Korpora로부터 추출한 단어 빈도 정보는 중요한 역할을 할 수 있다. Streiter et al. (1999)는 특정 주제 영역에서의 단어 빈도 정보가 올바른 대역어 선택의 가능성을 높임을 실험을 통해 입증하였다. 예를 들어 영-독 기계번역에서 영어 단어 'match'는 독일어로 'Spiel' 또는 'Streichholz'로 번역될 수 있다. 그러나 이 단어가 '스포츠'

9 '수신하다' 동사가 반드시 [+Nachricht] 자질을 갖는 명사와 결합하는 것만은 아니다. 이 동사는 익명의 심사위원께서 제시하신 '전파를 수신하는 안테나'의 예에서 볼 수 있는 바와 같이 '전파'와 같은 명사들과도 공기할 수 있다. 그러나 어떠한 경우이건, 위 예에서 보이고자 하는 바는 이러한 의미자질에 기반한 대역어 선택 방법은 언어학적으로 타당성이 결여된 의미자질을 임의로 상정해야 한다는 데 있으므로, '수신하다' 동사가 반드시 [+Nachricht] 자질의 명사만을 목적어로 취하는 지에 대한 논의는 피하기로 한다.

주제 영역에서 사용되었다면 독일어로는 'Spiel'로 번역될 가능성이 높다. 이 정보는 실제로 스포츠 영역의 코퍼스에서 사용되는 단어들의 빈도를 조사해보면 'Spiel'이 'Streichholz'보다 훨씬 빈번히 출연하기 때문에 단순 빈도 조사만으로도 얻어질 수 있다.

이 방법론은 주제 영역의 단어 빈도 정보를 고려하지 않는 다른 방법론들보다는 우수한 성능을 보이지만, 스크립트를 사용하는 지식기반 의미 모호성 해소 방법론과 유사한 문제점을 가지고 있다. 즉, 특정 주제 영역이 선택되면 어떤 단어에 대해 특정 대역어만이 추천되게 된다. 즉, 위의 예에서 보면 스포츠 영역에서는 'match'가 'Streichholz'로 번역될 가능성이 완전히 배제되는 문제점이 있다.

지금까지 대역어 선택에 관련된 두 가지 연구 흐름, 즉 선택 제약 기반 방법론과 단어 빈도 정보 기반 방법론을 Palmer et al. (1999)와 Streiter et al. (1999)의 연구를 통하여 살펴보았다. Palmer et al. (1999)의 방법론은 목표언어에 기반한 선택제약을 사용하고, 국소 문맥을 고려하지만 의미자질의 선정이 어렵고, 광역 문맥이 고려되지 못하는 문제점을 나타내었다. 이에 반해 Streiter et al. (1999)의 방법론은 단일어 코퍼스로부터 추출한 통계정보를 사용하여, 광역 문맥을 반영하지만, 국소 문맥이 고려되지 못하는 단점을 나타냈다.

지금까지의 논의를 종합해볼 때 새로운 방법론은 광역 문맥을 고려하면서도 통사적, 국소 의미적 문맥을 함께 고려해야할 것이다.

따라서 단어의 빈도수를 조사할 경우에도 단순히 출현 빈도만을 따질 것이 아니라 구조적 문맥도 고려한 상태에서 출현 빈도를 조사하여야 한다. 이와 관련하여 Lee et al. (1999)은 흥미로운 연구 결과를 발표하였다. 이들은 영-한 기계번역에서 대역어 선택을 위해 100만 어절 규모의 한국어 코퍼스를 사용하였다. 이들의 방법론에 따르면 영어 단어의 의미가 분석 단계에서 결정되면 이에 해당하는 한국어 대역어 후보들이 매칭된다. 대역어 선택을 위한 정보는 단일어 코퍼스로부터 추출한 통계값으로부터 가져오는데, Streiter et al (1999) 등과 다른 점은 단순 빈도 정보를 가져오는 것이 아니라 동사와 특정 구조 관계에 있는 단어들의 빈도만을 고려한다는 점이다. 즉 술어-주어, 술어-목적어 등의 관계만을 고려 대상으로 한다.

우리는 이러한 방법론을 한-독 기계번역에도 적용하고자 한다. 그러나 Lee et al. (1999)의 방법론을 완전히 수용하지는 않는다. 우선 Lee et al. (1999)은 단일

어 모호성 해소를 위해서 기계가독형 사전으로부터 추출한 연어 정보를 이용하지만, 본 논문에서는 앞장에서 소개한 SDL을 이용한다. 또한 대역어 선택을 위한 통계정보를 추출함에 있어서도, 단순히 구조 문맥 정보만을 이용하는 것이 아니라 이러한 구조 문맥 자체도 주제 영역이라는 광역 문맥에 다소 영향을 받기 때문에, 주제 영역별로 구조 문맥을 추출하여 사용하고자 한다. 앞서 반복되어 사용된 '만들다'의 예를 가지고 대역어 선택을 위한 우리의 방법론을 설명하면 다음과 같다.

- 한국어 동사 '만들다'가 SDL을 통해 의미 분석이 되면 (즉 단일어 모호성이 해소되면), 결정된 의미에 대한 대역어 후보가 변환단계에서 생성 모듈로 넘어온다 ('produzieren', 'herstellen', 'zubereiten', 'anlegen', 'erfinden', 'drehen', ...)

- 주제영역에 따라 태깅된 단일어 코퍼스로부터 구조 문맥에 따른 빈도 정보를 추출한다. 빈도 정보는 다음과 같은 형태를 띠게 된다: (pred-obj, herstellen, auto, 10), (pred-obj, herstellen, lkw, 3), (subj-pred, hyundai, herstellen, 2). 예를 들어 (pred-obj, herstellen, auto, 10)는 'herstellen'과 'auto'가 술어-목적어 관계로 코퍼스 상에서 10번 나타난다는 것을 의미한다.

- 만약 독일어 출력 문장의 목적어가 예를 들어 'Auto'라면 한국어 동사 '만들다'의 대역어는 위 통계결과에 따라 'herstellen'으로 결정된다.[10]

---

10 통계정보 추출을 위해 사용된 코퍼스에서 'Auto'와 'produzieren'이 'herstellen'보다 많이 사용되었다면, '만들다'의 대역어로 'produzieren'이 제시될 것이다. 이와 같이 통계정보는 도메인 별로 축적된 코퍼스로부터 추출되므로, '자동차를 만들다'에서 '만들다'의 대역어로서 'herstellen', 또는 'produzieren'이 선택될 것이다. 그러나 어떠한 경우에건, 'zubereiten', 'drehen', 'komponieren' 등은 대역어로 제시되지 않을 것이다.
현재의 사전구조에서는 대역 후보들이 상호 연관 관계에 대한 정보가 없이 평면적으로 나열되어 있다. 그러나 익명의 심사위원께서 제시하신 바와 같이 동사들 간의 상위, 하위, 동의 등과 같은 의미 관계를 설정함으로써 보다 효과적인 대역어 선택이 가능하리라고 본다. 이러한 관계를 설정하게 되면 통계 기반 방법론의 가장 큰 문제점인 데이터 부족 문제 (data sparseness problem)을 해결하는데 다소 도움이 되리라 생각한다. 즉, 예를 들어 코퍼스에서 'auto'와 공기하는 '만들다'의 대역 후보가 없거나 매우 적을 때, 신뢰하기 힘든 통계정보에 의존하여 번역을 시도하기 보다는 상위 개념어인 'machen'이나 'produzieren' 등을 사용하면 의미 전달에 충분한 번역 결과를 얻을 수 있을 것이다. 이러한 아이디어를 제공해주신 익명의 심사위원께 지면을 빌어 감사의 뜻을 전한다.

다음 장에서는 이러한 통계정보를 코퍼스로부터 뽑기 위한 구현 알고리즘과 간단한 실험 결과가 소개된다.

IV. 구현

III장에서 소개한 통계정보 추출 방법론에 기반하여 통계 정보 추출 프로그램을 구현하였다. 통계 정보 추출 프로그램은 Perl 프로그래밍 언어를 사용하여 구현되었다. 본 장에서는 간단한 프로그램 알고리즘과 통계추출 실험 결과만에 대해 언급한다.

구조 문맥을 고려한 통계정보를 얻기 위해서는 코퍼스를 형태소 태깅 Tagging하여야 한다. 이를 위한 독일어 형태소 태거로써 본 논문에서는 독일 자르브뤼켄 대학의 IAI 연구소에서 개발한 Mpro라는 형태소 분석기를 사용하였다.[11] Mpro는 입력문의 단어들에 대해 원형을 찾아 태깅하는 기능 이외에 복합명사 분석 및 간단한 구조 분석 Shallow Parsing 등과 같은 부가적인 기능을 가지고 있다. 지면 관계상 Mpro의 태깅 결과는 생략하도록 한다. 태깅 결과로부터 III장에서 언급한 통계정보를 얻기 위한 알고리즘은 다음과 같다.

- 문장의 시작점과 끝점을 찾는다
  - Mpro 분석 결과 모든 문장은 <I> 태그로 시작한다
- 문장의 주동사 Hauptverb를 찾는다
  - {c=hs}를 찾는다
  - {c=hs} 노드에 의해 지배되는 것 중 {c=verb, vtyp=fiv} 자질이 있는 노드를 찾는다
  - 통계 테이블의 동사 칼럼에 해당 단어의 lu 값을 적는다
- 문장의 목적어를 찾는다
  - 주동사의 오른쪽에 위치한 {c=np} 노드를 찾는다
  - {c=np} 노드에 의해 지배되는 것 중 {c=noun}를 찾는다

11 Mpro 시스템에 대한 자세한 내용은 Maas (1998) 참조.

- 통계 테이블의 목적어 칼럼에 해당 단어의 lu 값을 적는다
- 문장의 주어를 찾는다[12]
- 주동사의 왼쪽에 위치한 {c=np} 노드를 찾는다t
- {c=np} 노드에 의해 지배되는 것 중 {c=noun}를 찾는다
- 통계 테이블의 주어 칼럼에 해당 단어의 lu 값을 적는다

위와 같은 알고리즘의 성능을 실험하기 위하여 간단한 테스트를 실시하였다. 실험은 Daimler Chrysler사에서 제공한 자동차 수리 매뉴얼 코퍼스를 이용하여 이루어졌다. 코퍼스는 약 1만 2천 단어 분량의 비교적 작은 크기였다. 이 코퍼스로부터 '만들다'의 독일어 대역 후보에 속하는 'produzieren', 'herstellen', 'bauen', 'drehen', 'zubereiten' 등이 어떤 명사와 특정 구조 관계 (술어-주어, 술어-목적어)에 속하는지, 그리고 몇 번이나 그러한 관계가 출현하는 지에 대한 빈도를 추출하였다. 이 실험 결과 다음과 같은 통계 정보가 획득되었다.[13]

동사	목적어	빈도
herstellen	Active Body Control	8
herstellen	neuer CL	2
herstellen	s-Klasse	2
herstellen	mercedes-Benz	1
produzieren	luftfederung AIRMatic	5
produzieren	drehstab-Stabilisator	3
bauen	V8-Triebwerk	2
bauen	zwölfzylinder-Motor	1

12 Mpro가 기반한 문법은 LFG 등과는 달리 주어, 목적어 등의 개념을 사용하지 않는 간단한 구구조 문법 Phrasenstruktur Grammatik이므로, 문장의 주어와 목적어를 찾기 위해 어순만을 고려하는 비교적 간단한 방법을 취하고 있다.

13 실제 통계 정보를 추출할 때는 술어-주어의 관계도 고려하나, 본 논문에서는 명료한 예시를 위해 술어-주어 관계를 생략하였음. 또한 위 통계 결과에서도 분석 오류 등의 이유로 일종의 노이즈 정보가 추출된 경우도 있었으나, 결과에 영향을 줄 정도는 아닌 작은 분량이었으므로, 위 표에서는 제시하지 않았음을 밝혀둔다.

위 통계 결과에 따르면 물론 코퍼스가 매우 작은 크기였음을 감안하더라도, '자동차' 분야에서 '만들다'라는 동사의 대역어로 고려될 수 있는 것은 'herstellen', 'produzieren', 'bauen' 등임을 알 수 있다. 또한 만약 '다임러 크라이슬러사는 8기통 엔진을 만든다'라는 문장이 입력된다면, 이 시스템은 'produzieren'을 '만들다'의 대역어로 선택할 것이다. 이상으로 보인 바와 같이 대역어 선택은 코퍼스에 기반하여 다이내믹한 방식으로 이루어질 수 있다. 즉, 매번 코퍼스를 갱신하여 통계 결과를 추출한 후 이것을 변환 및 생성 정보로써 사용할 수 있다. 현재까지는 CAT2 시스템과 통계 추출 프로그램을 연결할 수 있는 API가 개발되지 않아, CAT2 시스템을 이용한 직접 번역 결과를 제시할 수는 없으나, 본 논문에서는 대역어 선택을 위한 방법론을 제시하는데 그 의의를 두었다.

V. 결론

본 논문에서는 기계번역의 가장 어려운 문제 중의 하나인 대역어 선택 문제를 다루었다. 대역어 선택의 문제는 일반적으로 입력문에 나타나는 단어의 의미가 다의적으로 사용될 때 발생한다. 이와 같은 이유로 많은 기존 연구에서 대역어 선택의 문제와 의미 모호성 해소의 문제가 혼동되어왔다. 그러나 본 논문에서는 대역어 선택의 문제를 일으키는 경우를 '단일어 모호성'이라는 개념과 '번역 모호성'이라는 개념을 도입하여 구별하였다. 단일어 모호성의 문제를 해결하기 위해 확장된 선택제약 방식인 SDL 이론을 도입하였다. 이 이론을 통해 단어의 가변적이고 생성적인 특성을 기술할 수 있게 하였다. 번역 모호성의 문제를 해결하기 위해 기존의 다른 연구들과는 달리 출발언어나 목적언어에만 부합하는 의미적 자질을 설정하지 않고, 주제 영역과 구조 문맥을 고려한 통계적 정보를 사용하였다. 또한 알고리즘의 제시와 실험을 통해 이러한 통계정보가 추출 가능하며, 실제 유용하게 응용될 수 있음을 보였다.

그러나 이러한 통계 기반 방법론이 갖을 수 있는 가장 큰 문제점은 잘 알려진 바와 같이 데이터 부족 문제이다. 이러한 문제점을 해결하기 위해 분야별로 잘 정제된 대용량 독일어 코퍼스를 확보하는 것이 본 논문에서 소개한 방법론의 효과를 극대화하기 위해 절대적으로 필요하다. 또한 술어-주어, 술어-목적어 등과

같은 구조 문맥을 고려한 통계정보를 정확히 추출하기 위해서는 앞서 소개한 휴리스틱 기반 구조 분석 방식이 아닌 문장 내의 문법 관계를 정확히 파악할 수 있는 독일어 분석기에 기반하여야 하리라고 본다.

참고문헌

Carbonell, J., R. Cullingford & A. Gershman (1981): Steps toward knowledge-based machine translation, in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 3 (4), S.376-392

Choi, S.K. (1995): Unifikationsbasierte Maschinelle Übersetzung mit Koreanisch als Quellsprache, IAI Working Papers 34, Saarbrücken

Hauenschild, C. (1986): KIT/NASEV oder die Problematik des Transfers bei der maschinellen Übersetzung, in Batori, I & H. Weber (Hgg.). Neue Ansätze in Maschineller Sprachübersetzung: Wissensrepräsentation und Textbezug (=Sprache und Information Band 13), Tübingnen, Niemeyer

Hong, M.P. (2001): Linguistische Probleme in der maschinellen Übersetzung Koreanisch-Deutsch, Dissertation, Universität des Saarlandes

Katz, J.J. & Fodor, J. (1963): The structure of a semantic theory, Language 39, S.170-210

Laffling, J. (1991): Towards high-precision Machine Translation- based on Contrastive Textology, 7 Distributed Language Translation, Foris Publications, Berlin/New York

Lee, H.A., J.C. Park & G.C. Kim (1999): Lexical Selection with a Target Language Monolingual Corpus and an MRD, in Proceedings of the 8thInternational Conference on Theoretical and Methodological Issues in Machine Translation (TMI99), Chester, England

Maas, D. (1998): Multilinguale Textverarbeitung mit MPRO, in. G. Lobin et al. (Hgg.). Europäische Kommunikationskybernetik heute und morgen, KoPäd, München

Palmer, M., C. Han, F. Xia, D. Egedi & J. Rosenzweig (1999): Constraining Lexical Selection Across Languages Using Tree Adjoining Grammars, in TAG3 Workshop Proceedings, CSLI

Pustejovsky, J. (1995): The Generative Lexicon, The MIT Press, Cambridge/London

Schank, R.C. (1982): Dynamic Memory: a Theory of Reminding and Learning in Computers and People, Cambridge-Press.

Streiter, O. (1998):A Semantic Description Language for Multilingual NLP, in Tuscan Word Centre-Institut für Deutsche Sprache Workshop on Multilingual Lexical Semantics

Streiter, O., L. Iomdin, M.P. Hong & U. Hauck (1999): Statistical Support for Rule-Based MT, in Proceedings of the 8thInternational Conference on Theoretical and Methodological Issues in Machine Translation (TMI99), Chester, England

Weber, H.J. (1981): Repräsentation von Alltagswissen und die Auffindung von Sinnzusammenhngen im Text, in Kohrt, M. & Lenerz, J. (Hgg.). Sprache: Formen und Strukturen. Akten des 15. Linguistischen Kolloquiums, S.327-336

Wilks, Y. (1975): An intelligent analyzer and understander of English, in Communications of the ACM 19, 5, S.264-274

Zusammenfassung

Über die Zielwortselektion in der maschinellen Übersetzung Koreanisch-Deutsch

Hong, Munpyo (ETRI)

Die vorliegende Arbeit widmet sich der Problematik bei der Zielwortselektion. Die Probleme bei der Zielwortselektion sind eine der größten Herausforderungen für die maschinelle Übersetzung. Schwierigkeiten werden einerseits dadurch bereitet, dass viele Wörter Mehrdeutigkeiten aufweisen. Um eine richtige Zielwortselektion zu gewährleisten, muss die richtige Bedeutung eines Wortes in einem Kontext ermittelt werden. Probleme ergeben sich andererseits daraus, dass ein Wort, auch wenn seine Bedeutung in einem Kontext geklärt ist, mehrere Übersetzungsmöglichkeiten haben kann. Hierbei wurde die Auffassung vertreten, dass die Disambiguierung eines Wortes

nicht automatisch zur Zielwortselektion führen kann. Aus diesem Grunde wurde ein dreistufiges Modell zur Zielwortselektion vorgestellt: Disambiguierung, Transfer in die Übersetzungskandidaten, statistisch gesteuerte Zielwortselektion. Zur Disambiguierung des Quellwortes wurde die SDL (Semantic Description Language) herangezogen. Nach der Disambiguierung wird das Quellwort nicht direkt zu einem Zielwort, sondern in eine Gruppe der Übersetzungskandidaten überführt. Bei der Auswahl unter den Übersetzungskandidaten wurde das empiristische Wissen herangezogen, das von monolingualen Korpora erworben worden war. Dabei erfolgt das Wortfrequenz-Zählen unter Bezugnahme auf die grammatischen Beziehungen. Die dynamische Zielwortselektion ist dazu geeignet, einen wesentlichen Beitrag zur Verbesserung der Übersetzung zu leisten. Im Rahmen der vorliegenden Arbeit konnte jedoch ausschließlich das Programm zur Erstellung der Statistik implementiert werden. Die entsprechende Modifikation des CAT2 Systems bleibt eine Herausforderung für die Zukunft.

주제어: 대역어 선택, 의미 모호성, 한-독 기계번역, CAT2

Schlüsselbegriffe: Zielwortselektion, Ambiguität, Koreanisch-Deutsch MÜ, CAT2

필자 E-Mail: hmp63108@etri.re.kr

투고일: 2004. 5. 31 / 심사일: 2004. 6. 21 / 심사완료일: 2004. 8. 29