

## 패턴 및 예문에 기반한 영한 변환

이 기 영 김 한 우  
한양대학교 전자계산학과

### English-Korean Transfer Based on Patterns and Examples

Ki Young Lee, Han Woo Kim  
Dept. of Computer Science & Engineering, HanYang University  
E-mail : kylee@cse.hanyang.ac.kr

#### Abstract

Conventional rule-based approaches have some problems caused by rule maintenance. Also they have some limitations to get the high quality translation results. This paper presents new English-Korean transfer approach that uses patterns and examples on limited domains. The use of patterns and examples can resolve the ambiguities and give high quality of MT. Proposed approach can be applied in various NLP related area. Experimental results with a test corpus are discussed.

#### 1. 서론

전통적인 규칙 기반의 기계번역 방법은 많은 수의 규칙을 사용하여 분석, 변환, 생성의 세 단계 번역 과정을 통해 번역을 수행한다. 이러한 규칙 기반의 기계번역 방법은 너무 많은 규칙의 사용으로 인해 해석에 있어서 많은 모호성을 생성할 뿐만 아니라, 이를 유지 및 관리하는데 많은 문제점을 드러내고 있다. 또한 규칙 기반의 기계번역 방법은 시스템의 도메인(domain)을 변경 또는 확장할 경우, 새로 도입되는 규칙과 기존의 규칙간의 충돌을 세심하게 신경 써야 할 뿐만 아니라, 전체 시스템을 개발하기까지 걸리는 시간과 노력이 엄청나다는 단점을 가지고 있다. 이러한 이유로 인해 전통적인 규칙 기반 기계번역 방법과는 다른 다양한 기계번역 방법론들이 제시되었으며, 이러한 방법론들로서 대표적인 것으로는 예제 기반 기계번역, 패턴 기반 기계번역, 통계 기반 기계번역 및 코퍼스 기

반 기계번역 등이 있다.

본 논문에서는 영한 병렬 코퍼스를 변환 지식으로 사용하여 코퍼스로부터 추출된 패턴과 코퍼스의 예문을 함께 사용하는 새로운 영한 번역 방법을 제안한다.

본 논문에서 제안된 방식은 현재 진행 중인 패턴 추출 과정의 자동화가 완료될 경우, 번역 도메인을 변경하거나, 전체 시스템을 개발하는데 걸리는 시간을 상당히 줄일 수 있다는 장점을 지닌다. 또한 제안된 방법은 예문을 사용함으로써 기준의 규칙 기반 기계번역 방법에 비해 보다 자연스러운 번역 품질을 얻을 수 있으며, 패턴을 사용함으로써 원시 언어 해석에서 생기는 모호성을 상당히 줄일 수 있을 뿐만 아니라, 기준의 예제 기반 기계번역 방법이 가지고 있는 문제점인 표증 표현의 비교로 인한 유사 문장 발견의 어려움을 해결할 수 있다.

#### 2. 관련 연구

기계번역에서 예문을 사용하려는 생각은 [1]에서 처음으로 제안되었다. [1]에서 제안된 방법은 원문과 번역문의 쌍(pair)으로 구성된 대규모 병렬 코퍼스를 사용하여 입력문과 유사한 예문을 찾아서 이를 이용하여 번역문을 생성한다. 이와 같은 번역 방법은 일치하는 예문이 존재할 경우에는 비교적 자연스러운 번역이 가능하다는 장점을 지닌다. 그러나 입력문의 길이가 길어지면 문장 단위의 비교에 의해서는 유사한 예문을 발견할 확률이 매우 떨어진다. 따라서 이러한 문제점을 코퍼스의 크기로 해결하기 위해서는 매우 방대한 양의 예문을 필요로 한다는 단점을 지니고 있다.

[2]에서는 코퍼스의 예문과 입력문과의 비교를 문장

단위가 아닌 문장의 일부, 즉, 청크(chunk)로 나누는 방법을 제안하였다. [2]에서는 문법은 별로 고려하지 않고 주로 표층 단어의 일치 여부에만 의존하기 때문에, 청크를 잘못 나누었을 경우에는 번역의 품질에 악영향을 끼치게 되는 단점이 있다.

[3]에서는 예문 데이터베이스를 의존 문법에 기반하여 원문과 번역문을 구문 분석한 결과인 의존트리로 구축하며, 각 노드 간의 대응관계를 따로 나타내었다. [3]에서는 하나의 입력문을 번역하는데 있어서 하나 이상의 예문을 사용한다. 이때, 번역 단위의 분할은 기준의 문법에 의해서 구현된다. 즉, 기준의 문법을 사용하여 문장을 구 단위로 분할한 뒤, 각각의 구를 서로 다른 예문을 사용하여 번역하고, 최종적으로 하나의 문장으로 통합해서 번역문을 생성한다. 이 방법의 문제점은 우선 데이터베이스를 구축하는데, 너무 많은 비용이 든다는 것이고, 기준의 규칙에 기반하여 원시 언어 및 목적 언어를 해석하므로 그 모호성도 무시할 수 없다는데 있다.

예문에 기반한 번역 방법과 함께 새로운 번역 방법으로서 제안된 것이 패턴에 기반한 방법이다. 패턴은 번역에 사용하려는 시도는 [4], [5], [6], [7] 등에서 제안되었다. [4], [5], [6]에서는 특정 도메인 상에서 자주 사용되는 빈출 표현들을 패턴으로 구축하였으며, 메타 부분에 올 수 있는 언어의 리스트를 따로 만들어 두고 적당한 번역을 시도하였다. [7]에서는 패턴 기반의 문맥자유문법을 설정하여 이를 기계번역에 사용하였다. [8]에서는 번역 패턴을 추출하는 방법을 제안하였다.

하지만 이러한 방법들도 역시 기준의 규칙 기반의 해석을 통해 패턴을 추출하므로 해석에서 발생하는 모호성 문제는 피하기 어렵다고 할 수 있다.

위에서 언급한 연구들이 공통으로 다루고 있는 것은 기준의 규칙 기반의 기계번역 시스템은 많은 양의 규칙을 사용하고 있으므로 그 유지 및 관리가 힘들며, 확장 또한 매우 어렵다는 것이다. 또한 기준의 규칙 기반의 기계번역 시스템들은 시스템 개발에 걸리는 시간이 기타 소프트웨어에 비해 특히 길어서, 상용 제품으로의 개발에 많은 문제점을 지니고 있다는 것이다. 이러한 점은 Sato에 의해 [9]에서도 규칙 기반 기계번역 시스템의 문제점으로 제기되었다.

### 3. 패턴

본 논문에서는 문장 패턴 및 구 패턴이라는 두 가지 패턴을 사용하여 병렬 코퍼스의 문장을 표현한다.

#### 3.1 문장 패턴

문장 패턴은 해당 문장에서 문법적·의미적 중심 부분을 의미한다. 일반적으로 문장의 핵심 부분은 용언에 해당하며, 격문법 또는 결합가 문법 등은 용언과 명사와의 의미 관계를 중심으로 문장을 해석하였다. 하지만 기준의 규칙에만 의존해서 번역을 수행할 경

우, 이미 서론에서 언급한 많은 문제점과 직면하게 된다. 따라서 본 논문에서는 병렬 코퍼스를 기반으로 해당 도메인에서 빈번하게 발생하는 고정된 표현 뿐만 아니라 자연스러운 번역을 위해서 하나의 번역 단위로 보아야 할 표현들을 문장 패턴으로 간주하였다.

(문장) The inner type is identified by means of its identifier.

(대역) 내부 유형은 식별자에 의해 명시된다.

(원문패턴) x1 be identified by means of x2

(대역패턴) x1은 x2에 의해 명시된다

(문장) The ASN.1 notation is referenced by other standards.

(대역) ASN.1 표기법은 다른 표준에 의해 참조된다.

(원문패턴) x1 be referenced by x2

(대역패턴) x1는 x2에 의해 참조된다

위의 예는 병렬 코퍼스의 영어 및 한국어 문장과 그에 해당하는 패턴 및 대역 패턴을 나타낸다. (원문패턴)의 패턴 표현에서 사용된 x1, x2, x3 등은 패턴의 메타 부분이며, 이 메타 부분에는 다양한 어구들이 적용될 수 있다. 이렇게 정의된 문장의 패턴은 제한된 도메인에서는 더욱 뚜렷하게 드러난다.

#### 3.2 구 패턴

구 패턴은 문장에 문장 패턴을 적용했을 때 메타 부분에 해당하는 구의 패턴을 말한다.

(구) a number of simple types

(대역) 다수의 단순 유형

(구패턴) a number of x1

(대역패턴) 다수의 x1

(구) the tables in this annex

(대역) 이 부기의 표

(구패턴) x1 in x2

(대역패턴) x2의 x1

위의 예는 구 패턴의 예를 나타낸다. 이렇게 메타 부분을 다시 패턴화 하는 이유는 메타 부분을 하나의 번역 단위로 봄으로써 예문을 찾을 경우, 웬만한 코퍼스 크기로는 동일 예문이 발견될 확률이 낮기 때문이다.

패턴을 통해서 원시 언어를 해석할 경우, 문법적, 의미적 모호성을 매우 효과적으로 줄일 수 있다. 즉, 문장의 패턴이 우선적으로 적용될 경우, 전역 파싱에 의해 발생할 수 있는 모호성을 상당히 줄일 수 있으며, 패턴 적용 이후의 과정인 변환 과정에서 메타 부분을 어느 정도 독립적으로 병렬 코퍼스의 예문을 사용하여 자연스럽게 변환할 수 있다. 또한 메타 부분의 사용은 유사한 예문을 발견할 확률을 높여준다.

#### 4. 패턴과 예문을 사용한 영한 변환

아래의 그림1은 본 논문에서 제안하는 변환 기법의 전체적인 흐름을 나타낸다.

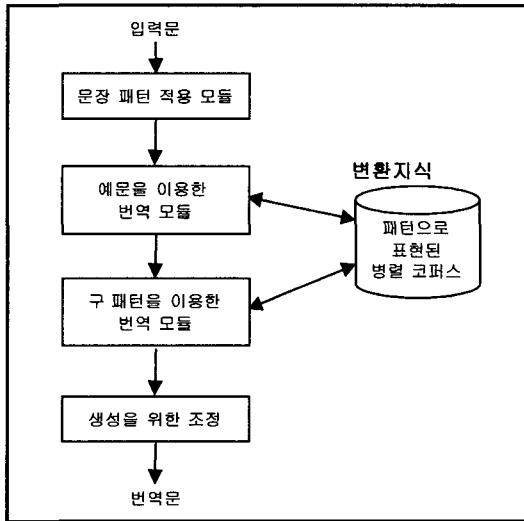


그림 1 영한 변환 순서도

Fig. 1 English-Korean transfer flowchart

본 절에서는 그림1에 나타나는 번역 과정을 예문을 통해서 상세히 설명한다.

##### 4.1 문장 패턴의 추출 및 검색

번역 대상인 영어 문장이 입력문으로 들어오면 시스템은 우선 입력문을 커버하는 패턴을 추출한다.

(1) The order of combining characters is defined.

예를 들어, 문장 (1)이 입력문으로 들어오면 시스템은 문장을 커버하는 문장 패턴을 병렬 코퍼스에서 찾는다. 시스템은 문장 패턴이 'x1 be defined'라는 것을 발견하고, 이에 대한 한국어 번역문은 대역 패턴에 의해 'x1은 정의된다'로 결정된다. 그리고 x1에는 'the order of combining characters'가 할당된다.

##### 4.2 예문을 사용한 메타 부분의 변환

계속해서, x1에 할당된 'the order of combining characters'를 번역하기 위해서 우선 x1 전체를 모두 포함하고 있는 예문이 존재하는지 병렬 코퍼스를 검색한다.

(DB1) Following represents the order of combining characters.

(다음은 합성 문자의 순서를 나타낸다.)

예를 들어, (DB1)과 같은 문장이 병렬 코퍼스에 존재한다면 x1의 번역을 일괄적으로 수행할 수 있다. 즉, x1의 대역문은 '합성 문자의 순서'로 결정된다. 그리고 최종적으로 조정 과정을 거친 뒤에, '합성 문자의 순서가 정의된다'가 문장 (1)의 번역문으로 생성된다.

#### 4.3 구 패턴을 사용한 메타 부분의 변환

그러나 만약, (DB1)과 같은 문장이 존재하지 않는다 고 가정하자.

(DB2) SaudiName uses a subset of combining characters.  
(SaudiName은 합성문자의 하부 세트를 사용한다.)

(DB3) A canonical order for tags is defined in 6.4.  
(태그에 대한 정규 순서는 6.4에서 정의된다.)

대신에, (DB2) 및 (DB3)와 같은 예문이 병렬 코퍼스에 존재한다고 가정하자.

'the order of combining characters'라는 x1에 할당된 어구 전체를 포함하는 예문이 존재하지 않으므로, 보다 작은 번역 단위로 나누기 위해 구 패턴 'x3 of x4'를 적용한다. 구 패턴의 적용에 의해 x3에는 'the order'가 할당되고, x4에는 'combining character'가 할당된다. 계속해서, x3과 x4를 포함하고 있는 예문으로 (DB1)과 (DB2)가 발견되고, 예문들을 통해서 x3과 x4를 각각 '순서'와 '합성 문자'로 옮바르게 번역할 수 있다. 결국 'the order of combining character'는 '합성 문자의 순서'로 번역되며, 마지막으로 최종적인 조정 과정을 거쳐서 입력문 (1)에 대한 올바른 해석인 문장 (2)을 얻게 된다.

(2) 합성 문자의 순서가 정의된다.

#### 4.4 생성을 위한 조정

본 논문에서 제안하는 방법의 최종 단계인 조정 단계에서는 하나 이상의 번역 단위로 나뉘어서 번역된 대역 부분들을 적당하게 결합하는 처리를 수행한다. 또한 조정 단계에서는 문장 패턴에 명시되어 있지 않은 모달리티나 시제 등의 처리를 수행한다.

### 5. 실험

#### 5.1 학습 코퍼스 및 테스트 코퍼스

실험을 위해 사용된 학습 코퍼스는 제한된 도메인으로서 'ITU-T X계열 권고'의 문장들로 구성되어 있다. 연구 과정이 초기 단계인 관계로 병렬 코퍼스의 영-한 문장간 정렬(alignment) 및 패턴의 추출 등을 거의 수동으로 처리하였으며, 이러한 이유로 인해 20개의 영어 동사가 본동사로 사용된 1000개의 문장만을 사용하였다.

실험을 위해 사용한 테스트 코퍼스는 학습 코퍼스에

서 대상으로 한 20개의 동사를 본동사로 사용하고 있는 192개의 문장을 사용하였다.

### 5.2 실험 결과

본 논문에서는 문장 패턴의 적용 범위를 실험해서 아래의 표1과 같은 결과를 얻었다.

표 1. 패턴 적용 결과

	패턴 발견	패턴 미발견
문장 수	176 (91.6%)	16 (8.4%)

표1을 보면 학습 코퍼스의 크기가 비교적 작음에도 불구하고 높은 패턴 적용 성공률을 보이고 있다. 이와 같은 이유는 제한된 도메인에서는 문장의 문법적·의미적 구조가 매우 제한적이라는 것을 암시한다. 이것은 현재 진행중인 문장 패턴 추출의 자동화가 어느 정도 이루어지면 기계번역 시스템의 새로운 구축이나 도메인의 변경 등에 필요한 비용을 상당히 줄일 수 있다는 것을 뜻한다.

아래의 표2는 문장 패턴이 적용된 이후 메타 부분에 대한 예문의 검색에서 메타 부분과 일치되는 어구를 포함하는 예문이 어느 정도 발견되는지에 대해 코퍼스의 크기를 달리해서 수행한 실험 결과이다.

표 2. 메타 부분의 일치도 실험

	메타 부분의 개수	코퍼스 크기		
		101K	640K	1.3M
고빈도어에 의한 분할 이전	317	68 (21.4%)	79 (24.9%)	84 (26.4%)
고빈도어에 의한 분할 이후	432	108 (25%)	131 (30.3%)	191 (44.2%)

표2는 메타 부분을 포함하고 있는 예문이 발견될 확률이 코퍼스의 크기와 상관없이 매우 낮다는 것을 보여주며, 이러한 이유로 인해 메타 부분을 구 패턴으로 다시 나눌 필요가 있다는 것을 보여준다.

### 5.3 문제점

참고로 본 논문에서는 언급하지 않았지만, 패턴 적용에도 약간의 모호성은 발생하게 된다. 이러한 이유는 병렬 코퍼스에서 발견되는 똑같은 패턴을 적용할 수 있는 문장이라고 하더라도 그 번역은 서로 다른 경우가 존재하기 때문이다. 또한 패턴 적용시에 두 가지 이상의 패턴을 적용할 수 있는 경우도 발생하였다. 이러한 문제점을 메타 부분에 할당된 중심 단어들의 의미 거리(semantic distance)를 이용해서 해결하려는 시도가 현재 진행중에 있다.

## 6. 결론

본 논문에서는 패턴과 예문을 사용하여 영한 변환을 수행하는 방법을 제안하였다. 패턴을 사용함으로써 해석에서 발생하는 모호성을 상당히 감소시킬 수 있으며, 동시에 유사 예문 발견의 확률을 높일 수 있다. 또한 예문을 사용함으로써 해당 도메인에 대한 자연스러운 번역 품질을 얻을 수 있다.

제안된 방법은 현재 연구중인 패턴의 자동 추출이 완료될 경우, 시스템 개발 및 확장시 발생되는 비용을 상당히 줄일 수 있다. 또한 기존의 규칙으로 처리하기에는 예외적인 경우가 많았던 대화체 등에 적용할 경우, 기존의 대화 시스템에 비해 보다 견고한(robust) 시스템의 개발이 가능할 것으로 여겨진다. 이러한 패턴 및 예문에 기반한 방법의 적용 분야로는 대화 시스템 뿐만 아니라 자연어 인터페이스 시스템, 자연어 검색 시스템 등에도 충분히 활용될 수 있을 것이다.

향후 연구 과제로는 본 논문에서 진행 과제로 언급했던 패턴 추출의 자동화에 대한 연구가 반드시 필요하며, 본 논문에서는 언급하지 않은 패턴 적용에서 발생하는 모호성 해결에 대한 연구가 필요하다.

## 참고문헌

- [1] Nagao. M., "A Framework of a mechanical translation between Japanese and English by Analogy principle," *A. Elithorn and R. Barnerji, eds., Artificial and Human Intelligence*, pp.173-180, 1984.
- [2] S. Nirenburg, C. Domashnev, D. J. Grannes. "Two Approach to Matching in Example-Based Machine Translation," *Proceeding of TMI'93*, pp.47-57, 1993.
- [3] 佐藤 理史, "實例に基づく翻譯," 情報處理, Vol.33, No.6, pp.673-681, 1992.
- [4] 古瀬 藏, 隅田 英一郎, 飯田 仁, "變換主導型機械翻譯の實現手法," 自然言語處理80-8, pp.1-8, 1990.
- [5] 古瀬 藏, 飯田 仁, "變換と解析の協調的處理による翻譯手法," 自然言語處理87-4, pp.27-34, 1992.
- [6] Osamu Furuse and Hitoshi IIDA, "Constituent Boundary Parsing for Example-Based Machine Translation," *Proceeding of COLING94*, pp.105-111, 1994.
- [7] Koich Takeda, "Pattern-Based Context-Free Grammars for Machine Translation," *Proceeding of ACL96*, 1996.
- [8] Hideo Watanabe, "A Method for Extracting Translation Patterns from Translation Examples", *Proceeding of TMI'93*, pp.292-301, 1993.
- [9] 佐藤 理史, アナロジーによる機械翻譯, 共立出版社株式會社, 1997.