사전 재구성과 대역어 정보를 통한 동사구 패턴의 확장 및 관리

홍문표[0] 김영길 류철 최승권 박상규
한국전자통신연구원 휴먼정보처리연구부
{hmp63108, kimyk, ryuch, choisk, parksk}@etri.re.kr

Extension and Management of Verb Phrase Patterns based on Lexicon Reconstruction and Target Word Information

Munpyo Hong[0], Young-Kil Kim, Chul Ryu, Sung-Kwon Choi, Sang-Kyu Park
Dept. of Human Information Processing, ETRI

요 약

데이터 기반 기계번역의 성공여부는 대량의 데이터를 단기간에 구축하는 방법과, 또 구축된 데이터에 대한 효과적인 관리 방법이 좌우한다고 할 수 있다. 대표적인 데이터 기반 기계번역 방법론인 예제 기반 기계번역 방식이나 패턴 기반 기계번역 방식에서는 최소한의 학습 내지는 학습과정 없이 데이터를 구축하는 데에 연구가 중점적으로 이루어져왔으나, 데이터의 관리 문제에 대해서는 많은 연구가 이루어지지 못하였다. 그러나 데이터의 확장 못지않게 데이터의 효율적인 관리도 데이터 기반 기계번역 시스템의 개발에서 매우 중요하다.

이 논문에서는 사/피동 링크 등을 이용하여 사전을 재구성하는 것이 데이터의 일관성과 관리성을 향상시키고, 이론적인 면에서는 정보 기술상의 잉여성을 줄인다는 점을 보인다. 또한 이러한 정보에 기반하여 기구축된 동사구 패턴으로부터 대역어 정보를 이용하여 새로운 패턴을 만들어내는 방법론도 제시한다.

1. 서론

규칙 기반 방식 (Rule-based MT)과 더불어 기계번역 시스템 구축의 주요한 방법론인 데이터 기반 기계번역 (Data-driven MT)의 성공여부는 데이터 구축의 속도와 데이터의 품질이 좌우한다. 대표적인 데이터 기반 기계번역 방법론인 예제 기반 기계번역 방식 (Example-based MT)이나 패턴 기반 기계번역 방식 (Pattern-based MT)에서는 최소한의 학습이나 학습과정 없이 데이터를 구축하는 데에 연구가 중점적으로 이루어져 왔으나, 데이터의 관리 문제에 대해서는 많은 연구가 이루어지지 못하였다.[1] 그러나 데이터의 확장 못지않게 데이터의 효율적인 관리도 데이터 기반 기계번역 시스템의 개발에서 매우 중요하다고 할 수 있다.

본 논문에서는 현재 한국전자통신연구원 (ETRI)에서 개발중인 한-중 기계번역 시스템에서 사용되는 번역 지식들을 관리하고 확장하는 방법론에 대해 논의한다.[2]

한-중 기계번역 시스템에서 사용되는 형태소 사전과 동사구 패턴과 같은 번역 지식은 중국어 전공의 사전편집자 (lexicographer)들에 의해 구축되어 왔으나, 수작업이라는 한계 때문에 속도와 비용, 작업 결과의 일관성, 작업의 효율성이라는 문제가 있었다.[3] 이러한 문제들을 해결하기 위해 기 구축된 지식을 이용한 동사구 패턴 등의 (반) 자동 생성 및 지식 관리 프레임워크의 재구성 등을 시도하고 있다.

본 논문에서는 동사의 기본형과 사/피동형들 간의 상호 링크를 도입하고 사/피동 자동변환 규칙을 이용하면 패턴의 수동 구축 과정을 반 자동화 할

[1] 패턴 기반 기계번역 시스템과 예제 기반 기계번역 시스템의 지식 확장 방안에 대해서는 각각 [6] 과 [4] 참조

[2] ETRI 한-중 기계번역 시스템에 대해서는 [7] 참조

[3] 중국어 전공자들에 의해 구축된 지식은 언어학 전공의 한-중 이중언어 화자들에 의해 감수과정을 거쳤다

수 있고 작업 결과의 일관성을 보장할 수 있음을 보인다. 또한 이러한 접근 방법은 정보 기술 시 잉여성 (Redundancy)을 줄일 수 있다는 장점이 있다. 또한 [5]와는 달리 사/피동 링크를 도입하게 되면 대역어 정보를 이용하여 기존 동사패턴에서 자동으로 새로운 패턴을 만들어낼 때 패턴의 정확률을 높일 수 있음을 실험을 통해 보인다.

2장에서는 패턴 확장과 관리 방안에 대한 기존의 연구들을 소개하고 각각의 문제점들을 지적한다. 3.1장과 3.2장에서는 한-중 기계번역 시스템에서 사용되는 지식 포맷에 대해 간략히 언급하고 3.3장에서는 사/피동 링크의 구축에 대해 설명한다. 또한 3.4장에서는 사/피동 링크와 대역어 정보에 기반한 패턴 확장방안도 소개한다. 3.5장에서는 패턴 확장방안에 대한 실험이 이루어진다. 끝으로 4장에서는 본 연구의 결론이 제시된다.

2. 관련연구

[1]은 ALT-J/E 시스템의 사전을 동사의 사/피동 현상, 목적어 현상들과 같은 동사의 교대(Alternation)현상에 기반하여 계층적으로 구성하였다. 이러한 계층적인 사전 구성은 정보기술에 있어서 최소한의 잉여성만을 허용하며 정보의 공유를 가능하게 한다는 장점이 있음을 주장하였다. 그러나 계층적으로 구성된 사전에서 격틀이 동사의 태(Voice)에 따라 구체적으로 어떻게 표상화 되는지에 대한 언급이 누락되어 있는 문제가 있다

[2]와 [8]에서는 각각 일본어와 한국어에서 사/피동 현상들을 이용해 패턴을 (반)자동으로 확장하는 방법론을 제안하였다. 이 연구들은 기존의 패턴에서 어떠한 알고리즘을 통해 사/피동형 동사들의 패턴을 생성해내느냐를 다루고 있으며, 사전의 재구성 필요성에 대해서는 다루지 않고 있다.

[5]는 일본어 패턴에서 대역어 (영어)의 정보를 이용하여 기존의 패턴을 확장하는 방법을 제안하였다. 그러나 단순한 대역어 정보만으로는 한국어와 중국어의 경우에서처럼 언어쌍에 따라 자동 생성 패턴의 정확률이 낮아질 수 있는 문제가 있다.

3. 사/피동 링크와 대역어를 이용한 패턴 확장

본 장에서는 형태소 사전을 파생동사의 원형과 사.피동 동사간에 상호링크를 통하여 재구성한 후 대역어 정보를 이용하여 한-중 동사구 패턴을 확장하는 방안에 대해 설명한다. 구체적인 방법론에 들어가기에 앞서 3.1 과 3.2에서 현재 한-중 기계번역 시스템에서 사용하는 지식 포맷에 대해 언급한다.

3.1 형태소 사전

한중 기계번역 시스템의 형태소 사전은 키워드, 형태소 코드, 빈도수, 영어 대역어 (EROOT), 중국어 대역어 (CROOT), 의미코드 (SEM), 중국어 수량사 (NC) 정보 등을 담고 있다:

```
'개'
120021 247
{ [ (EROOT dog) (CROOT 狗) (SEM 포유류) (NC 条) (NC 条 마리) ]
[ (EROOT mouth_of_a_river) (CROOT 江口) (SEM 지리)]}
```

예 1: 명사 엔트리의 예

```
'먹다'

600014 1047
{ [ (EROOT eat) (CROOT 吃) (SEM ) (EPOS verb) (ETYPE ) ] }
```

```
'먹이다'

600012 61
{ [ (EROOT feed) (CROOT 喂) (SEM ) (EPOS verb) (ETYPE ) ] }
```

```
'먹히다'

600012 29
{ [ (EROOT be_eaten) (CROOT 吃) (SEM ) (EPOS verb) (ETYPE ) ] }
```

예 2: 동사 엔트리의 예

동사 엔트리의 예에서 볼 수 있는 바와 같이 기본형 동사 '먹다' 와 이로부터 파생된 사동형 '먹이다' , 피동형 '먹히다' 는 서로 독립된 엔트리로서 등록이 되어 있다. 이러한 사전의 구성은 지식 (동사구 패턴)을 추가할 때 지식의 일관성을 해치며 기술상의 잉여성 (descriptive redundancy)을 가져온다.

3.2 동사구 패턴

한중 기계번역 시스템에서 동사구 패턴은 대역어의 선택뿐만 아니라 한국어 구조 분석에도 사용되는 지식이다. 동사구 패턴은 일종의 확대된 하위범주화 틀 (subcategorization frame)로 볼 수 있는데, 일반 언어학 이론의 하위 범주화 틀과의 가장 큰 차이점은 필수 논항뿐만 아니라 번역에 영향을 미칠 수 있는 수의 논항 (optional argument) 및 부사등과 같은 부가어들도 기술한다는 점이다. 또한 대역어 부분이 항상 달려 있다는 점에서도 다

르다고 할 수 있다: [7].

형태소 사전에서와 마찬가지로 동사구 패턴에서도 파생관계에 있는 동사들은 별도의 엔트리로서 존재한다:

```
'먹다'

먹다13 : A=사람!가 B=식품!를 먹!다 > A 吃:v B [그가 밥을 먹었다]
먹다14 : A=사람!가 B=액체!를 먹!다 > A 喝:v B [우리는 우유를 먹었다]
먹다15 : A=사람!가 B=약품!를 먹!다 > A 服:v B [그가 소화제를 먹었다]
```

```
'먹이다'

먹이다6 : A=사람!가 B=재화!를 C=사람!에게 먹이!다 > A 給 C 塞:v B [김 대리는 거래처 직원에게 뇌물을 먹였다]
먹이다7 : A=사람!가 B=식품!를 C=사람!에게 먹이!다 > A 給 C 喂:v B [그가 나에게 소꼬리곰탕을 먹였다]
먹이다8 : A=사람!가 B=약품!를 C=사람!에게 먹이!다 > A 給 C 吃:v B [엄마가 나에게 감기약을 먹였다]
```

```
'먹히다'

먹히다6 : A=재화!가 B=사건!에 먹히!다 > B 费:v 不少 A [이번 일에 돈이 많이 먹혔다]
먹히다12 : A=포유류!가 B=사람!에게 먹히!다 > A 被 B 吃掉:v [토끼가 사람에게 먹히다]
```

예 3 : 동사구 패턴의 예

동사구 패턴의 예에서 A,B,C 등은 변수로서 명사구들이 나올 수 있는 격슬롯 (case slot)을 의미하며 격슬롯에 대한 의미제약이 의미코드 (재화, 사건, 포유류등)에 의해 기술되어 있다. 격 정보는 대표격 (*가, 를*)등에 의해 표시된다.

이러한 기술 방법의 문제점은 많은 부분의 격슬롯에 대한 의미제약 정보를 공유하는 동사의 원형과 사/피동 형태 동사들간의 정보 공유 (Information Sharing)가 일어나지 않으므로 동사패턴 구축 작업의 효율성이 떨어지고 작업 결과의 일관성이 떨어진다는 데에 있다. 예를 들면 기본형 동사 '먹다' 의 'A=사람!가 B=액체!를 먹!다' 패턴에 대해 사동형 패턴 'A=사람!가 B=액체!를 C=사람!에게 먹이!다' 패턴이 존재하지 않을 수도 있다.

3.3 사/피동 링크의 구축

기존의 형태소 사전을 사/피동 링크등에 따라 재구성하기 위해 새로운 사전의 포맷이 필요하다. [3]과 [1]에서는 사전을 단어레벨 (word level), 의미레벨 (sense level), 격틀레벨 (frame level)과 같은 계층적 (hierarchical) 구조로 나누었다. 단어 레벨에는 기본형과 사/피동 파생형들이 공유하는 정보들이 기록되어 있다. 사/피동 파생형에 대한 링크는 기본형 동사의 단어레벨에 기록되어 있어 사/피동 파생형들도 일부 정보를 기본형 동사와 공유할 수 있도록 되어 있다.

현재 개발중인 한-중 기계번역 시스템에서는 형태소 사전과 동사구 패턴 사전이 분리되어 있으므로 형태소 사전에서 동사 엔트리들간의 링크 관계를 표시하고 각 동사 엔트리들과 동사구 패턴 엔트리들을 연결해주는 방식으로 접근한다.

형태소 사전에서 사/피동형태가 파생될 수 있는 동사의 경우 사동 (CAUS), 피동 (PASS) 동사의 어휘 정보를 기록한다. 사/피동형 동사의 엔트리에는 파생이 비롯된 원형 (BASE)에 대한 어휘정보를 포함한다[4]:

```
'먹다'

600014 1047
{ [ (EROOT eat) (CROOT 吃) (CAUS 먹이) (PASS 먹히) ] }
```

```
'먹히다'

600012 29
{ [ (EROOT be_eaten) (CROOT 吃) (BASE 먹)] }
```

```
'먹이다'

600012 61
{ [ (EROOT feed) (CROOT 喂) (BASE 먹)] }
```

예 4 : 사/피동 링크가 첨가된 엔트리의 예

이렇듯 상호 링크가 필요한 이유는 기본형에서 파생형으로만 링크를 할 경우 다음과 같은 문제가 생기기 때문이다:

> 놀다 (play) → 놀리다 (let play)/ 놀리다 (make fun of)

[4] 논지 전개상 불필요한 사전 자질을 생략하였음을 밝힘

사동화 접미사 '-리' 가 붙어서 된 '놀리다 (let play)' 와 어휘 자체로서의 형태인 '놀리다 (make fun of)' 의 경우처럼 동일한 형태를 가지는 어휘들이 존재할 때 파생형으로부터 기본형으로의 링크가 없으면 모호성이 발생하여 동사구 패턴의 작성시 등에 문제가 있다.

기본형 동사의 사/피동 값으로 들어가는 어휘들에는 대부분 '이,히,리,기,우,구,추' 등과 같은 접미사로 이루어진 형태들과 형태론적인 관련성은 없지만 의미적으로는 사/피동 관계를 이루는 '사다-팔다', '주다-받다' 등의 어휘가 포함된다. 다만 사동형의 경우 '게-하' 로 이루어진 형태들은 그 생산성(productivity)으로 인해 엔진에서 처리한다.[5]

3.4 사/피동 링크를 이용한 패턴의 반 자동 구축

동사구 패턴을 여러 명의 작업자들에 의해 수동으로 구축할 경우 생기는 문제는 작업 속도와 비용의 문제뿐 아니라 3.2에서 보인 바와 같이 작업 결과의 일관성 및 효율성이 떨어진다는 단점이 있다.

이러한 문제를 해결하기 위해 3.3에서 제안한 형태소 사전의 사/피동 링크와 사/피동 자동 변환 규칙등을 이용할 수 있다. 작업자들이 동사구 패턴을 구축 시 해당 동사가 사/피동이 가능한 동사일 경우 워크벤치는 사/피동 링크와 사/피동 자동 변환 규칙을 사용해 사/피동형에 임의의 패턴을 만들어 작업자에게 제시한다. 이 패턴을 아무 수정 없이 사용할 수 있는 경우에는 작업자는 패턴에 해당되는 예문만을 추가한 뒤 저장한다. 이 패턴에 약간의 수정이 필요한 경우에는 수정한 후 예문을 첨가한 뒤 저장한다:

(작업자) "사람!가 식품!를 먹다"를 추가할 경우
➔
(UI) "식품!가 먹히다"를 추가하시겠습니까?

(작업자)
예 ➔ 그대로 채택/ 수정후 채택
그대로 채택 ➔ 작업자에게 예문 추가를 요청한 후 저장
수정후 채택 ➔ 패턴을 수정하고 예문을 추가한 후 저장

예 5: 동사구 패턴의 반 자동 구축의 예

[5] 한국어 사/피동 분류는 [9] 참조

사/피동 링크를 도입하면 그림 1과 같은 통합 지식 관리/확장 프레임워크의 모습을 지니게 된다.

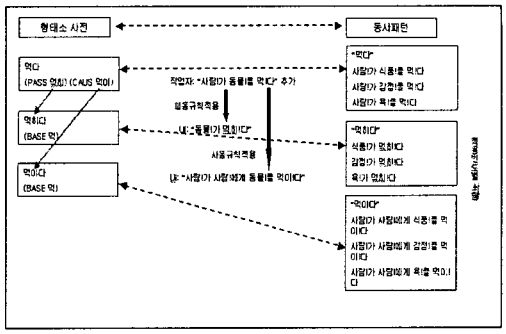

그림 1: 통합 지식 관리/확장 프레임 워크

3.5 사/피동 링크와 대역어 정보를 이용한 패턴 확장

[5]의 연구에서는 대역어 (영어)정보를 이용하여 일본어 패턴을 확장하는 방법론을 제안하였다. 대역어를 이용한 동사의 의미별 분류는 의의가 있다고 본다. 왜냐하면 동사를 의미별로 분류하는 것 자체가 어려운 작업이고 분류를 하더라도 분류의 기준에 의해 그 통사적 실현이 완전히 달라질 수 있기 때문이다. 그 분류의 한 기준으로서 대역어 정보는 패턴 확장을 위해 좋은 기준이 될 수 있음은 [5]의 연구에서 이미 입증되었다.

현재 한-중 번역시스템에는 총 16065개의 동사 키워드에 대해74,880개의 동사구 패턴이 구축되어 있어, 키워드 1개당 평균 약 4.7개의 패턴을 보유하고 있다. 대역어를 이용한 패턴 확장 방법론을 한-중 시스템에도 적용할 수 있는가를 테스트하기 위해 다음과 같은 실험을 실시하였다. 우선 전체 동사구 패턴 DB에서 100개의 중국어 동사 (즉, 대역어) 를 무작위로 추출하고 이 중국어 동사를 대역어로 취하는 한국어 동사구 패턴을 추출하였다. 추출된 동사구 패턴을 동사별로 소팅한 후 동사구 패턴에 대해 서로 비교하여 누락된 패턴을 자동으로 추가하는 방식으로 패턴을 확장하였다 (예6 참조).

단순히 대역어 정보만을 가지고 패턴을 자동 확장하는 방법은 40.5%의 정확률만을 기록하였다. 대역 정보만을 사용하는 방법의 문제점은 한국어에서는, 예를 들어 "전개하다", "벌어지다" 와 같이, 태 (Voice)가 다른 동사들이, 중국어에서 많은 경우에 동일한 동사, '展开', 로 번역되는 경우가 많기 때문이다. 이러한 경우에 능동과 피동이라는 구조적인 차이에 의해 동사구 패턴 (격틀)의 실

현 양상이 상이하게 보여질 것이라는 점은 명확하다.

'參加'를 대역어로 가지는 동사구 패턴

[가담하다]
A=사람!가 B=조직!에 가담하!다 > A 參加:v B [그가 의병에 가담하다]

[참가하다]
A=사람!가 B=싸움!에 참가하!다 > A 參加:v B [우리나라는 걸프전에 참가했다]

확장 결과
➔

[가담하다]
A=사람!가 B=조직!에 가담하!다 > A 參加:v B [그가 의병에 가담하다]
A=사람!가 B=싸움!에 가담하!다 > A 參加:v B [우리나라는 걸프전에 가담했다]

[참가하다]
A=사람!가 B=싸움!에 참가하!다 > A 參加:v B [우리나라는 걸프전에 참가했다]
A=사람!가 B=조직!에 참가하!다 > A 參加:v B [그가 의병에 참가하다]

예 6: 대역어 기반 동사구 패턴 확장의 예

따라서 대역어 정보를 사용한 자동확장의 경우에는 태(Voice)정보가 같은 경우끼리만 비교, 확장하는 방법을 고안하였다. 이 결과 61.23%의 정확률을 기록할 수 있었다:

	태(Voice)정보/대역어 정보 기반	대역어 정보 기반
정확률	61.23%	40.50%

표 1: 동사구 패턴 자동 확장 실험 결과

정확률을 저하시키는 요인 중의 하나는 동사가 연어 (Collocation)의 성격을 가지고 있을 때이다. 예를 들어 '육상부에 참가하다' 라는 의미의 '육상부에 들다' 에서 '들다' 동사는 조직을 나타내는 명사구들과 결합할 때만 '參加' 의 의미를 지니므로 다른 패턴들에서 복사해 온 패턴들에 적용되면 많은 잘못된 패턴들을 생성해내는 결과를 낳는다.

4. 결론

이 논문은 동사구 패턴 기반 기계번역 시스템에서 한국어 분석과 대역어 선택의 핵심적인 역할을 하는 리소스인 동사구 패턴을 효율적으로 관리, 확장하는 방안에 대해 다루었다. 본 논문에서 주장된 것은 첫째, 동사구 패턴의 일관성 있는 구축과 관리성의 향상을 위해 동사의 사/피동 링크를 도입해야 함을 보였다. 사/피동 링크는 모호성의 해소를 위해 상호 링크의 모습을 띄어야 함도 주장되었다. 둘째로는, 사/피동 링크가 도입될 경우 패턴의 사전 작업자들이 동사구 패턴을 추가할 때 관련된 사/피동형에 대한 패턴 구축도 연계될 수 있어서 지식구축의 측면에서 일관성과 정보 기술상의 효율성이라는 장점을 얻게 됨을 보였다. 마지막으로 대역어 정보를 이용한 동사구 패턴의 확장에서도 태(Voice)정보를 이용하는 방법이 [5]에서 시도된 것과 같은 대역어 정보만을 이용하는 방법보다 더 높은 정확률을 올릴 수 있음이 주장되었다.

5. 참고 문헌

[1]. Baldwin & Bond (2002): Alternation-based Lexicon Reconstruction, in TMI 2002

[2]. Baldwin & Tanaka (2000): Verb Alternation and Japanese – How, What and Where?, in PACLIC 2000

[3]. Baldwin, Bond & Hutchinson (1999): A Valency Dictionary Architecture for Machine Translation, in TMI 1999

[4]. Carl, M. (1999): Inducing Translation Templates for Example-Based Machine Translation, in Proceedings of MT Summit

[5]. Fujita & Bond (2002): A method of Adding New Entries to a Valency Dictionary by Exploiting Existing Lexical Resources, in TMI 2002

[6]. Imamura K. (2002): Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT, in TMI 2002

[7]. Kim, Hong, Huang, Kim, Yang, Seo & Choi (2002): Korean-Chinese Machine Translation Based on Verb Patterns, to appear in the Proceedings of AMTA 2002

[8]. Yang, Hong, Kim, Kim, Seo & Choi (2002): An Application of Verb-Phrase Patterns to Causative/Passive Clause, in IASTED2002

[9]. 이익섭 & 임홍배 (1983): 국어 문법론, 학연사