2000년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

한영 번역 시스템을 위한 문틀 기반 번역 방식의 실현성 분석

김영길, 서영애, 서광준, 최승권
한국전자통신연구원 컴퓨터소프트웨어연구소
언어공학연구부 지식처리팀
e-mail : kimyk@etri.re.kr, yaseo@etri.re.kr seokj@etri.re.kr, choisk@etri.re.kr

An Analysis of Feasibility of Sentence Frame Based Method for Korean to English Translation System

Young-Kil Kim, Young-Ae Seo, Kwang-Jun Seo, Sung-Kwon Choi
Knowledge Procssing Team, Dept. of Language Engineering , ETRI

요 약

지금까지의 한영 번역 방식은 규칙 기반 방식이 주를 이루었지만 현재 패턴을 이용한 번역 방식이 활발히 연구되고 있다. 그러나 패턴 기반 방식은 그 적용성(Coverage)에 대한 치명적인 단점을 지닌다. 따라서 본 논문에서는 한국어 패턴을 어절 단위의 일반 문틀과 동사구를 중심으로 하는 용언 중심의 문틀로 나누어 각 패턴들에 대한 적용성 및 실현성을 조사한다. 실험은 기존의 형태소 분석기를 이용하여 방송 자막 문장 351,806 문장을 대상으로 자동으로 문틀을 구축하여 4,995 문장의 테스트 데이터에 대한 적용성 검사를 실시하였다. 즉 본 논문에서는 방송 자막 문장을 대상으로 한영 번역을 위한 일반 문틀 및 용언 중심의 문틀 방식의 적용성을 조사하여 문틀 기반 방식의 실현성을 평가하고 앞으로의 한영 번역 시스템 개발 방향을 제시한다.

1. 서론

현재 우리나라를 포함한 일본, 미국, 유럽 등의 세계 각국에서 기계번역에 관한 활발한 연구와 실용화가 이루어지고 있다[1-2]. 국내에서는 영한, 일한 번역 시스템의 개발이 활발하게 진행되고 있으며 현재 한일 번역 시스템의 상품화가 진행되고 있다. 그러나 한국어 구조 분석의 어려움으로 인하여 어순이 다른 영어를 목표로 하는 한영 번역 시스템의 개발은 일부 학교 및 연구소를 중심으로 제한적으로 진행되고 있는 실정이다[3]. 서강대 및 KAIST 를 중심으로 제한된 영역하에서의 한영 대화체 번역 시스템이 개발되었다[4]. 그리고 ETRI 에서는 한-일 호텔 예약 음성 번역 시스템[5]과, 규칙 기반 방식을 이용한 에서로/KE 한영 번역 시스템을 개발한 바 있으며[1-2] 현재 의존 규칙 및 동사구 번역 패턴을 이용한 방송 자막용 한영 번역 시스템을 개발중이다.

지금까지의 한영 번역 방식은 규칙 기반의 자동번역이 주를 이루었다. 그러나 기존의 규칙기반 자동번역은 규칙의 조합적 작동에 의한 번역을 실행하였는데 이 방법의 단점은 규칙의 조합성이 모든 언어현상을 반영하지 못하기 때문에 번역품질의 개선과 유지를 보장할 수 없다. 따라서 이를 극복하기 위한 통계적 기법이나 지식처리 기법을 기반으로 한 언어이해 메커니즘의 규명에 노력이 모아지고 있다[6-7]. 이 중 대표적인 것이 대규모 코퍼스로부터 통계적 기법을 이용하여 언어 현상의 일반화 규칙 추출에 대한 연구와 기존의 번역 예제로부터 번역 패턴을 얻는 패턴 기반 번역에 대한 연구이다. 그러나 패턴 방식은 그 적용성(Coverage)에 대한 치명적인 단점을 지닌다.

영한 번역 시스템을 비롯하여 몇몇 번역 시스템을 대상으로 문틀 패턴을 이용한 번역 방식이 시도되고 있지만 한영 번역 시스템을 위한 한국어에 대한 문틀의 실현성은 아직 검증되지 않았다. 따라서 본 논문에서는 방송 자막 문장을 대상으로 한국어를 대상으로 문틀 방식의 패턴 기반의 번역 방식의 실현성을 조사하고 이에 대한 결과를 분석하여 한영 번역 시스템의 개발 방향을 제시한다.

2. 문틀을 이용한 패턴 번역 방식

문틀을 이용한 패턴 번역 방식은 주로 예제 기반의 구 단위의 문틀과는 달리 문장 전체를 하나의 번역 패턴으로 기술하는 방식이다. 그림 1 은 문틀을 이용한 영한 번역 패턴의 일례를 보이고 있다. 한 패턴은 문장 단위로 구성되며 패턴 매칭을 위한 조건부와 대역 정보 기술을 위한 행위부로 구성된다. 모든 단어를 하나의 패턴 성분으로 기술하지는 않으며 애매성이 없는 단계까지의 구문 분석을 통하여 가장 기본적인 구문단위가 하나의 슬롯을 구성하고 그 슬롯들의 나열로 전체 문틀이 이루어진다.

```
They're working with NATO, but they technically answer to the Russian commanders.
그들은 북대서양 조약 기구와 일하지만, 기술적으로 러시아 사령관에게 대답합니다.

NP1:[ (etype == [cent])] VERB1:[ (etype == [t1]) _AND (eform == [vb]) ] NP2:[ (etype == [prop])] PUNCT1:[ (etype == [cpunct])] CONJ1:[ (etype == [coorv])] NP3:[ (etype == [cent])] VERB2!:[ (etype == [t1]) _AND (eform == [vb]) ] NP4:[ (etype == [comm])]  ->  NP1:[ kcase := [topic] ] NP2:[ kcase := [obj] ] VERB1 CONJ1 PUNCT1 NP3:[ kcase := [topic] ] NP4:[ kcase := [obj] ] VERB2!
```

(그림 1) 영한 번역을 위한 문틀의 일례

한국어 원시 문틀 기술 형태는 다음과 같다. 용언은 VP+인덱스+연결어미 형태로 기술하며 명사 체언구는 NP+인덱스+조사 형태로 기술한다. 그리고 원시문틀은 괄호를 이용하여 구문 구조를 표현한다. 즉 용언이 가지는 명사 체언구들은 하나의 괄호 안에 위치하며 이는 의존 트리 구조와 동일하다.

그림 2 는 한영 번역을 위한 문틀의 일례이다. 한국어 문틀은 하나의 어절이 하나의 슬롯을 구성하는 일반 문틀과 용언구를 중심으로 하는 용언 중심의 문틀로 이루어 질 수 있다. 용언 중심의 문틀에서는 동사구 슬롯외에도 인접하는 동사에 의존하지 않는 명사구 슬롯도 패턴에 나타날 수 있다.

```
입력문 1:
(김대통령은 오늘 (청와대에서 김종필 총리를 만나) 의견을 나누었습니다.)
일반문틀: NP1가 (NP2에서 NP3를 V1)아서 NP4를 V2
용언중심문틀: NP1가 (VP1)아서 VP2

입력문 2:
((김대통령은 오늘 청와대에서 김종필 총리를 만났고) 그들은 서로 의견을 나누었습니다.)
일반문틀: (NP1가 NP2에서 NP3를 V1)고 NP4가 NP5를 V2
용언중심문틀: (VP1)고 VP2
```

(그림 2) 한영 번역을 위한 문틀의 일례

일반 문틀은 번역 패턴의 상세한 기술이 가능하여 구조적인 모호성을 해결할 수 있는 자질들을 기술할 수 있지만 그 적용성이 상당히 떨어진다. 반면 용언 중심 문틀은 그 적용성은 일반 문틀에 비해 높지만 대량의 문틀을 구축하여도 실제 입력문을 대상으로 구조 분석의 애매성을 해결할 수 있는 정보를 제공하지 못한다는 단점이 있다.

3. 한국어 문틀의 실현성 분석

한국어는 언어 특성상 영어와는 달리 부분적으로 구문 분석될 수 있는 부분이 적고 실질 형태소 이외에 기능 형태소(조사, 어미)에서 나타나는 어휘의 변화가 심하기 때문에 문형 판단 기준이 되는 슬롯의 종류가 상당히 많은 편이다.

그리고 본 논문에서 대상으로 하는 뉴스 문장은 한 문장으로 사건의 경위를 압축적으로 설명하는 관계로 그 문장의 길이 또한 일반 텍스트 문장보다 긴 편이다. 따라서 뉴스에서 나타나는 길고 복잡한 문장들에 대한 문형의 수렴성 여부는 문형 기반 번역 방식의 성공 여부를 결정짓는 중요한 요소가 된다. 우리는 이러한 수렴성 여부에 대한 조사를 실현성(Feasibility) 분석이라고 정의하고 현재 확보된 뉴스 코퍼스를 대상으로 문틀을 구축하고 그 데이터에 대해서 문형의 수렴성 여부를 검증한다.

문형 구축 및 구축된 데이터에 대한 수렴성 조사는 형태소 해석기 및 고정 표현 처리기 등에 의해 문틀을 자동으로 구축하는 방법과 문형 구축 도구를 이용하여 제한된 양의 데이터를 구축하여 그 수렴성 여부를 조사하는 방법으로 나눌 수 있다.

이와 같이 문형의 수렴성을 위한 조사를 위해서 두 가지 방법이 제시될 수 있지만 각각 장단점을 지닌다. 수동적인 문형 구축에 의한 방법은 패턴 구축자가 형태소 해석 및 태깅 에러를 수정해 주고 고정표현 처리의 에러 또한 수정할 수 있으므로 정확한 문형의 구축이 가능하다는 장점이 있지만 한국어 문형의 실현성을 조사할 수 있을 정도의 대량의 데이터는 구축할 수 없다는 단점을 지닌다.

그리고 자동적인 문형 구축 방법은 형태소 해석기의 품사 태깅 오류가 존재함으로 인해서 그 문형의 정확성이 떨어지지만 문장 데이터만 있다면 자동으로 모두 문형 패턴으로 만들어 실험의 모집합이 되는 문틀을 대량으로 확보할 수 있다는 장점이 있다. 따라서 한국어 문형 실현성 조사라는 측면에서 자동적인 구축 방법이 타당하다. 그리고 자동적으로 구축된 경우 나타나는 문형의 에러 유형과 그 에러 유형의 일관성으로 인하여 실현성 조사에 미치는 영향이 적은 지 등에 관한 조사가 선행되어야 한다.

본 실현성 조사를 위해서 실제 자동으로 구축한 문형 데이터와 수동으로 구축한 문형 데이터 각 100 문장씩을 분석하여 나타나는 문형 에러 유형을 조사하였고 이러한 에러 유형들로 인하여 실현성 조사에 크게 영향을 받지 않는다는 사실을 알 수 있었다. 따라서 본 실현성 조사는 정확성이 다소 떨어지더라도 대량의 문형을 구축하기 위하여 자동적인 문형 구축 방

법을 사용하며 문형의 수렴성을 조사하여 문틀 방식의 실현성을 판단한다.

자동적인 문형 구축기를 구현하기 위하여 우리말 정보처리 소프트웨어 개발 과제에서 개발한 한영 번역 시스템 에서로 K/E 에서 사용하는 한국어 형태소 및 태거를 이용하고 고정표현 처리기를 추가하여 자동적으로 모집합 및 테스트 데이터로 사용할 문틀을 구축한다.

4. 실험 및 분석

문틀 방식의 실현성 여부를 조사하기 위해서 사용한 모집합 데이터는 MBC 방송자막 대본 (1997 년, 1998 년), KBS 방송 자막 대본 (1998 년 7 월 1 일 - 1999 년 3 월 8 일) 351,806 문장이다. 문틀의 적용성 조사를 위해서 사용한 테스트 데이터는 모집합 데이터에 포함되지 않는 4,995 문장의 데이터로 역시 같은 KBS 방송 자막 대본으로부터 구축되었다. 데이터를 분석한 결과 문장들은 크게 방송체와 대화체 2 가지 형태로 분류할 수 있었으며 방송체는 앵커 문장, 기자 문장, 캐스터 문장으로 구성되어 있으며 대화체는 인터뷰 문장으로 구성되어 있으며 2 가지 종류의 데이터를 모두 사용하였다.

<표 1> 실험 대상 코퍼스

	모집합	테스트
문장 수	351,806	4,995
원문틀 수	351,806	4,995
부분 원문틀 수	254,408	0
평균 문장 길이 (단어수)	12.455	12.736
평균 원문틀 길이 (슬롯수)	6.898	7.162
평균 부분문틀 길이 (슬롯수)	4.467	0

문틀에 나타나는 패턴의 유형을 분석하기 위하여 실험 코퍼스를 대상으로 한국어 원문틀에 나타나는 슬롯들의 통계치를 추출하였다. 전체 슬롯 개수는 3,757,306 개이고 총 슬롯 종류는 946 개 였다. 이는 영한 번역에 사용되는 슬롯에 비하여 상당히 많은 종류의 슬롯이 사용되고 있다는 것이며 한영 문틀의 실현성에 치명적인 결과를 야기시킬 수 있다. 다음 표는 출현하는 프로텍터의 횟수가 70,000 번이 넘는 후보들을 나타내고 있다. "NP 가" "V 다" "NP 를"이 예상대로 가장 많은 빈도로 출현하고 있었으며 한 슬롯이 20% 를 넘어가지 않고 상위 순위에서는 그 빈도가 비교적 고루 분포되고 있으며 이들 슬롯이 전체 빈도의 대부분을 차지하고 있다. 즉 상당히 빈도가 낮은 슬롯들이 다수 존재함을 알 수 있다.

<표 2> 고빈도 슬롯 비율

순위	슬롯 종류	빈도수	비율
1	NP 가	675,324	0.1797
2	V 다	507,279	0.1350
3	NP 를	477,322	0.1270
4	V 는	400,162	0.1065
5	NP 에	190,022	0.0506
6	NP	179,253	0.0477
7	NP 으로	166,877	0.0444
8	V 아	104,622	0.0278
9	NP 이다	100,643	0.0268
10	NP 에서	85,870	0.0229
11	V 고	77,930	0.0207
12	V 를	74,385	0.0198

자동적인 문형 추출 후 새로운 4,995 개의 뉴스 문장들에 대한 완전 매칭, 부분 매칭을 시도하여 문형의 수렴성 여부를 테스트하였다. 즉 적용성 분석을 위해 모집합의 원문틀(351,806)과 부분 원문틀(254,408)의 합인 606,214 개의 문틀에 대해, 테스트 데이터인 4,995 개의 문틀의 매칭 여부를 조사하였다.

전체 4,995 개의 테스트 문틀에 대해 모집합의 원문틀 및 부분 문틀과 완전 매칭된 것은 1,541 개 (30.9%) 이며, 매칭에 실패한 문틀은 3,454 개 (69.1%)이다. 매칭에 실패한 3,454 개의 문틀에 대해서 부분적인 매칭을 테스트하기 위해서 동사 슬롯에 어미가 연결어미인 슬롯을 중심으로 부분 문틀과의 매칭 시작점으로 이용하였다. 이는 단순하게 최장 일치로 테스트 문틀에 대해 매칭을 시도할 때 발생하는 상당수의 의미없는 결과를 배제하기 위함이다.

기본 매칭에 실패한 문틀에 대한 부분 매칭 결과로는 3,454 개의 문틀에 대해 부분 매칭에 성공한 문틀의 수가 752 개 (17.15%)로 나타나며, 부분 매칭에도 실패하는 문틀의 수가 2,702 개 (54.1%)로 나타났다. 즉 부분 매칭을 포함하여 전체적으로 매칭되는 경우가 약 54% 정도인 것으로 판단되어 구축 문형의 적용성이 상당히 떨어지는 것을 알 수 있다. 따라서 문형에서 슬롯들의 종류들을 줄일 수 있는 Grouping 등의 기법이 도입되지 않는 한 매칭율의 향상을 기대하기 어렵다.

<표 3> 일반 문틀 길이별 매칭 결과

슬롯 수	빈도	완전 매칭	부분 매칭	완전+부분
1	609	609	0	609(12.24%)
2	252	245	0	245(4.92%)
3	297	270	2	272(5.47%)
4	328	236	29	265(5.33%)
5	407	236	48	284(5.71%)
6	418	125	111	236(4.74%)
7	434	43	132	175(3.52%)
8	410	10	125	135(2.71%)
9	387	2	108	110(2.21%)
10	342	1	76	77(1.55%)

11	311	0	48	48(0.96%)
12	240	0	33	33(0.66%)
13	157	0	17	17(0.34%)
14	149	0	10	10(0.2%)
15	79	0	6	6(0.12%)
16	58	0	3	3(0.06%)
17	38	0	2	2(0.04%)
18	28	0	2	2(0.04%)
19	18	0	0	0
20 이상	33	0	0	0
총계	4995	1541	752	2293
매칭율		30.9%	15.1%	46%

결론적으로 일반 문틀의 적용성에 대한 분석을 보면 자동으로 구축한 351,806 여개 수준의 원문틀 구축으로 5 단어 이하의 간단한 단문만이 매칭된다는 사실을 알 수 있었다. 따라서 본 문형 매칭에 의한 번역 방식은 실제 슬롯의 개수를 줄이지 않는 한 이러한 적용성 문제는 결국 나타날 수 밖에 없다. 즉 10 단어 이상의 장문의 경우 완전 매칭 과 부분 매칭을 합쳐도 5%에도 미치지 못하고 있다.

실제 뉴스에서 사용하는 문장들이 10 단어 이상의 중문 및 복문인 경우가 73% 정도임을 감안할 경우 일반 문틀의 문형 매칭 방법의 적용성 문제로 인하여 실현성이 없는 것으로 판단된다. 물론 자동 문형 추출 과정에서의 오류가 고려되어야 하겠지만 전체 문형의 완전한 매칭에 의한 번역 방식은 현실적으로 한계가 있을 것으로 판단되며 단문의 적용성 충분히 가능하다는 사실이 입증되었기 때문에 이러한 단문 중심의 부분 문형의 조합으로 전체 문형을 매칭하는 방식이 신중히 검토되어야 하겠다.

<표 4> 용언 중심 문틀 길이별 매칭 결과

슬롯 수	빈도	완전매칭	부분매칭	완전+부분
1	1207	1207	0	1207(24.16%)
2	856	846	1	847(16.96%)
3	938	856	37	893(17.88%)
4	791	585	143	728(14.57%)
5	572	267	217	484(9.69%)
6	322	57	202	259(5.19%)
7	151	3	105	108(2.16%)
8	100	1	66	67(1.34%)
9	33	0	20	20(0.4%)
10	13	0	7	7(0.14%)
11	9	0	4	4(0.08%)
12	0	0	0	0
13 이상	3	0	0	0
총계	4995	3822	802	4624
매칭율		76.5%	16.1%	92.6%

표 4는 용언 중심의 문틀 길이별 매칭 결과를 나타낸다. 완전 매칭된 문장의 개수가 총 3,822(76.5%) 문장이며 부분문틀에 의한 매칭은 802(16.1) 문장이다. 따라서 전체 92.6%의 적용성을 보임으로써 패턴의 적용성은 상당히 높다고 볼 수 있다. 그러나 이러한 패턴에서 각 슬롯의 의미 및 어휘 등에 대한 정보를 상세히 기술할 수 없어 비록 패턴이 매칭이 된다고 해도 이에 대한 구조적인 모호성을 해결할 수 없다는 치명적인 단점을 안고 있다.

5. 결론

어절을 단위로 하는 문틀 매칭 방법은 실험에 의하여 그 매칭률이 현저히 떨어짐을 알 수 있었고 용언 중심의 문틀 방식은 그 매칭률은 높은 반면 실제 적용시 여러 후보문틀 중 입력문틀과 가장 유사한 문틀을 제시할 수 있지만 원시문틀의 수가 많아지는 경우 결국 구조 분석의 모호성은 피할 수 없다. 따라서 한영 번역 방식은 한국어 구문 구조 모호성이라는 특성상 규칙 기반 번역 방식을 중심으로 어휘상호간의 정보인 MI, 단문 중심의 대역 격틀 패턴, 의미 정보 등의 번역 데이터를 이용하여 규칙에 의하여 분석되는 다수의 구조들 중 최적의 구조를 선택해 나가는 방식이 타당하다.

따라서 한국어를 원시언어로 사용하는 한영 번역 방식에서 사용 가능한 번역 패턴은 문장 전체를 표현하는 문틀보다는 동사를 중심으로 하는 의미 정보 및 대역어 정보가 포함된 단문 단위의 대역 패턴이라고 판단된다. 앞으로의 과제는 이러한 동사구 중심의 대역 패턴에 대한 체계적인 지식 구축과 이를 어떻게 효율적으로 구문구조 애매성 해소에 적용하는가 하는 문제이다.

참고문헌

[1] 시스템공학연구소, 영한 한영 텍스트 자동 번역 기술 개발, 연구 보고서, 1997.

[2] 여상화 외 5 인, "에서로/KE: 한영 기계번역 시스템, 정보과학회 추계 학술대회 논문집, 1997.

[3] 서울대학교, 한영 기계번역을 위한 한국어 구문 분석과 변환에 관한 연구, 연구보고서, 1996.

[4] 서정연, 조정미, 김길창, "한=>영 대화체 기계번역 시스템," 제 11 회 음성 통신 및 신호처리 워크샵 논문집 제 11 권 1 호, pp.65-70, 1994

[5] 이영직 외 9 인, "한-일 호텔예약 음성번역시스템," 제 12 회 음성 통신 및 신호처리 워크샵 논문집 제 12 권 1 호, pp.204-207, 1995

[6] 김종혁, 한-영 기계 번역에서의 예문을 이용한 변환, 석사학위논문, 한국과학기술원, 1994.

[7] 김형근, 확률적 의존 문법과 한국어 구문 분석, 석사학위논문, 한국과학기술원, 1994.