영한 자동번역에서의 한국어 분류사의 반자동 구축 방법

이기영○ 최승권 김영길
한국전자통신연구원 언어처리연구팀
{leeky, choisk, ygkim}@etri.re.kr

Semi-Automatic Building of Korean Classifiers in English-Korean MT

Ki-Young Lee○, Sung-Kwon Choi, Young-Gil Kim
Natural Language Processing Team, Electronics and Telecommunications Research Institute

요 약

본 논문은 영한 기계번역에서 영어 수사가 포함된 영어 명사구를 한국어로 번역할 때, 영어 명사에 대응되는 한국어 명사의 적절한 분류사를 반자동으로 구축하는 방법에 대해 기술한다. 영한 번역의 측면에서, 분류사는 목표언어인 한국어에서만 나타나는 현상이다. 따라서 영어를 한국어로 번역할 때, 적절한 분류사를 생성하지 않으면 한국어 어법에 맞지 않는 부자연스러운 번역 결과를 생성한다. 본 논문에서는 한국어 태그드 코퍼스와 한국어 의미코드 체계에 따라 한국어 분류사를 반자동으로 구축하는 방법을 제안한다. 제안하는 방법에 따라 한국어 명사에 대해서 한국어 분류사가 구축되었으며, 이렇게 구축된 분류사는 영한 기계번역시스템의 번역 사전에 'KCOUNT'라는 자질을 할당하여 부가하였다. 제안하는 방법의 검증을 위해 수동평가와 자동평가를 수행하였으며, 그 결과, 영한 기계번역의 문장 생성에 있어서 자연스러움(fluency)의 측면에서 번역률 향상이 있었다.

1. 서 론

국가간 통신 기술의 발달과 함께 기계번역에 관한 요구가 증가하였으며, 이러한 추세와 함께, 규칙 기반 방식을 기반으로 하여 패턴 기반 및 통계 기반과 같은 다양한 방식의 기계번역시스템이 개발되어 왔다. 특히, 영한 기계번역의 경우, 다른 언어셋과 비교할 때 가장 먼저 개발이 시작되었다고 볼 수 있다. 이와 함께, 영한 기계번역시스템을 개발하는데 있어서, 그 개발 이슈(issue)도 점차 언어 분석 기술에서 이제는 영한 변환 및 한국어 생성 기술로 옮겨가고 있는 추세이다.

이러한 맥락에서, 본 논문은 영한 기계번역에서 영어 수사가 포함된 영어 명사구를 한국어로 번역할 때, 영어 명사에 대응되는 한국어 명사의 적절한 분류사를 생성하기 위해, 한국어 분류사를 반자동으로 구축하는 방법에 대해서 기술한다. 영한 번역 관점에서, 분류사는 목표언어인 한국어에서만 나타나는 현상이다. 따라서 영어를 한국어로 번역할 때, 적절한 분류사를 생성하지 않으면 한국어 어법에 맞지 않는 부자연스러운 번역 결과를 생성한다.

(원문1) We maintain two computers.
(번역문1-1) 우리는 2 컴퓨터를 운영하고 있다. (X)
(번역문1-2) 우리는 2대의 컴퓨터를 운영하고 있다. (O)

(원문2) I saw two dogs.
(번역문2-1) 나는 2개를 보았다. (X)
(번역문2-2) 나는 개 2마리를 보았다. (O)

(원문1)은 ‘two computers’라는 수사를 포함하는 명사구(‘수사 + 명사’)를 갖는다. 이러한 수사를 포함하는 명사구에 대해서 적절한 수량사를 생성하지 못하면 (번역문1-1)과 같은 어색한 번역 결과를 생성한다. 즉, (원문1)에 대한 올바른 한국어 번역 결과를 생성하기 위해서는 ‘computer’의 대역어를 생성할 때, ’computer’의 대역어인 ‘컴퓨터’와 함께 ‘컴퓨터’에 대한 적절한 한국어 수량사 ‘대’를 함께 생성해 주어야 한다. 마찬가지로, (원문2)의 경우도 ’dog’에 대한 한국어 분류사 ‘마리’가 ‘dog’에 대한 대역어와 함께 생성되어야 올바른 번역 결과를 얻을 수 있다. 이러한 분류사는 각 명사에 따라 서로 다르다. 즉, ’computer’의 대역어 ‘컴퓨터’에 대한 분류사는 ‘마리’가 아니라 ‘대’이며, ’dog’의 대역어 ‘개’에 대한 분류사는 ‘대’가 아니라 ‘마리’인 것이다. 따라서 영한 기계번역에서 보다 우수한 번역품질을 얻기 위해서는 한국어 어법에 맞는 올바른 분류사를 생성하는 것이 중요하다. 본 논문은 한국어 분류사를 반자동으로 구축하고, 이를 활용하여 영한 기계번역에서 자연스러운 결과를 생성하는 것을 목표로 한다.

본 논문의 구성은 다음과 같다. 2장에서는 한국어 분류사의 유형 및 특성에 대해 살펴보고, 3장에서는 한국어 분류사를 한국어 태그드 코퍼스와 한국어 의미코드 체계를 이용해 반자동으로 구축하는 방법을 기술한다. 4장에서는 구축된 분류사를 영한 기계번역에 적용하는 방안에 대해서 설명하고, 5장에서는 분류사 적용에 따른 번역률 변화에 대한 실험 결과에 대해서 설명하며, 마지막으로 6장에서 결론을 맺는다.

2. 한국어 분류사의 유형 및 특성

한국어 분류사에 대한 선행 연구에서 전형적인 분류사의 성격으로 논의된 것은 다음과 같다 [1].

① 셈의 단위가 된다.
② 명사 특히 의존명사의 하위 부류이다.
③ 구성상 수량사의 수식을 받는다.
④ 용법상의 의미를 지닌다.
⑤ 대상 명사의 의미 유형을 분류한다.

이와 같은 한국어 분류사의 성격에 따라, 본 절에서는 한국어 분류사의 문법적 범주와 특성 및 그 의미적 역할을 설명한다.

[2]는 분류사의 유형을 '수사', '분류사' 및 '명사'의 순서에 따라 두 가지로 나누었다. 아래의 표 1은 이에 따라 분류사의 유형을 나눈 예를 보인다. 소유격 유형은 소유격 조사 '의'가 첨가된 형태로 주로 한국어 문어체에서 사용되며, 명사우선 유형은 소유격 조사를 사용하지 않으며, 주로 한국어 구어체에서 사용된다. 이와 같은 한국어 분류사의 유형은 한국어 코퍼스에서 품사 관계를 이용하여 한국어 분류사의 사용 패턴을 추출할 때 사용된다. [3]은 이러한 한국어 수명사구의 구조를 HPSG에 따라 구조적으로 기술한 바가 있다. 또한 [4]는 한국어 분류사가 의존명사에 한정되지 않으며, 양수 표현 뒤에 위치하고, 특유의 어휘의미를 지닌다는 한국어 분류사의 구조적 특성을 정의하였다.

표 1. 분류사의 유형

	품사 배치	예
소유격 유형	수사+분류사+소유격 조사+명사	두 *대*의 컴퓨터, 세 *개*의 시스템
명사우선 유형	명사+수사+분류사	컴퓨터 두 *대*, 시스템 세 *개*

[5]는 분류사를 도량성(mensural)과 부류성(sortal)의 두 가지로 구분하였다. 도량성 분류사는 대상 명사를 양(quantity)에 의해 개체화하며, 부류성 분류사는 그것이 지시하는 실체의 부류(class)를 나타낸다. 따라서, 한국어 분류사가 용법상의 의미를 지니고, 대상 명사의 의미 유형을 분류한다는 특성에 비추어 볼 때, 한국어 분류사는 관련 명사의 지시물을 부류화하고 개체화한다고 볼 수 있다. 이러한 부류화 기능은 의미적인 성격을 띠는 것으로 볼 수 있다. 따라서 분류사는 부류화가 1차적인 기능이고, 수량화 기능은 부류화 기능에 따르는 부차적인 것이 된다 [1]. 이러한 점에서 한국어 분류사는 관련 명사에 따라 의미적 계층 구조를 형성할 수 있다.

그림 1은 이러한 특성을 이용하여 현재 한영 자동번역 시스템에서 사용하고 있는 의미코드 체계의 일부와 해당 의미코드에 할당할 수 있는 한국어 분류사를 보인다. 그림 1을 보면, 현재 한영 자동번역시스템에서 사용하는 한국어 의미코드 체계가 8레벨로 표시되고, 각 의미코드 노드에 대해서 <분류사 리스트>의 올바른 분류사가 서로 대응되어 있는 것을 알 수 있다. 즉, 의미코드 '포유류', '나무', '화초'에 대한 한국어 분류사로서 각각 '마리', '그루', '송이'가 대응되는 것을 알 수 있다.

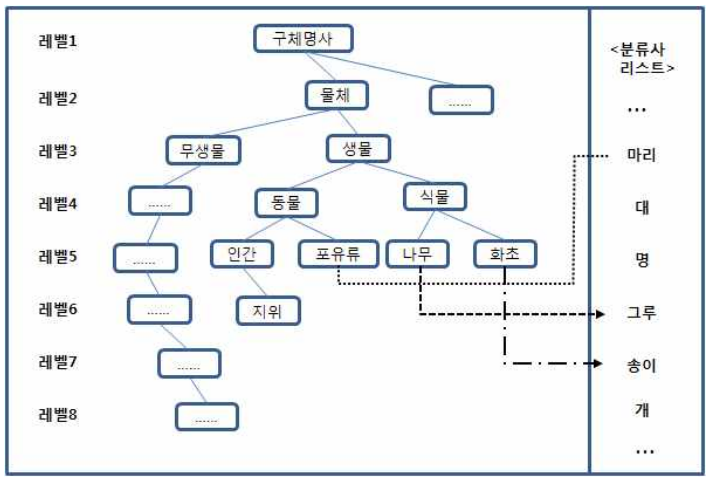

그림 1. 의미코드 체계와 분류사의 대응

한국어 분류사가 명사에 단지 하나만 적용되는 것은 아니다. 그 이유는, 발화 상황에서 명사가 지시하는 대상이 어떻게 인식되는가에 의존하기 때문이다 [1]. 발화 상황에 따른 분류사의 선택적 사용에 대한 예는 다음과 같다.

- 나무를 생물로 인식한 경우: 나무 한 {그루, 포기, 뿌리}
- 나무를 무생물/낱개로 인식한 경우: 나무 한 {개비, 도막}
- 나무를 무생물/집합체로 인식한 경우: 나무 한 {단, 묶음, 짐}

본 논문에서는 분류사가 발화 상황에 따라 선택적으로 사용되는 경우는 다루지 않도록 한다.

3. 한국어 분류사의 반자동 구축

본 절에서는 한국어 태그드 코퍼스와 의미코드 체계[1]를 사용하여, 영어 명사의 대역어에 대응되는 한국어 분류사를 반자동으로 구축하는 방법을 기술한다.

1) 의미코드 체계는 ETRI의 한영 자동번역 시스템에서 번역패턴이나 대역사전에 어휘 의미를 부착하기 위해 사용하는 계층화된 어휘 의미를 말하며, 이런 의미코드 체계는 명사나 동사의 대역어 선택을 위하여 사용되고 있다.

3.1 한국어 태그드 코퍼스의 사용

한국어 태그드 코퍼스로부터 한국어 분류사를 반자동으로 추출하는 방법은 다음과 같이 이루어졌다.

- 단계1. 한국어 코퍼스 수집
- 단계2. 수집된 한국어 코퍼스에 대해서 한국어 태거를 이용하여 한국어 태그드 코퍼스 구축
- 단계3. 한국어 분류사 추출 패턴을 정의하여, 한국어 태그드 코퍼스를 대상으로 한국어 명사 및 대응되는 한국어 분류사 추출
- 단계4. 한국어 명사와 분류사에 대한 수동 검증 작업

한국어 분류사를 반자동으로 추출하기 위해 사용된 한국어 코퍼스는 과학기술 논문 분야에 해당하는 600만 문장이 사용되었다. 한국어 형태소 분석기 및 태거의 경우, 높은 정확률을 보이지만, 100%가 아니라는 점 때문에, 단계3에서, 가능하면 모호성이 없는 특정 패턴에 해당하는 한국어 품사열 만을 대상으로 분류사와 관련 명사 정보를 추출하였다. '단계3' 에서 언급한 한국어 태그드 코퍼스로부터 한국어 분류사를 추출하기 위한 분류사 추출 패턴에 대해 더욱 상세히 기술하면 그림 2와 같다. 한국어 태그드 코퍼스에 나타나는 명사와 분류사를 포함하는 수명사구의 패턴에 있어서 모호성이 없는 품사 배열을 조사하면 그림 2와 같이 몇 가지 유형으로 분류할 수 있다.

<패턴 1> 수사+분류사+관형격조사 명사+조사
예) 2대의 시스템을
<패턴 2> 수사+분류사+관형격조사 형용사+관형사형전성어미 명사+조사
예) 2대의 효율적인 시스템을
<패턴 3> 수사+분류사+관형격조사 성상관형사 명사+조사
예) 3가지의 기본적 방법을
<패턴 4> 수사+분류사+관형격조사 동사+관형사형전성어미 명사+조사
예) 3개의 운영되는 프로세서가
<패턴 5> 수사+분류사 명사+조사(관형격조사 제외)
예) 2대 시스템이
<패턴 6> 명사 수사+분류사+조사(관형격조사 제외)
예) 컴퓨터 3대를…

그림 2. 한국어 분류사를 추출하기 위한 분류사 추출 패턴의 유형

그림 2에서 <패턴 1>에서 <패턴 4>까지는 한국어 수명사구의 구조에서 소유격타입의 수명사구 유형이며, <패턴5>와 <패턴 6>은 명사우선타입의 통사 구조 유형이다. 그림 2에서 제시된 분류사 추출 패턴에 기반하여 600만 문장을 대상으로 한국어 분류사와 관련 명사를 추출한 결과는 표2와 같다.

표 2. 한국어 태그드 코퍼스에서 추출된 명사 및 분류사 통계

구분	개수
명사-분류사 전체 수	9,021
수동 검증 과정에서 수정된 분류사	1,702
수동 검증 과정에서 삭제된 분류사	329

표 2에서, "수동 검증 과정에서 수정된 분류사 ' 및 ' 수동 검증 과정에서 삭제된 분류사 '는 추출된 전체 명사-분류사에 대해서 전문 번역사의 검증 과정에서 수정되거나 삭제된 것을 나타낸다. 이러한 이유는 대부분이 한국어 태깅 과정에서 잘못된 태그가 부착되어서 잘못 추출되었기 때문이다.

3.2 한국어 의미코드 체계의 사용

한국어 의미코드 체계를 구성하는 의미코드는 416개이며 구조적으로는 8레벨로 구성되어 있다 [6]. 한국어 의미코드 체계에 한국어 분류사를 할당하는 작업은 전문번역가가 의미코드 체계에 부착된 용례를 보고 한국어 분류사를 부착하였다. 한국어 의미코드 체계의 어휘 의미와 대응되는 한국어 분류사의 구축 결과는 표 3과 같다.

표 3. 한국어 의미코드 체계를 사용하여 부착된 분류사

의미코드	레벨	부착된 한국어 분류사
인간	4	명
화초	5	송이
나무	5	그루
가공식품	6	개
양말	6	켤레
요리	6	가지
밥류	7	그릇,가지
자동차	8	대

상기에 설명한 방식으로, 의미코드 체계에 대해 한국어 분류사를 부착한 결과, 일반적으로 상위 계층의 의미코드에는 '개', '가지' 와 같은 총칭성 분류사와 '건', 번 '과 같은 사건성 분류사가 부착되는 경향이 있었으며, 하위 계층의 의미코드에 대해서는 보다 구체적인 ' 마리 ', ' 명 ', ' 대 '와 같은 개체성 분류사가 부착되는 경향이 있었다.

한국어 의미코드 체계에서 사용되는 416개의 의미코드는 현재 한영 자동번역 시스템의 한영 번역 사전에 각 엔트리에 대한 의

미로써 할당되어 있으며 그 통계치는 표 4와 같다. 명사 엔트리 당 어휘 의미가 1개 이하인 이유는 엔트리가 복합 명사인 경우 그것의 핵심어에 의해 어휘 의미가 결정되기 때문에 복합 명사에 대해서는 어휘의미를 부여하지 않았기 때문이다.

표 4. 한영 자동번역 사전의 의미코드 통계

구분	개수
전체 명사 엔트리 수	342,812개
의미코드가 할당된 명사 엔트리 수	217,051개
명사 엔트리 당 의미코드 개수	0.63

4. 영한 기계번역시스템에서의 적용

본 논문에서 구축된 한국어 분류사는 영한 기계번역시스템의 영한 사전에 'KCOUNT'라는 자질값으로 추가되었다. 기본적으로 한국어 태그드 코퍼스를 이용하여 구축된 분류사가 우선 적용되었으며, 영어 어휘의 해당 대역어에 대한 분류사가 한국어 태그드 코퍼스로부터 구축되지 않은 경우에는 어휘 의미 체계로부터 구축된 분류사가 적용되었다. 또한 이 과정에서 구축되지 않은 명사의 분류사에 대해서는 분류사 생성시에 가장 보편적인 분류사에 해당하는 '개'를 적용하였다. 분류사 중에서 '개'는 가장 넓은 범위로 나타나는 형상성에 대한 보편 분류사라 할 수 있다 [7].

아래의 그림 3은 이렇게 사전에 부가된 KCOUNT 자질값의 예를 나타낸다. 그림 3에서는 설명을 위해 기타 번역에 필요한 자질을 생략하였다. 그림 3을 보면 'car'와 'peony'에 대한 한국어 대역어와 대응되는 분류사 정보가 각각 'KROOT' 및 'KCOUNT'라는 자질의 값으로 할당되어 있는 것을 알 수 있다. 또한 복합어에 대해서는 복합어의 헤드 명사에 대해 동일한 처리를 수행하였다.

```
[KEY]
car@NOUN
[CONTENT]
{변환
[ ..... (KROOT 자동차)(KCOUNT 대) .....]

[KEY]
peony@NOUN
[CONTENT]
{변환
[ ..... (KROOT 작약)(KCOUNT 송이) .....]
}
```

그림 3. 분류사를 포함하는 영한 사전의 예

기본적으로 영한 기계번역시스템은 위에서 기술한 바와 같이, 일련의 과정을 통해서 구축된 분류사 정보를 사전에 'KCOUNT'라는 자질의 자질값으로 부가하였으며, 이렇게 부가된 'KCOUNT' 정보를 이용하여 한국어 생성을 수행한다.

5. 실험

본 절에서는 영한 기계번역시스템에서 수량사 생성이 번역률에 미치는 영향에 대해 실험한 결과를 설명한다. 제안하는 한국어 수량사 구축 및 적용 방법의 평가를 위해 수동평가와 자동평가를 수행하였다.

5.1 수동 번역률 평가

수동평가를 위해 사용한 테스트셋, 평가 방법, 평가 기준을 기술하면 다음과 같다.

- 테스트셋: 다양한 과학기술논문에서 추출한 400 문장으로 구성되었으며, 전체 테스트셋의 평균 단어수는 17.05 단어이다.
- 평가 방법: 5인의 전문 번역가에게 정확성에 관한 점수를 교육한 후 평가 기준에 따라 각자 평균 점수를 부여하게 하고 5인의 점수 가운데 최고/최저를 제외한 3인의 점수에 대한 평균으로 번역률을 계산하였다.
- 평가 기준: 표 5 참조.

표 5. 수동평가 평가 기준

점수	평가 기준
4.0	원어문의 의미가 그대로 전달된 경우
3.5	복문에서, 문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1-2 단어의 대역어가 잘못된 경우
3.0	문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달되는 경우
2.5	하나의 동사절이라도 정확히 번역되어 부분적으로 문장의 의미를 전달할 경우
2.0	하나 이상의 구가 정확히 번역되지만 전체적인 문장의 의미를 파악하기 어려운 경우
1.0	문장 중에 하나의 단어 또는 구라도 정확히 번역된 경우
0.0	번역문 출력이 안 된 경우

위와 같은 평가 기준에 따라 400문장으로 구성된 테스트셋에

대해 영한 기계번역시스템을 평가한 결과는 표 6과 같다.

표 6. 수동평가 결과

	전체 번역률	체감 번역률
분류사 미생성	81.10%	75.75%
분류사 생성	81.23%	78.25%

표 6은 수동평가 결과를 나타내며, 분류사를 생성함으로써, 전체번역률 및 체감 번역률의 향상이 있었음을 볼 수 있다. 또한 분류사의 생성이 번역 결과의 자연스러움에 기여하는 효과가 크기 때문에 3점 이상으로만 구성되는 체감 번역률의 향상에 영향을 많이 주고 있음을 알 수 있다.

수동평가를 위한 테스트셋의 경우, 모든 문장이 수명사구를 포함하는 것은 아니기 때문에, 수명사구를 포함하는 문장 100문장을 추출하여 번역을 수행한 후, 생성된 한국어 분류사에 대한 평가도 수행하였다. 그 결과는 표 7과 같다. 분류사 생성이 잘못된 18문장의 오류는 분류사를 생성할 필요가 없는 경우와 앞서 언급한 '발화상황에 따른 선택 오류'에 해당하였다. 특히, 분류사를 생성할 필요가 없는 명사의 경우, 해당 명사를 수집하여 분류사 생성 정확률을 개선할 필요가 있다.

표 7. 분류사 생성 정확률

수명사구 포함 문장	100 문장
분류사 생성 정확률	82%

5.2 자동 번역률 평가

우리는 또한 1000문장으로 구성된 5인 레퍼런스를 포함하는 정답셋에 대해서 BLEU(Bilingual Evaluation Understudy) 스코어를 측정하였다 [8]. 표 8은 자동평가 결과를 나타낸다. 표 8에서 BLEU-EOJ은 어절 단위의 자동평가 결과를 나타내며, BLEU-MA는 형태소 단위의 자동평가 결과를 나타낸다. 자동평가 결과 두 가지 측정 단위에 있어서 모두 번역 정확률의 향상이 있었음을 알 수 있다.

표 8. 자동평가 결과

	BLEU-EOJ	BLEU-MA
분류사 미생성	0.3359	0.5603
분류사 생성	0.3368	0.5612

수동평가와 자동평가를 통한 실험 결과, 의미코드에 기반하여 분류사를 적용할 경우, 일부 어휘들에 대해서는 모호성이 발생하였다. 즉, 특정 의미코드의 경우에 대해서는 분류사 적용에 있어서, 모호성이 발생하며, 이러한 문제를 해결하기 위해서는 특정 레벨의 의미코드에 대해서는 세분화가 필요함을 알 수 있었다.

6. 결론

본 논문에서는 영한 기계번역시스템에서 영어 수사가 사용된 명사구를 한국어로 자연스럽게 번역하기 위해 한국어 생성에 필요한 한국어 분류사를 반자동으로 구축하여 시스템에 적용하는 방안을 기술하였다.

한국어 분류사를 반자동으로 구축하는 방법은 두 가지 방법에 의해 이루어졌다. 첫 번째 방법은 한국어 형태소 분석기 및 태거를 사용하여 구축된 한국어 태그드 코퍼스로부터 한국어 분류사를 반자동으로 추출하여 구축하는 방법이며, 두 번째 방법은 한국어 의미코드 체계에 한국어 분류사를 수작업으로 부착하여 구축하는 방법이었다. 이렇게 구축된 한국어 분류사는 시스템의 영한 사전에 KCOUNT라는 부가 자질을 통해 추가되었으며, 해당 수명사구의 번역 과정에서 KCOUNT 자질의 정보를 이용하여 분류사가 번역 결과에 생성되었다.

실험은 수동평가와 자동평가의 두가지로 수행되었으며, 실험 결과 번역의 정확률 뿐만 아니라 번역 결과의 자연스러움 측면에서 많은 향상이 있었다. 본 논문의 한국어 분류사 생성에 사용된 코퍼스는 과학기술논문에 한정되었었으나 향후에는 한국어 분류사의 추출 및 그 대상을 웹문서를 대상으로 확장할 계획이다. 또한 의미코드 기반의 분류사 생성에 대해서는 현재의 의미코드 체계를 추가적으로 검토하여 분류사 적용 단계에서 발생하는 모호성을 최소화하는 것이 필요하다.

<참고문헌>

[1] 우형식, 한국어 분류사의 기능과 범위, 한글, 한글, 한글학회, 2000.

[2] 이남석, 명사구의 통사적 구성에 있어 수량 단위 표현의 기능, 독일문학, 43권 1호, 2002.

[3] Kim et al., Processing Korean Numeral Classifier Constructions in a Typed Feature Structure Grammar, Lecture Notes in Computer Science, Vol. 4188, 2006.

[4] 임홍빈, 국어 분류사의 성격에 대하여, 국어학의 새로운 인식과 전개(김완진 선생 회갑 기념 논총), 1991.

[5] Lyons, J., Semantics I, Cambridge University Press, 1979.

[6] Hong et al., Customizing a Korean-English MT System for Patent Translation, MT Summit, 2005.

[7] 우형식, 2001, 형상성 분류사의 범주화 분석, 외대어문논문집, 16권, 2001.

[8] Kishore Papineni et al., BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the ACL, 2002.