韓國 컴퓨터情報學會 論文誌
第10卷 第4號, 2005. 9.

2005-10-4-1-20

한-일 기계번역에서 '하다'용언의 번역 방법

문 경 희*

Translation Method of '-hada' verb in a Korean-to-Japanese Machine Translation

Kyong-hi Moon*

요 약

한국어와 일본어는 문법 체계의 유사성으로 인하여, 양언어의 형태소들 간에 1대1 매핑 만으로도 높은 번역 성능을 얻을 수 있다. 따라서, 대부분의 한-일 기계번역에서는 한국어와 일본어 형태소 사이에 1대1 매핑을 기본으로 하고 있다. 명사와 '하다'로 구성되는 한국어 '하다'용언도 대부분 명사와 'する'로 구성되는 일본어 'する'용언에 대응되므로, 일반적으로 1대1 매핑을 관계를 적용한다. 그러나, 한국어 '하다'용언이 일본어 'する'용언에 대응되지 않는 경우, 1대1 매핑 만으로는 정확한 번역 결과를 얻지 못하는 경우도 자주 발생하게 된다. 이 경우 명사와 '하다'를 하나의 번역 단위로 다루어 주는 것이 필요하다. 따라서, 본 논문에서는 한국어 '하다'용언의 특성을 조사하고, 명사와 '하다' 사이에 삽입된 어휘들에 의한 비연속성 문제, 피동화, 관형어 수식 등 입력 문장에서의 다양한 상황에 따른 '하다'용언의 변환 기법을 제안하였다. 실험 결과, 높은 번역 성능을 보임으로써 제안한 방법이 한-일 기계번역에서 '하다'용언을 다루는데 효율적임을 볼 수 있었다.

Abstract

Due to grammatical similarities, even a one-to-one mapping between Korean and Japanese morphemes can usually result in a high quality Korean-to-Japanese machine translation. So most of Korean-to-Japanese machine translation are based on a one-to-one mapping relation. Most of Korean '-hada' verbs, which consist of a noun and '-hada', also correspond to Japanese '-suru' verbs, which consist of a noun and '-suru', so we generally use one-to-one mapping relation between them. However, the applications only by one-to-one mapping may sometimes result in incorrect Japanese correspondences in some cases that Korean 'hada' verbs don't correspond to Japanese 'suru'verbs. In these cases, we need to handle a noun and '-hada' as one translation unit. Therefore, this paper examined the characteristics of Korean '-hada' verb and proposed transfer rules of Korean 'hada' verb, applying for various states of input sentences such as discontinuity due to inserted words between a noun and '-hada', passivization, and modification of '-hada' verb. In an experimental evaluation, the proposed method was very effective for handling '-hada' verb in a Korean-to-Japanese machine translation, showing high quality of translation results.

▸ Keyword : '하다'용언('hada' verb), 한-일 기계번역(Korean-to-Japanese Machine Translation)

• 제1저자 : 문경희
• 접수일 : 2005.07.29, 심사완료일 : 2005.08.22
* 신라대학교 컴퓨터정보공학부

Ⅰ. 서 론

한국어와 일본어는 문법 체계가 거의 유사하고 아무런 정보처리 없이 양언어의 단어들을 1대1로 대응시킨 경우에도 약 65% 정도의 번역 성능을 얻을 수 있으므로[1], 대부분의 한 일 기계번역 시스템은 직접번역 방식에 의해 구축되고 있다[2, 3, 4]. 직접번역 방식은 구문분석이나 의미분석 없이, 원시언어를 형태소 분석하고 이를 직접 목적언어로 변환하여 생성하는 방식이며, 이는 한국어와 일본어 형태소 사이에 1대1 매핑을 기본으로 하고 있다.

한국어에서 빈번히 사용되는 '하다'용언(명사 + '하다')의 경우도 일본어 'する'용언(명사 + 'する')에 해당되어 일반적으로1대1 매핑만으로 올바른 대역어를 얻을 수 있지만, 일본어에서 하나의 형태소에 해당된다든지 구조가 전혀 다른 대역어에 해당하는 경우도 자주 발생한다. 따라서, 한국어와 일본어 대역어 간에 1대1 매핑이 불가능하거나, 또는 1대1 매핑이 가능은 하더라도 "대답(-을) 하다"의 일본어 대역어로 "答える"와 "答え(-を) する" 모두 가능하지만 하나의 대역어로 구성된 "答える"가 훨씬 선호되는 것처럼 좀 더 자연스러운 번역 표현이 있는 경우에 대한 처리가 필요하다.

또한, '하다'용언의 경우 명사와 '하다'가 조사나 다른 어절들의 삽입으로 비연속적으로 나타나는 경우가 빈번히 발생하므로, 이 경우 명사와 '하다'를 하나의 '하다'용언으로 인식하여 번역해 주는 것이 필요하다. '하다'용언의 비연속성에 의해 최악의 경우, 하나의 '하다'용언 인식을 위하여 입력문장 전체를 살펴 보아야 할 수도 있다. 이는 많은 인식시간을 요구하며 또한, 불필요한 부분과의 비교로 인하여 인식오류가 발생할 수도 있다. 따라서, 입력문장의 상황과 '하다'용언의 특성에 맞게 최대한 인식범위를 축소하여 인식하는 방법이 요구된다.

그러나, 기존 한 일 기계번역 시스템에서는 이러한 '하다'용언의 특성과 입력문장의 상황에 따른 번역 기법을 제시하지 않고 있으며, 실제로 기존 상용 시스템들에서 "거짓말하다"와 "거짓말을 하다"와 같이 '하다'용언이 직접 인접하여 나타난 경우와 기본격을 가지고 나타난 경우가 다르게 번역되는 경우도 있으며, 이 두 경우는 같이 번역되더라도 "거짓말을 그가 하다"와 같이 약간의 변형만 일어나도 번역 결과가 달라지는 경우가 발생하고 있다.

따라서, 본 논문에서는 명사와 '하다'가 모두 나타났을 때, 1대1 매핑관계에 의해 일본어 명사 대역어+'する'로 번역되어야 하는 일반적인 번역의 경우와 전혀 다른 형태의 일본어 단위로 번역되는 통합 의미적 번역의 경우를 구분하여 번역하기 위해, 명사와 '하다' 사이에 삽입된 어휘와 명사를 수식하는 관형어에 의한 번역 규칙, 그리고 피동화된 경우의 번역 규칙을 설정함으로써, 입력 문장에서의 상황에 따라 일반적인 번역과 통합 의미적 번역을 수행할 수 있도록 제안하였다. 또한, 번역 후, 명사 + '하다'가 하나의 통합 의미적 단위로 번역되는 경우 나타나는 어순 재조정 문제를 다루었다.

본 논문의 구성은 다음과 같다. 2장에서 '하다'용언의 특징 및 처리 필요성을 살펴보고, 3장에서는 '하다'용언 처리에 대한 기존 연구를 살펴보고, 4장에서는 한 일 기계번역에서 '하다'용언의 특성 및 변형에 따른 변환 규칙들을 설정하고, 5장에서 그에 따른 변환 알고리즘, 6장에서 삽입어휘에 대한 어순 조정 방법을 다루며, 7장에서는 실제 한 일 기계번역 시스템에 본 연구의 방법을 적용하여 그 유용성을 검증하고, 8장에서 결론을 맺는다.

Ⅱ. '하다' 용언의 특징 및 처리 필요성

한국어 '하다'용언은 명사와 '하다'로 구성되며, 명사 뒤에 조사가 삽입 또는 삭제된 형태로 나타난다. '하다'용언의 선행요소인 동작성 명사는 의미적으로 볼 때, (그림 1)과 같이 분류할 수 있다[5].

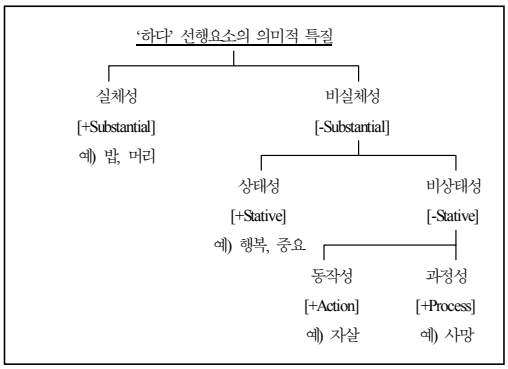

그림 1. 동작성 명사의 의미적 특질
Fig 1. Semantics of Verbal Nouns

한국어 '하다'용언은 대부분 일본어 'する'용언에 대응된다. 그러나, 일본어 'する'용언은 선행요소인 명사가 거의 다 비실체성이며, 비실체성 중에서도 비상태성만을 선행요소로 갖는다. 따라서, 한국어 '하다'용언 중 실체성 명사를 선행요소로 갖는 것은 대부분 일본어 'する'용언에 대응되지 못하고 다른 형태의 대역어를 가지게 된다. 또한, '하다'용언 중 상태성 명사를 선행요소로 갖는 것은 형용사적 쓰임을 보이는 것으로서, "정확하다"가 "正確だ"로 번역되는 것과 같이 대부분 '하다'의 대역어로서 '-だ/な/に/…'등의 활용을 하는 일본어의 형용동사에 대응되며, "피곤하다"가 "疲れている"로 번역되는 것과 같이 전혀 다른 형태로 대응되는 경우도 종종 있다. 그 외에도, 동작성, 과정성의 선행명사를 가지는 '하다'용언도 일본어 'する'용언에 대응되지 않는 경우가 있으며, 한국어 '하다'용언이 일본어 'する'용언에 대응되는 경우도 삽입조사의 격이 달라져서 1대1 매핑으로는 번역이 불가능한 경우가 있다. 따라서, 이러한 경우들의 '하다'용언은 통합 의미적 번역이 이루어야 한다. (그림 2)은 '하다'용언에 대한 일본어 대역어의 가능한 형태들을 일부 보여준다[5].

(ⅰ) 명사(조사) + 'する'
　한국어: 여행 (-을) 하다
　일본어: 旅行 (-を) する
(ⅱ) 명사 조사 + 'する'
　한국어: 사형 (-을) 하다
　일본어: 死刑 -に する
(ⅲ) 용언
　한국어: 시작 (-을) 하다
　일본어: 始める
(ⅳ) 명사(조사) + 용언
　한국어: 살인 (-을) 하다
　일본어: 殺人 (-を) 犯す

그림 2. '하다'용언의 일본어 대역어 구조
Fig 2. Japanese Correspondences of 'hada' verb

첫번째 타입의 '하다'용언은 일본어 'する'용언에 대응된다. 이 경우는 한국어에서 명사를 뒤따르는 조사가 생략되면 대응되는 일본어에서도 생략될 수 있다. 따라서, 첫번째 타입의 '하다'용언은 1대1 매핑에 의해서 정확히 번역될 수 있다.

그러나, 두번째 타입의 '하다'용언은 한국어에서 조사가 생략되더라도 일본어 대역어에서는 특별한 조사가 반드시 나타나는 경우이며, 세번째 타입은 하나의 일본어 용언에 대응되는 경우이다. 네번째 경우는 한국어 '하다'가 일본어 'する'에 대응되지 않고 다른 용언으로 대응되는 경우로서, 이때, 한국어와 일본어의 구문구조는 같지만, 한국어 '하다'가 매우 많은 의미를 가지므로 그에 적당한 일본어 용언을 찾기 힘들다. 따라서, (ⅱ), (ⅲ), (ⅳ) 타입의 '하다'용언은 1대1 매핑에 의해 정확히 번역될 수 없으므로 통합 의미적 번역이 이루어져야 한다.

한국어 '하다'용언에 대응되는 일본어 대역어가 [명사(조사)+'する'] 형태도 가능하고 하나의 용언 형태도 가능할 때, 일본어에서는 하나의 용언 형태를 선호하는 경향이 있다. <표 1>은 한국어 '하다'용언 "대답(-을) 하다"에 대응될 수 있는 일본어 대역어들과, 그들이 약 13,000여 문장의 일본 신문기사에서 출현하는 빈도를 나타낸 것이다. "答えする" 와 "返事する"는 (그림 2)의 첫번째 타입에 해당하는 대역어로서 1대1 매핑에 의해 번역될 수 있다. 그러나, 이들은 극히 적은 빈도로 출현하며, 대부분 1대1 매핑에 의해서는 제대로 번역될 수 없는 표현인 "答える"의 형태로 나타났음을 알 수 있다. 따라서, 보다 자연스러운 번역 결과를 위해서는 "대답(-을) 하다"가 "答える"로 대응되도록 해야 하므로 이는 통합 의미적 번역으로 처리되어야 한다.

표 1. "대답(-을) 하다"의 일본어 대역어 후보
Table 1. Japanese Candidates of "daidap(-ul) ha(-da)"

일본어 대역어	출현빈도 (%)
答える	105 (95.45%)
答え(を) する	2 (1.82%)
返事(を) する	3 (2.73%)
합계	110 (100%)

'하다'용언은 구성형태소인 명사의 바로 뒤에 조사가 자유롭게 삽입 또는 삭제 가능하며, 여러 어절들이 구성형태소인 명사와 '하다' 사이에 자유롭게 삽입될 수 있다. 따라서, 한국어와 일본어간에 1대1 매핑으로 번역이 불가능한 경우에, 비연속적으로 나타나는 명사와 '하다'를 하나의 '하다'용언으로 인식하여 통합 의미적 번역이 이루어질 수 있도록 하여주는 방법이 필요하며, 삽입된 어휘들에 대한 처리도 필요하다.

III. 기존 연구

한국어와 일본어의 문법적 유사성으로 인하여, 한-일 기계번역 시스템은 다른 언어들 간의 기계번역 시스템보다 훨씬 많은 시스템들이 개발되어 있으며, 특히, 약간의 지역적 어순 불일치 외에는 어순이 거의 일치하여[6], 어느 정도의 처리만으로도 높은 번역 성능을 보여주고 있다. 이러한 점에 착안하여 대부분의 한-일 기계번역 시스템이 한국어와 일본어의 형태소 간 1대1 매핑관계를 적용한 직접번역 방식을 따르고 있다. 그러나, (그림 2)에서 설명한 (ii), (iii), (iv) 타입의 '하다'용언과 <표 1>에서 설명한 '하다'용언은 1대1 매핑관계에 의해서 부자연스럽거나 잘못된 번역 결과를 내는 경우에 해당되며, 보다 고품질의 한-일 기계번역 시스템을 위해서는 1대1 매핑이 이루어지지 않는 이러한 문제들에 대한 처리가 필요하다. 그러나, 기존의 시스템들은 '하다'용언의 문제를 세부적으로 나누어 다루고 있지 않다. 또한, 1대1 매핑 관계가 이루어지지 않는 경우들을 인식하더라도, 입력문장에서 명사와 '하다'가 비연속적으로 나타난다든지, 관형형 수식을 받는다든지 하는 문장 변화에 적절히 대응을 하고 있지 못하다. 다음 <표 2>는 '하다'용언의 여러 가지 특성들을 보여주고 있는 (그림 2)와 <표 1>의 예들을 웹 상에서 제공되고 있는 기존 한-일 기계번역 시스템들에 적용시켜 본 결과이다.

기존 시스템들에서는 2장에서 설명한 '하다'용언의 여러 가지 일본어 번역 형태를 고려하지 않고, <표 2>의 (1), (3)에서 보는 바와 같이 '하다'를 'する'로 단순히 번역시킴으로써 문제를 쉽게 다루어 버리고 있다. <표 2>의 (2), (4)에서 보는 것처럼 명사와 '하다'가 연속하여 나타나는 경우는 일본어 단일어 대역이 가능하도록 하기도 하지만, 그 사이에 '을/를'과 같은 조사가 삽입된다든지, 다른 어휘가 삽입되는 경우는 역시 (그림 2)나 <표 1>에서 설명한 '하다'용언의 여러 가지 대역 형태 처리를 무시하고 모두 'する'로 번역되어짐을 볼 수 있다. 따라서, 본 논문에서는 보다 고품질의 한-일 번역 시스템을 위하여 '하다'용언의 특성에 맞는 일본어 대역어를 선정해 줄 수 있는 번역 방법을 제시하고자 한다.

표 2. 기존 한-일 기계번역 시스템에서의 번역 결과
Table 2. Translation Results in Existing Korean-to-Japanese Machine Translation Systems

한국어		일본어	
		A시스템	B시스템
(1)사형하다	사형하다	死刑する	死刑する
	사형을 하다	死刑をする	死刑をする
(2)시작하다	시작하다	始める.	始める.
	시작을 하다	手始めをする	開始を する
(3)살인하다	살인하다	人殺しをする	人を殺す
	살인을 하다	殺人をする	殺人を する
(4)대답하다	대답하다	答える	答える
	대답을 하다	返事をする	返事 (返答) を する

IV. '하다'용언의 특성 및 변형에 따른 변환 규칙

한국어 입력문장에서 명사와 '하다'가 모두 나타났더라도 그들 간의 의존 관계가 성립하여 '하다'용언으로서의 의미를 가지는 경우와 그렇지 못한 경우가 있다. 또한, 한 일 기계번역에서 다루어져야 하는 '하다'용언의 경우 1대1 매핑을 기본으로 하지만 자연스러운 번역 형태인 통합 의미적 번역이 따로 존재하는 경우 그것을 따르는 것이 일반적이다. 그러나, 입력문장에서의 변형, 즉, 삽입어휘나 관형어 수식, 피동화 등에 따라 일반적인 번역 방법을 따라야 하는 경우도 발생한다. 따라서, '하다'용언의 변환규칙 설정은 문장에서 명사와 '하다'가 모두 나타났을 때, '하다'용언으로 인식하여 다룰 것인가 결정하는 것과, 통합 의미적 번역과 일반적인 번역 중 보다 올바른 번역 방법을 선택하는 것을 목적으로 한다. '하다'용언으로 인식되었을 때, 그 '하다'용언을 위한 통합 의미적 번역이 존재하는 경우는 그것을 우선으로

하며, 관형어 수식 등 특수한 경우 일반적 번역을 선택해서 변환해야 하는 경우도 발생한다.

'하다'용언의 변환은 명사와 하다를 따로 분리함으로써 생기는 불필요한 의미 중의성 해소 과정을 피하기 위하여 품사태깅이 일어난 바로 다음에 수행한다.

4.1. 삽입 조사에 의한 변환 규칙

명사 뒤에 후접하는 조사는 그 명사의 격을 결정한다. 그 명사의 기본격을 결정하는 조사를 기본격조사라 하고, 이는 앞 명사의 유/무/ㄹ종성형과 격에 따라 하나의 표층형태로 결정되므로 '-가'(주격 무종성)나 '-이'(주격 유/ㄹ종성), '-을'(목적격 유/ㄹ종성), '-를'(목적격 무종성) 등과 같이 직접 표층정보로 기술한다.

'하다'용언은 주로 목적격조사 '-을/를'을 기본격조사로 가지며, 이는 "대답을 하다" 처럼 명사 뒤에 후접하여 나타날 수도 또는 "대답하다"처럼 명사 뒤에 조사가 생략될 수도 있다[7]. 이 경우 모두 '하다'용언으로 인식되어 변환되어야 한다.

또한, 격조사와는 달리 일정한 격을 표시하는데 사용되지 않고 특별한 의미를 더해주는 역할을 하며 여러 격에 통용될 수 있는 특성을 나타내는 보조사류(-는, -도, -만, -마다, -부터, -까지 등)는 이러한 기본격조사들의 일부를 대행하여 나타날 수 있으며, 목적격 조사 '-을/를'은 대부분의 보조사에 의해 대행이 가능하다[8, 9]. 따라서, '하다'용언의 경우 기본격 조사 '-을/를'이 나타난 경우, 생략된 경우, 그리고 보조사들이 삽입된 경우도 '하다'용언으로 인식하여 변환을 수행한다.

또한, "행복하다"와 같이 선행명사로 상태성 명사를 갖는 경우는 명사와 '하다'가 합쳐져서 하나의 형용사 역할을 하는 것으로 기본격조사를 갖지는 않으나 "행복(-은/도/조차) 하다" 등과 같이 보조사의 삽입이 허용될 수 있다[10]. 따라서, 이 경우는 보조사가 삽입된 경우도 '하다'용언으로 다루도록 한다.

4.2 삽입 어절에 의한 인식 규칙

하다'용언에서 구성하는 두 구성형태소 사이에는 의존관계가 성립한다. 즉, '하다'용언의 구성형태소 중 '하다'는 명사의 지배소 역할을 하며, 따라서, 입력문장에서 그들의 구성형태소인 명사와 '하다'가 모두 나타났더라도, 그 구성형태소간의 의존관계가 유지된 경우만을 '하다'용언으로 다룬다. 이러한 의존관계는 구성형태소들이 바로 연속하여 나타난 경우는 거의 유지되지만, 그 사이에 다른 어절이 삽입된 경우는 그렇지 못한 경우가 종종 발생한다. 구문분석이 정확히 수행된다면 이러한 의존 관계를 밝힐 수 있으나, 앞서 기술한 바와 같이 한 일 기계번역 시스템에서는 한국어와 일본어의 문법적 유사성으로 인하여 직접번역방식을 사용하고 있으며, 따라서, 구문분석을 수행하지 않고 있다. 또한, 아직까지 구문분석 기술이 만족할만한 수준에 이르지 못하였고, 한 일 기계번역에서 구문분석 정보가 필요한 경우는 아주 적으므로 구문분석을 수행하는 것이 오히려 역효과가 날 수 있다. 따라서, 이러한 의존관계가 유지되지 못한 경우를 조사하기 위하여 삽입어절에 대해 다음과 같은 휴리스틱을 적용한다.

명사와 '하다'가 나타났더라도 그 사이에 다른 서술형명사가 삽입된 경우는 '하다'용언으로 다루지 않는다. 이 경우 삽입된 서술형 명사에 의해 구성형태소인 명사와 '하다'가 의존관계를 갖지 않을 확률이 높기 때문이다. (그림 3)에서, "애국(-을) 하(-다)"의 구성형태소들 사이에 서술형 명사인 '운동'이 삽입되었으며, '하다'는 '애국'이 아닌 '운동'의 지배소 역할을 하게 되므로 '애국'과 '하다'는 무관하게 된다. 또한, 명사와 '하다' 사이에 다른 용언이 나타나는 경우도 '하다'용언으로서의 의미가 없으므로 이 경우 '하다'용언으로 인식하지 않는다. 이 조건은 '하다'용언을 찾는 범위를 제한하여 시간을 절약하고, 잘못된 인식이 이루어지지 않도록 하는 규칙이 될 수 있다.

애국(-을) 하(-다)
Ex) 애국 운동 -을 하다.
시작(-을) 하(-다)
Ex) 시작 은 보 -려고 하 -였지만,

그림 3. 삽입어절에 의한 변환규칙 적용 예
Fig 3. Examples Applying the Transfer Rules by Inserted Words

4.3 관형어에 의한 인식규칙

'하다'용언의 앞에서 '하다'용언의 구성형태소 중 체언을 수식하는 관형어는 피수식어로 체언을 반드시 필요로 하는 의존형식이다. 그러나, "대답(-을) 하(-다)"가 "答え(-る)"로 대응되는 것처럼 한국어 '하다'용언에서는 체언이 나타났더라도 일본어 대역어에는 대응하는 체언이 용언과 결합하여 하나의 용언 형태로 나타나서, 관형어에 의한 수식을 받는 경우 일본어에서는 피수식어가 생성되지 않으므로 비문법적인 일본어 번역문을 생성하는 경우가 생긴다. "대답(-을) 하(-다)"는 자연스러운 번역을 위하여 통합 의미적 번역이 이

루어져야 함을 5.1.2에서 설명한 바 있다. 그러나, (그림 4)에서 처럼 그 구성형태소 '대답'이 "정확한"에 의해 수식되는 경우 "대답(-을) 하(-다)"가 통합 의미적 번역인 "答え(-る)" 로 변환된다면, "정확한"에 대응되는 "正確な"가 수식할 명사가 없어지므로 비문법적인 일본어 번역 결과가 생성된다.

```
한국어 : 정확하 -ㄴ 대답 -을  하 -ㄹ 수
         가 없 -었 다.
일본어 : 正確 -な  答え -られ な -い.(X)
       /* ungrammatical expression*/
```

그림 4. 관형어 수식에 의한 비문 생성 예
Fig 4. Examples of Ungrammatical Expressions by Modificaton

관형어가 체언을 수식하는 방법은 다음 4가지이다[9].

❶ 관형사가 그대로 관형어가 되는 경우
❷ 체언에 관형격 조사 '-의'가 결합된 경우
❸ 용언 어간에 관형사형 전성어미가 결합된 경우 (관형사형 전성어미: -ㄴ, -던, -은, -는, -ㄹ, -을, -라는, -으라는)
❹ 관형격 조사 '-의'가 생략되어 체언+체언의 구성으로 된 경우

이들 중, 형용사어간에 관형사형 현재형어미 '-ㄴ/은'이 붙는 경우는 관형사형 현재형어미를 부사형어미로 바꾸어주기만 하면 일본어 대역어에 체언이 없을지라도 대부분 자연스러운 번역이 이루어질 수 있다. 그러나, 그 외의 경우는 일반적인 문장 재구성 규칙을 도출하기 어려우며 기계번역에서 그때그때 달라지는 이러한 문장 재구성을 수행하는 것은 극히 어려운 문제이다. 따라서, 통합 의미적 번역으로 비문법적인 문장을 생성하기 보다는, 통합 의미적 번역으로 다루지 않고 부자연스러운 번역일지라도 일반적인 번역으로 다루는 것이 낫다.

관형어 수식에 의한 인식 규칙을 정리하면 다음과 같다.

- 규칙 1: '형용사어간+관형사형 현재형어미'의 형태로 관형어 수식을 받고 일본어 대역어에 체언성분이 없는 경우는 통합 의미적 번역을 수행하되 관형사형 현재형어미를 부사형 전성어미로 대치한다.
- 규칙 2: 관형어가 '형용사어간+관형사형 현재형어미'가 아니고 일본어 대역어에 체언성분이 없는 경우는 통합 의미적 번역 대신 일반적 의미로 번역한다.

(그림 5)은 규칙 1과 규칙 2가 적용된 예이다.

```
(ⅰ) 규칙 1 적용 예 : 통합의미적 번역
한국어 : 정확하 -ㄴ 대답-을  하 -ㄹ 수 -가
         없 -었 -다.
일본어 : 正確 -に  答え -られ -な-かった

(ⅱ) 규칙2 적용 예 : 1대1 매핑에 의한 번역
한국어 : 그 -가 바라 -는 대답 -을 하 -ㄹ
         수 -가 없 -었 -다.
일본어 : 彼-が 望 -む 答え -が でき -な
         -かった.
```

그림5. 관형어 수식에 의한 변환규칙 적용 예
Fig 5. Examples Applying the Transfer Rules by Modification

4.4 피동화에 의한 변환 규칙

'하다'용언에 피동화가 일어나는 경우 격전이가 발생할 수 있으므로 기본격인 목적격 '-을/를'대신 '-이/가'가 나타난 경우도 '하다'용언으로 인식하여 변환하여 준다. 이때 한국어에서 '하다'용언 뒤에 따르는 피동의 양상을 일본어의 양상에 대응시켜 줌으로써 정확한 번역이 이루어질 수 있다.

Ⅴ. '하다'용언의 변환 알고리즘

이 절에서는 앞서 기술된 변환 규칙들에 따라 '하다'용언의 변환 과정을 보인다. '하다'용언의 변환은 한국어 형태소 분석과 품사태깅 후에 수행된다. 따라서, '하다'용언의 변환을 위해 형태소분석 및 품사태깅이 완료된 하나의 결과가 입력문장으로 주어진다.

먼저 입력문장에서 형태소 분석, 품사태깅 과정을 통해 '하다'용언을 얻을 수 있다. (그림 6)는 '하다'용언(대답 + '하다')의 변환 과정을 보여준다.

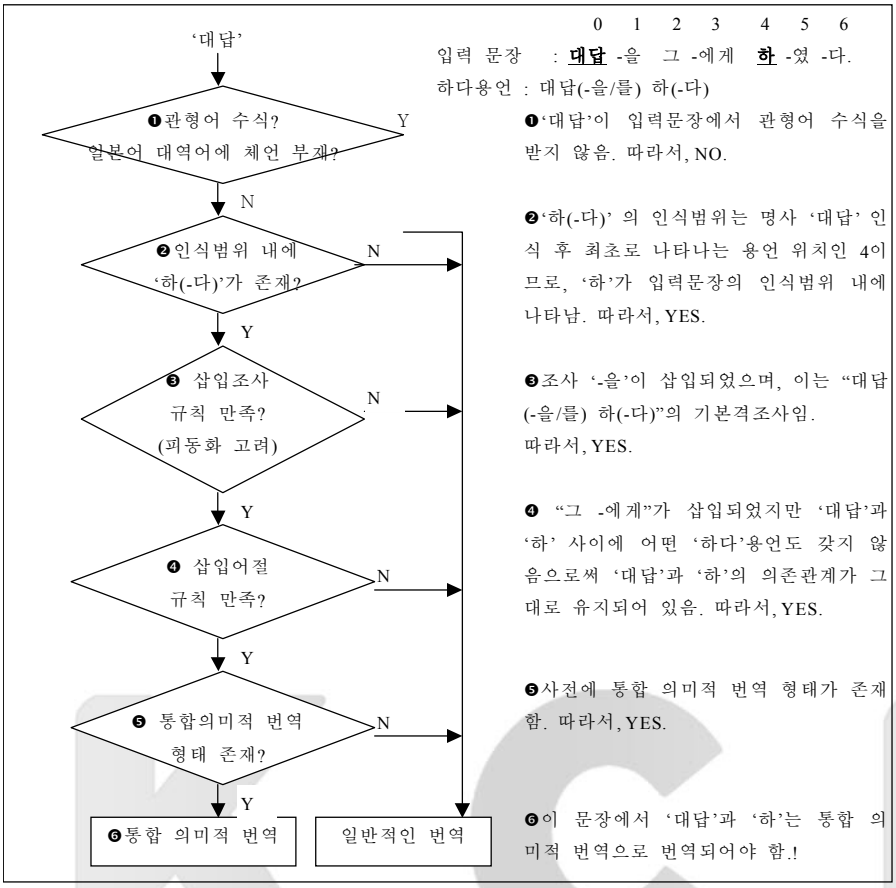

그림 6. '하다' 용언 변환 알고리즘
Fig 6. Transfer Algorithm of 'hada' verb

(그림 6)에서 ❶은 4.3에서 기술된 관형어에 의한 인식 규칙을 검사하기 위한 모듈이다. 즉, 일본어 대역어에서 체언 성분이 나타나지 않고 하나의 용언으로 대응되는 경우 통합 의미적 번역으로 변환이 이루어져야 하나, 한국어 '하다'용언에서 체언성분이 관형어에 의해 수식을 받으면, 통합 의미적 번역으로 변환하지 않고 1대1 매핑에 의한 일반적인 번역이 이루어진다. 단, 관형어가 '형용사어간+관형사형 현재형어미'의 구조이면 통합 의미적 번역을 따르되 관형사형 현재형 어미를 부사형전성어미로 대치한다.

❷에서는 '하다'용언의 특성에 따라, 명사가 나타난 후 '하(-다)'가 나타날 수 있는 유효 범위를 검사한다. 중간에 다른 용언이 나타나는 경우 '하다'용언으로서의 의미를 부여하지 않는다.

❸은 4.1에서 기술된 삽입조사에 의한 인식규칙을 검사하는 모듈이다. 구성형태소인 명사에 직접 인접하여 삽입된 조사가 있는 경우, 그 조사가 기본격 조사나 허용 가능한 보조사인 경우만 '하다'용언으로 인식한다. 또한, '하다'용언이 피동형으로 나타난 경우 기본격 조사 검사에서 '-을/를' 대신 '-이/가'를 선택한다.

❹는 4.2에서 기술된 삽입 어절에 의한 인식 규칙을 검사하는 모듈이다. 구성형태소들의 의존관계가 유지되어 있는지를 조사하기 위하여 4.2에서 기술한 휴리스틱을 검사한다. 즉, 명사와 '하다' 사이에 다른 서술형 명사가 삽입되면 '하다'용언으로 다루지 않으므로 일반적인 번역이 이루어진다.

❺는 위 단계들에서 '하다'용언으로 인식되었을 때, 통합 의미적 번역 형태가 사전에 존재하는지 검사하여 존재하면 통합의미적 번역으로, 그렇지 않으면 일반적인 번역을 수행하도록 결정하는 최종 단계이다.

이렇게 '하다'용언의 특성과 입력 문장에서의 나타난 형태에 따라 통합 의미적 번역과 일반적인 번역으로 구분하여 번역하는 방법을 채택함으로써 보다 자연스러운 번역을 수행할 수 있다.

Ⅵ. 삽입 어휘들의 어순 조정

앞서 기술된 변환 방법에 의해 '하다'용언을 변환할 때, 명사와 '하다'가 비연속적인 '하다'용언이 하나의 일본어 형태소로 변환되는 경우, 삽입 어휘들의 어순 조정이 불가피하다. 따라서 다음과 같은 방법으로 삽입 어휘들의 어순을 재배치한 후 일본어로 변환한다.

'하다'용언의 명사에 직접 인접하여 삽입된 조사는 무시하고, 명사와 '하다' 사이에 삽입된 대부분의 어절들은 명사의 앞으로 재배치하여 해당 일본어로 변환한다. (그림 7)의 '-을'이 전자의 경우이며, "그렇게"가 후자의 예가 된다. 또한, "대답을 하다/대답을 안하다/대답을 못하다"처럼 부정소인 '안'과 '못'이 술어성분 앞에 삽입될 수 있으며, 한국어에서는 부정소인 '안'과 '못'이 술어의 앞에 위치할 수 있지만 일본어에서의 부정표현은 'ない'나 'ず'등 술부 양상류로만 표현할 수 있으므로 이는 술어의 뒤로 위치하여 양상류로 대치되어야 한다. '안'은 순수 부정표현을 나타내는 반면, '못'은 부정과 가능의 표현이 합쳐진 불가능의 의미를 내는 경우가 많다. 따라서, 술어 앞에 삽입된 부정소 '안'이나 '못'은 술어 '하다'의 바로 뒤로 이동시키고 일본어 변환과정에서 안'은 부정의 양상정보로, '못'은 불가능의 양상정보로 대치한다.

(그림 7)은 '하다'용언의 사이에 삽입된 어휘들에 대한 어순 조정 예를 보여준다.

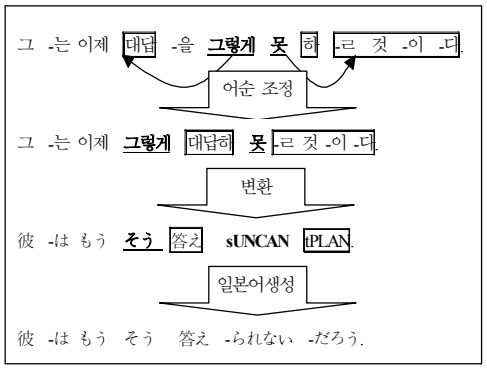

그림 7. 삽입어휘의 어순 조정 예
Fig 7. Example of Word Order Adjustment

Ⅶ. 실험 및 평가

제안한 한 일 기계번역에서의 '하다'용언의 처리 방법의 효율성을 증명하기 위하여, 한 일 기계번역 시스템(COBALT K/J)[3, 11]에 적용하여 번역률을 평가하여 보았다. '하다'용언 변환 모듈은 앞서 기술한 바와 같이 형태소분석, 품사태깅이 완료된 후, 하나의 품사태깅된 결과에 대해 적용하였으며, 실험 말뭉치로서는 약 천만어절 말뭉치인 국어정보베이스에서 임의로 2,808문장을 추출하였다. 추출된 말뭉치는 각 문장 당 평균 10여개의 어절로 구성되어 있다.

제안한 방법을 평가하기 위하여 <표 3>에서와 같이 1대1 매핑에 의한 번역 결과와 연속된 '하다'용언만을 인식하여 일반적인 번역과 통합 의미적 번역으로 구분하여 번역한 결과, 그리고 제안한 방법에 의해 비연속적이거나 입력문장에서 변형되어 나타나는 경우도 인식하여 일반적인 번역과 통합 의미적 번역으로 구분하여 번역한 결과를 비교하였다.

한국어와 일본어의 구조적 차이에 의해 통합 의미적 번역이 이루어져야 하는 '하다'용언은 모두 922개가 나타났으며, 1대1 매핑에 의한 일반적인 번역을 수행한 경우는 '하다'를 모두 'する'로 번역해 버림으로써 모두 잘못된 번역이나 부자연스러운 번역 결과를 내었다. 연속적으로 명사와 '하(-다)'가 나타난 경우만 인식하여 '하다'용언을 일반적인 번역과 통합 의미적 번역으로 구분하여 번역한 경우 약 97.3 % 정도의 번역 정확성을 보였으며, 제안한 방법에 의하여 비연속적으로 나타나는 '하다'용언을 인식하고 문장에서 나타나는 변형에 따라 일반적인 번역과 통합 의미적 번역으로 구분하여 번역한 경우 약 99.6%의 번역 정확성을 보임으로써 제안한 방법이 '하다'용언의 자연스러운 번역에 효과적임을 알 수 있었다.

표 3. 실험 결과 및 번역 예
Table 3. Experimental Results and Transfer Examples

실험 방법	번역률	번역 예
1대1 매핑에 의한 처리	0%	거짓말하다 ➔ 嘘する 거짓말을 하다 ➔ 嘘をする 거짓말을 그가 하다 ➔ 嘘を彼がする
연속된 '하다'용언만 처리	97.3%	거짓말하다 ➔ 嘘つく 거짓말을 하다 ➔ 嘘をする 거짓말을 그가 하다 ➔ 嘘を彼がする
제안한 방법에 의한 처리	99.6%	거짓말하다 ➔ 嘘つく 거짓말을 하다 ➔ 嘘をつく 거짓말을 그가 하다 ➔ 嘘を彼がつく

Ⅷ. 결론

한국어에서 빈번히 사용되는 '하다'용언의 경우 일반적으로 일본어 'する'용언에 해당되어 1대1 매핑으로 올바른 대역어를 얻을 수 있지만 일본어에서 하나의 형태소에 해당된 다든지 구조가 전혀 다른 대역어에 해당하는 경우도 자주 발생한다. 따라서, 한국어와 일본어 대역어 간에 1대1 매핑이 불가능하거나, 또는 1대1 매핑이 가능은 하더라도 좀 더 자연스러운 번역 표현이 있는 경우에는 통합 의미적 번역이 필요하다. 그러나, '하다'용언의 경우 명사와 '하다'가 조사나 다른 어절들의 삽입으로 비연속적으로 나타나는 경우가 빈번히 발생하므로, 이 경우 명사와 '하다'를 하나의 '하다'용언으로 인식하여 번역해 주는 것이 필요하다. 이를 위하여 명사와 '하다' 사이에 삽입된 어휘와 명사를 수식하는 관형어에 의한 번역 규칙, 그리고 피동화된 경우의 번역 규칙을 설정함으로써, 입력 문장에서의 상황에 따라 일반적인 번역과 통합 의미적 번역을 수행할 수 있도록 제안하였다. 실험 결과, 제안한 방법에 의해 거의 대부분의 '하다'용언이 자연스러운 일본어로 번역됨을 볼 수 있었다.

참고문헌

[1] 김정인, 문경희, 이종혁, 이근배, 일 한 기계번역에 있어서 한국어 술부의 생성과 평가, 한글 및 한국어 정보처리 학술 발표 논문집, pp.329 337, 1996.

[2] 허남원, 일 한 기계번역에서 접속어미의 애매성 해소를 위한 대조 연구 : 일본어 접속조사 「~te」를 중심으로, 포항공과대학교 박사학위 논문, 2000.

[3] 음성 번역 시스템을 위한 번역기술 연구, 한국통신 정보통신기초 연구보고서, 1999.

[4] 김선호, 한 일 기계번역에서 다어절 변환 단위의 인식 및 변환을 위한 사전적 접근 방법, 포항공과대학교 석사학위논문, 1997.

[5] 요꼬오 사에꼬, 한국어의 '하다'용언과 일본어 용언의 비교 연구, 한양대학교 석사학위 논문, 1987.

[6] 김연숙, 김창완, 퍼지 이론을 이용한 한국어 및 일어 화자 인식에 관한 연구, 한국컴퓨터정보학회 논문지, 제 5권 3호, pp. 51-57, 2000.

[7] 송재관, 박찬곤, 기계번역용 한국어 품사에 관한 연구, 한국컴퓨터정보학회 논문지, 제 5권 4호, pp. 48-54, 2000.

[8] 홍사만, 국어특수조사론 - 의미분석, 학문사, 1987.

[9] 이관규, 학교 문법론, 월인 출판사, 1999.

[10] 남기심, 국어 문법의 탐구 III - 국어 통사론의 문제와 전망, 태학사, 1996.

[11] Moon Kyeonghee, Jong Hyeok Lee, Korean to Japanese MT System, COBALT K/J, 18th International Conference of Computer Processing of Oriental Languages (ICCPOL1999), pp.2, 1999.

저 자 소 개

문 경 희

2002년 2월 포항공과대학교 컴퓨터 공학과 공학박사

2004~현재 신라대학교 컴퓨터정보 공학부 교수

<관심분야> 기계번역, 정보처리, 시맨틱웹