한영 기계번역 시스템을 위한 효율적인 한국어 용언 처리

Efficient Korean Predicates Processing for Korean-English Machine Translation System

박홍원, 정경진, 네기시 켄이치로, 임유정
(주)두레소프트 부설 휴먼인터페이스 연구소
walts123@hitel.net, samagu96@hanmail.net, kennegishi@hotmail.com, tadami@hanmail.net

Hong-won Park, Kyung-Jin Jung, Kenichiro Negishi, Yoo-Jung Lim
DooReSoft Corp. Human Interface Laboratory

[논문요약]

한영 기계번역 시스템을 구현하기 위해서는 다양하게 활용하는 한국어 용언을 보다 효율적으로 처리해야 할 필요가 있다. 한국어 용언은 그 활용이 매우 다양하여 활용에 따라 문장 내에서 다양하게 기능하게 된다. 한영 기계번역 시스템에서는 용언의 활용이 가지는 여러 정보를 효율적으로 분석하여 해당 정보를 보다 효과적으로 역문에 반영시키는 연구가 필요하다.

본 논문에서는 용언의 활용에 따른 여러 정보-시제에 관한 정보(선어말어미 관련), 문종에 관한 정보(어말어미 관련), 양상에 관한 정보(보조용언, 어말어미 관련) 등-를 통일된 코드를 이용하여 일괄적으로 처리하는 방법론과 그 과정을 제시한다.

1. 서론

1-1. 연구의 필요성

한국어 문장에서의 용언[1]은 어미활용이 있으며 다양한 형태론적인 교착이 일어나고 보조용언이 사용된다는 점에서 형태적인 정보뿐만 아니라 통사적인 정보를 포함하고 있다.

한영 기계번역 시스템을 구현하기 위해서는 용언 활용에 따른 여러 정보를 역문의 생성에 효과적으로 반영하여 영어 문장의 특성을 충분히 살려주는 처리가 요구된다.

그러나 한국어의 용언은 어미의 활용에 따른 정보뿐이 아니라 자립성이 희박한 보조용언 등의 결합에 의해 양상이 변하는 경우가 있기 때문에 한국어의 용언 처리는 그 양(量)이 막대하다고 할 수 있다.

본 논문에서는 한영 기계번역 시스템을 구현하기 위해서 한국어의 다양한 용언 활용에 나타나는 정보를 미리 준비된 특정의 코드를 사용하여 일괄적으로 처리함으로써 보다 효과적으로 한국어 용언을 처리하는 방법을 제시한다.

1-2. 연구의 범위

한국어 용언 처리를 위해서는 우선 용언의 어간, 보조용언, 어미의 구분을 위한 정확한 형태소 분석이 이루어져야 한다. 한국어의 특성상 어간에 후접하는 보조용언, 선어말어미, 어말어미 등에 문장에 대한 통사적인 정보가 대부분 포함되어 있기 때문에 용언이 가지고 있는 의미를 정확하게 분석하여 일괄적으로 처리할 수 있는 방법이 요구된다. 그러나 형태소 분석기는 본 연구의 직접적인 논의 대상이 아니므로 본 논문에서 구체적인 언급을 하지 않는다.

본 논문에서는 한영 기계번역 시스템의 구현을

1) 여기서 '용언'은 '서술어(predicate)'와 동일한 개념으로 사용되었다

위해 한국어 용언이 가지고 있는 정보를 문법적인 분석을 통해 추출하고, 이를 코드화하여 해당 정보를 참조함으로써 한국어와 상이한 통사 구조를 가지고 있는 영어 문장이 문법에 맞게 올바르게 생성될 수 있도록 하였다.

이를 위해 활용된 용언이 가지고 있는 다양한 정보를 인식과정에서 코드화하게 된다. 본 연구에서는 이때 코드화의 대상이 되는 양상정보, 화행정보, 문종정보, 시제정보, 높임정보 등의 정보들에 대한 코드화 방법론과 절차 등에 대해 집중적으로 다룬다.

1-3. 이론적 배경

본 연구에서는 어미의 활용 부분을 제외한 본용언의 어간을 해당 용언의 표제어로 인식하거나, 종결어미, 연결어미, 전성어미, 선어말어미 등을 기능적인 특성에 따라 통일된 번호로 코드화하는 등의 대부분의 절차를 정통문법의 문법범주 구분에 기준하여 처리하고 있다.

	예문(먹을 필요가 있다)의 처리
정통문법	먹(어간)+을(관형사형전성어미)+필요(명사)+가(조사)+있(어간)+다(종결어미)
본 연구	먹(어간)+을 필요가 있(다어절어)+다(종결어미)

<표1> 보조용언 처리의 예

한편, 정통문법과 본 연구에서의 방법론의 상이점을 비교하여 표로 나타내보면 <표2>와 같다.

	보조용언의 문법범주 및 문법적 특성
정통문법	문법적인 뜻을 나타내며 자립성이 희박한 용언으로 2개이상이 연결되어 사용될 수 있다
본 연구	용언의 활용에 의해 양상, 화행을 표현하는 모든 다어절어 범주로 그 처리 범위를 확대 적용

<표2> 정통문법과의 문법범주 구분상의 상이점

본 연구에서는 한국어 보조용언의 처리에서 정통문법과 상이점을 보이는데 이를 '다어절어'라는 비교적 큰 범위의 문법 범주로 확대 적용하여 취급하고 있다.

기존의 보조용언 설정의 근거와 범주에 대한 다양한 시각이 있다[1][2][3][4]. 그러나, 본 연구에서는 정통문법의 다양한 시각에 관계없이 보조용언을 '본용언의 어간에 후접하고 어미활용을 동반하는 용언'으로 정의하고, 이와 같은 조건에 부합되는 용언은 모두 보조용언의 부류로 취급하여 '다어절어'라는 새로운 개념의 용언을 제안한다.

정통문법에서는 양태(modality), 상(aspect), 화행(speech act)을 각각 구분하고 있는 경우가 많고, 한영 기계번역 시스템 연구에서도 양상, 문종, 화법 등을 각각 구분하여 처리하는 것이 일반적이다[1][4][5].

그러나 본 연구에서는 양태, 상, 화행(화법) 등을 구분하지 않고 '양상'으로 통일시켜 그 개념을 사용하고 있으며, 문법범주가 정통문법과 다소 차이가 있다고 하더라도 시스템의 효율과 효과적인 역문 생성을 위해 이들을 모두 동일한 개념으로 간주한다.

2. 한국어 용언 처리의 방법론

2-1. 코드화를 통한 용언 처리

하나의 용언(보조용언이 결부되지 않은 형태일 경우)은 [어간]+[어말어미]의 형태 또는 [어간]+[선어말어미]+[어말어미]의 형태로 용언의 활용을 이루고 있다. 용언이 활용을 할 때, 그 형태가 변하지 않는 어간 부분을 제외하면 어미에 의한 용언의 변화는 매우 다양하다. 어간의 종류와 그에 결합하는 어미의 종류에 따라 활용법 또한 다양하다.

예를 들어 '먹'이라는 어간과 '가'라는 어간에 '-었(과거시제)'라는 선어말어미와 '다'라는 종결어미가 결합할 경우, 각각 '먹었다'와 '갔다'로 어미가 다르게 활용한다. 용언 활용을 처리하기 위해서는 이러한 활용의 차이에 따라 그룹을 만들어 처리할 필요가 있다. 어미를 어간에 자유롭게 결합시키기

위해 어간에 해당 그룹의 활용 정보를 주는 코드화의 첫단계라고 할 수 있다.

해당 그룹의 활용 정보를 가진 어간은 다양한 어미와 결합함으로써 시제정보, 높임정보, 양상정보, 화행정보, 문장의 종류 등 여러 정보를 포함하게 된다. 한국어의 어미는 동일한 정보를 가지고 있는 어미가 단 하나만 존재하는 것이 아니라 문맥에 따라 사용가능한 어미의 수가 매우 다양하므로 어미의 종류에 따라 코드화시켜서 일괄적으로 처리할 필요가 있다.

과거, 현재, 미래 등의 시제정보를 나타내는 시제정보 코드, 높임을 나타내는 높임정보 코드(한영 기계번역에서는 필요없기 때문에 문종을 나타내는 해당 어미와 동일 처리가 가능), 평서형, 의문형, 명령형, 감탄형, 청유형 등을 나타내는 문종 코드를 해당 어미에 부여하여 다수의 어미를 간단한 코드로써 정보를 갖게 하는 것이다.

어미와 어미 사이에 실질적인 뜻을 나타내지 못하는 보조용언이 결합하여 문장의 양상을 결정짓는 경우도 있는데, 세분화된 코드를 사용하여 다양한 양상을 표현할 수 있게 한다.

개인의 상황, 시대적 상황, 매스컴의 영향 등에 의해 한국어 어미의 수는 지속적으로 늘어나고 있기 때문에 기계번역에 있어서 막대한 양의 어미를 정해진 규칙에 의해 코드화시켜 일괄적으로 커리하는 것은 매우 효율적인 방법이라고 생각된다.

2-2. 보조용언 분석을 통한 정보 추출

보조용언은 [용언의 어간]+[보조적 연결어미]+[보조용언]+[어말어미]의 구조에서 그 역할을 수행하며 문장 내에서 실질적인 뜻을 나타내지는 못하지만, 문법적으로 보조적인 뜻을 지니고 있어서 문장을 완성시키는데 큰 영향을 미친다.

본 연구에서는 1-3에서 보조용언을 정통문법과는 달리 "양상과 화행을 표현하는 모든 범위의 다어절어"로 그 문법적 특징을 정의하였다. 따라서 [용언의 어간]+[보조적 연결어미]+[다어절어]+[어말어미]의 구조를 가지게 된다.

"다어절어"가 결합한 경우, 어간 이하를 모두 어미로 처리하여 정보를 주게 되면 양상과 어미를 동시에 가지게 되어 어미사전은 무한대로 늘어나게 된다. 이런 비효율성을 줄이기 위해 "다어절어"와 "어말어미"를 나누어 데이터베이스를 구축한다.

어간에 결합된 다어절어의 수(數)	다어절어의 예
1개	-수 있-
2개	-수 있는 척을 하-
3개	-수 있는 척을 하는 것 같-
4개	.....

<표1> 다어절어의 예

한편, "다어절어"는 2개 이상의 양상을 포함하고 있을 가능성이 있으므로 다양한 양상코드가 필요하다. '-수 있'과 같이 양상이 1개인 경우도 있을 수 있지만, '-수 있는 척을 하'와 같이 양상이 2개인 경우, 혹은 '-수 있는 척을 하는 것 같'과 같이 3개 이상인 경우도 있을 수 있다.

따라서, 본 연구에서는 1개 이상 또는 그 이상의 특정 양상에 특정 양상코드를 부여하는 방식이 아니라, 최소로 필요한 양상에 특정 양상코드를 부여하여 그 이상의 양상을 필요로 할 때, 조합하는 방식을 사용하였다.

'-수 있'에 해당하는 양상코드, '-척을 하'에 해당하는 양상코드, '-것 같'에 해당하는 양상코드를 각각 필요한 시점에 부여하여 '-수 있는 척을 하는 것 같'이라는 양상이 필요할 경우, 3개의 양상을 어미활용에 의해 연결시키고 각각의 양상이 가지고 있는 값을 조합하여 최종적인 양상의 집합구조를 만들어 내게 된다.

부가코드	다어절어
A	-수 있-
B	-척을 하-
C	-것 같-

<표3> 다어절어 사전

조합코드	조합된 다어절어	영어생성
A+B	-수 있는 척을 하-	pretend to be able to -
A+C	-수 있는 것 같-	seem to be able to -
B+C	-척을 하는 것 같-	seem to pretend to -
A+B+C	-수 있는 척을 하는 것 같-	seem to pretend to be able to -
....	....	....

<표3> 다어절어 생성 사전

2-3. 선어말어미 분석을 통한 정보 추출

용언의 어간에 어말어미나 선어말어미가 결합하여 활용의 형태에 따라 '첨가현상' 혹은 '탈락현상'이 일어나는 경우에 대해 각각의 어미 그룹을 만들어 코드화하였다.

시제와 높임을 나타내는 선어말어미는 어미코드의 의해 일괄적으로 처리되는데, 과거, 현재, 미래에 해당하는 시제에 대한 어미코드를 부여하여 시제를 나타내는 선어말어미와의 결합이 가능한 모든 어미에 시제정보, 높임정보를 부여하게 된다.

'먹는 사람'과 '먹은 사람'을 분석해 보면, 전자는 '먹(어간)'+'는(현재형 관형사형 어미)', 후자는 '먹(어간)+은(과거형 관형사형 어미)'로 시제표현이 현재와 과거로 다르게 나타나는데, 시제정보와 수식형 어미에 대한 정보도 동시에 코드화하여 해당 용언에 대한 정보를 부여한다.

2-4. 어말어미 분석을 통한 정보 추출

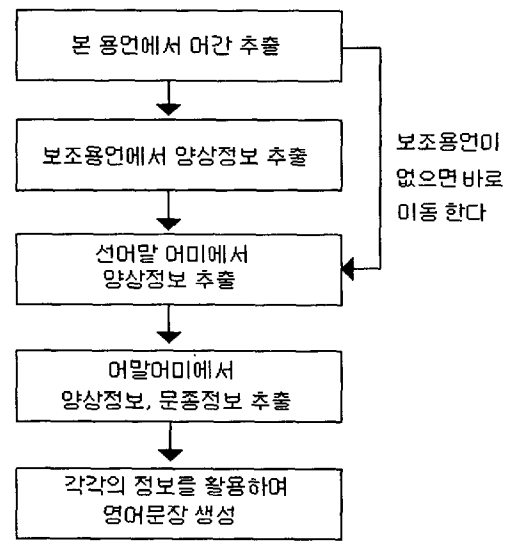
종결어미의 종류는 평서형, 감탄형, 의문형, 명령형, 청유형 등이 있는데, 본 연구에서는 평서형과 의문형을 통사적인 구조에 영향을 미치는 문종으로 가정하여 처리하고, 명령형과 청유형은 양상으로 처리하여 보조용언이 필요로 하는 데이터베이스를 이용하여 양상으로 처리하였다. 감탄형은 평서형의 범위에 포함시키기로 한다.

평서형이나 의문형 등의 종결어미만을 문종으로 취급하지 않고, 연결어미나 전성어미도 문종으로 취급하여 어미의 역할에 따라 어미코드를 부여한다. 구나 절에 동일한 정보를 주어야 하는 어미의 그룹을 코드에 의해 일괄적으로 처리할 수 있다.

예를 들어, '먹는 사람'은 '먹(어간)'+'는(관형사형 어미)'이라는 분석을 할 수 있는데, 이러한 '-는'이라는 관형사형 어미에 특정 코드를 부여함으로써 관형사형 어미를 가지는 수식문장을 간단한 코드에 의해 효율적으로 처리할 수 있게 된다.

3. 한국어 용언 처리의 과정

<그림1>에서 보는 바와 같이 형태소 분석을 통해 용언에서 어간을 추출하고 활용번호를 부여받는다. 보조용언(다어절어)이 결합된 문장이면, 다어절어 데이터베이스에서 그 다어절어가 가지고 있는 양상정보를 추출한다.

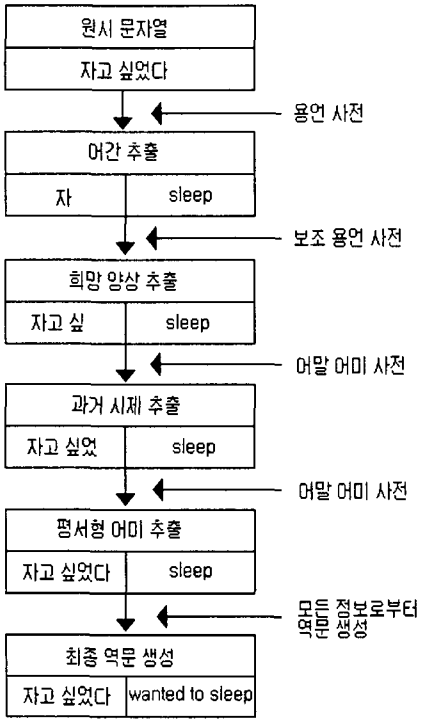


<그림1> 용언 처리 과정

보조용언이 없으면 어미사전으로 이동한다. 시제나 높임을 나타내는 선어말어미가 포함되어 있으면 그에 대한 정보를 추출하고, 어말어미와 결합하는 시점에서 선어말어미와 어말어미의 문종정보, 양상정보를 코드화하여 최종적인 영어 생성 문장을 얻게 된다.

4. 한국어 용언 처리의 예

예문1) 나는 자고 싶었다.

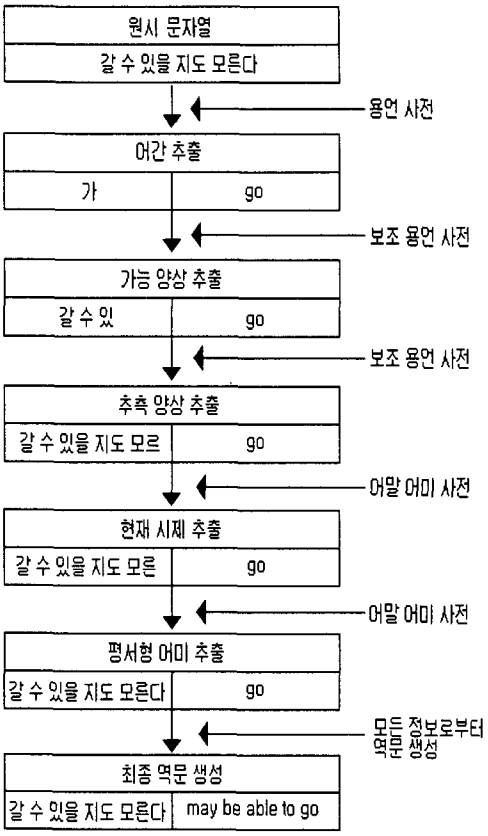


<그림2> 용언 처리 과정의 예 1

형태소 분석을 하여 '자(sleep)'를 용언사전에서 찾는다. 보조용언사전에서 '-고 싶'을 찾아 희망양상을 부여받고, 어말어미를 결합시키는 과정에서 선어말어미의 과거 시제와 평서형 어말어미 코드를 인식한다. 역문 생성을 위한 다어절어 조합 사전에서 생성에 필요한 정보를 받아 최종 역문을 생성한다.

'-기를 원하', '-고 싶어 하'등의 보조용언도 '-고 싶'과 같은 희망양상을 가지므로 다양한 표현을 코드에 의해 일괄적으로 처리할 수 있다. '-어요', '-어', '-ㅂ니다'등의 같은 문종에 대한 정보를 포함하고 있는 어미도 하나의 코드로 일괄 처리하므로 매우 효율적이다.

예문2) 학교에 갈 수 있을 지도 모른다.



<그림3> 용언 처리 과정의 예 2

형태소 분석을 하여 '가(go)'를 용언사전에서 찾는다. 다어절어 사전에서 '-수 있'을 찾아 가능양상을 부여받고 다시 활용하여 '-지도 모르'를 찾아 추측양상을 부여받는다. 어미사전에서 현재 시제와 평서형 어미 정보를 가진다. 역문 생성을 위한 다어절어 조합 사전에서 조합된 두 개의 양상에 생성 정보를 받아 최종 역문을 생성한다.

5. 결론

한영 기계번역 시스템에서 한국어의 용언을 효과적으로 처리하기 위하여 한국어 용언이 포함하

고 있는 다양한 정보들을 분석하여 의미에 맞는 영어 문장을 생성해 줄 수 있는 방법론의 개발은 매우 중요한 과제이다.

본 연구에서는 한국어 용언의 다양한 활용을 영어 문장의 생성에 반영하기 위해 통일된 코드를 사용하여 용언이 가지고 있는 정보를 일괄적으로 처리하는 방법을 사용하였다.

용언의 어간에 후접하는 보조용언, 선어말어미, 어말어미 등의 정보를 생성언어의 특성에 맞게 분석하여야 하므로 문법범주의 처리방법에 있어서 정통문법의 그것과 약간의 상이점이 있었다.

보조용언을 '다어절어'라는 개념으로 확대 적용하여 본용언에 부과되는 다양한 표현들을 양상코드를 부가하여 일괄적으로 처리하였다. 이때 양상은 정통문법과는 달리 양태(modality), 상(aspect), 화행(speech act) 등을 양상이라는 하나의 단일화된 개념으로 취급하였다.

본 연구에서 다루고 있는 코드화를 통한 용언처리 방법론은 한영 번역뿐만 아니라, 한중·한일 번역 시스템에서도 다양하게 활용될 수 있기 때문에, 다국어 번역 시스템을 구현하는 유용한 방법론으로 사용될 수 있을 것으로 본다.

[참고문헌]

[1]손세모돌, "국어 보조용언 연구", 한국문화사, 1996.

[2] 박덕유, "國語의 動詞相 硏究", 한국문화사, 1998.

[3] 이을환, 이길주, "韓國語文法論", 개문사, 1985.

[4] 김지은, "우리말 양태용언 구문 연구", 한국문화사, 1998.

[5] 이현호, "한영 사전에서 얻어진 패턴을 이용한 한영변환", 전북대학교 대학원 석사학위논문, 2001.