병렬 코퍼스로부터의 대역 표현쌍 추출: 과정, 원리 및 교훈

노용균
(충남대학교)

1. 서론

자연언어의 처리 능력을 모형화하는 일은 그 기본 가정과 방법에 있어서 뚜렷이 구별되는, 언어를 대하는 두 관점으로 나뉘어져 있다. 하나는 오랜 역사를 가진 `부호체계로서의 언어'라는 관점이고 다른 하나는 `발화 행위의 산물로서의 언어'라는 관점이다. 후자는 비교적 근래에 연구가 활성화되었는데, 이 관점을 취하는 연구자들은 주로 다양한 대규모의 코퍼스를 조성하고 코퍼스를 이용해서 경험적으로 언어현상을 관찰하고 기술하는 데에 치중한다. 이 두 관점은 긴장-대립관계에 있는 것은 사실이지만, 점차적으로 이들이 상호보완적 성격을 띠는 것으로 이해하는 학자들이 늘어 나고 있는 추세다.

두 언어 사이의 구조 면에서의 공통점이나 차이점을 찾아 낸다든지, 두 언어 사이의 텍스트의 번역을 시도한다든지 할 때에는 병렬 코퍼스의 이용이 유용하다. 역사상 가장 방대한 병렬 코퍼스는 캐나다 의회 회의록 The Canadian Hansard이다. 최근의 코퍼스 언어학과 기계번역 분야에서 가장 빈번히 이용되는 이 회의록의 모든 글들은 영어와 프랑스어 두 언어로 쓰여 있다. 캐나다의 공용어가 이 두 언어라는 역사적 우연이 이 두 언어 사이의 대응관계 연구를 유별나게 활성화시킨 셈이 된 것이다.

텍스트의 자동 번역과 관련된 연구에서 기본적으로 필요한 것이 두 언어의 낱말들 사이의 동의 관계에 관한 정보이다. 즉, 양어 사전 (bilingual dictionary)이다. 양어 사전은 기계 번역에서뿐 아니라 평범한 인간 사용자들에게도 유용하다. 양어 사전이 어떤 미시구조 (microstructure)와 거시구조 (macrostructure)를 갖는 것이 이상적이냐 하는 문제는 사전학의 연구과제일 테지만, 그 어떤 구조의 사전이라 하더라도 두 언어의 대역관계에 놓이는 낱말 쌍들을 제시해야 할 것이다.

이 논문은, 병렬 코퍼스가 대규모로 만들어져 있음직한데도 전혀 만들어져 있지 않은, 그러면서도 두 언어가 영어 -- 프랑스어처럼 인도유럽어족의 언어가 아닌, 한국어와 중국어 양어 코퍼스를 조성하는 과정과 이렇게 조성된 중한 병렬 코퍼스를 바탕으로 대역표현쌍을 자동 추출하는 기법을 기술한다. 제2절에서는 이 작지만 자체적으로 완전한 병렬 코퍼스의 구축 과정과 내부 형식을 기술하고 제3절에서는 문장정렬체들로부터 자동으로 대역표현쌍들을 얻어 내기 위한 기준 역할을 하는 상관계수 (correlation factor) 하나를 논하며, 제 4절에서는 이 기준을 이 코퍼스에 적용할 경우에 결과물이 갖는 정확도를 평가한다. 마지막 절에서는 쓸모가 큰, 상대적으로 완전한 한--중 중--한 사전의 편찬을 반자동화하기 위해서는 어떤 코퍼스가 어느 규모로 구축되어야 할지, 그리고 그러한 대규모 코퍼스를 구축하는 데에 어떤 자원이 소요되는지를 간략히 논의한다.

2. 병렬 코퍼스의 조성

이 절에서는 중한 병렬 코퍼스의 한 모델로서 저자가 조성한 "충실코 중한 병렬 코퍼스 1

(허삼관매혈기)"의 조성과정과 내부 구조, 그리고 그 크기를 기술하고자 한다.

2.1. 정렬체의 종류

이 코퍼스는 중국의 당대 문학 작가 余华의 1995년 소설 [许三观卖血记]와 이것의 한국어 번역본인 [허삼관 매혈기]로 이루어져 있다.[1] 중국어 원본과 한국어 대역본의 디지털화된 텍스트를 획득하고 이 두 글을 장(chapter)으로 각각 나누어 정렬하였다. 모두 스물 아홉 개의 장들로 이루어진 소설이니만큼, 가장 간단한 정렬로는 스물 아홉 개의 정렬 단위가 생긴다.

그러나 이들 장들은 모두 각각 수천 개의 낱말들을 담기 때문에 이 정렬을 바탕으로 얻어낼 수 있는 대역표현쌍은 그 크기가 너무 커서 아무 쓸모가 없는 것이다. 우리는 언어의 의미층위에서의 최소 단위인 낱말들 사이의 대역 관계를 추출하는 것을 기본 목표로 하는 만큼, 그리고 장 단위의 대역관계로부터 낱말 단위의 대역관계를 추출하려면 이미 대역 관계가 확립된 장들의 수효가 수십 개여서는 안된다. 적어도 수만 개는 있어야 할 것이다.

각 장은 문단들로 이루어져 있다고 보고, 장마다 문단들을 분리해서 문단간 대역관계를 수립한다. 문단들을 다시 더 작은 단위인 문장들로 분리해서 정렬한다. 장의 문단으로의 분리와 문단의 문장으로의 분리, 그리고 그들 사이의 대응관계 표시는 자동화가 불가능한 것이 우리 능력의 현재의 한계이다.[2] 이 과정들에서는 컴퓨터 소프트웨어의 도움을 받는 수작업이 중요한 역할을 한다. 문장 정렬이 늘 단순하지만은 않음은 손현정 (2007)의 연구에서 지적된 바 있다.

결국 문장 단위의 정렬체들이 확보되면 그 다음 단계들은 전적으로 컴퓨터 프로그램이 일을 하는 자동화된 소단위 대역쌍 추출이다. 이 작업의 전체적인 윤곽은 수작업으로 정렬되는 대단위 정렬체들로부터 자동으로 소단위 정렬체들을 얻어 내는 것이다. 우리는 소설 전체로부터 스물 아홉 개의 장을 얻고 그것들을 562개의 문단으로 나눈 후에 결국 오천삼백오십삼 개의 문장들로 분리했다. 이들 문장들의 길이는 낱말 수로 평균 15.3개 (중국어 부문), 17.1개 (한국어 부문)이다. 짧은 문장은 낱말 두 개로 이루어져 있고 긴 문장은 서른 개를 넘는 낱말들로 이루어져 있다. 일부 긴 문장들에 대해서는 추가적인 분리가 행해질 수도 있을 것이지만, 언젠가는 이 작업이 끝 나야 하기 때문에, 그리고 원칙적으로는 정렬체의 수가 전체적으로 중요하지 단일한 텍스트 안에서의 정렬체의 수는 중요하지 않기 때문에 위에서 언급한 오천삼백여개의 문장들을 분리한 상태에서 분리 작업은 더 이상 추구하지 않기로 결정했다.

분리되어 대응관계가 확립된 문장들은 XML 파일 안에 담긴다. 이 파일의 전체적인 구조는 다음과 같다.

---

1) 한국어 번역본은 박용만에 의해 번역되고 (주)푸른숲이 2002년에 발행한 것이다.

2) Brown, Lai and Mercer (1991)는 통계적인 방법으로 문장정렬체들을 얻어 내는 알고리듬을 소개한다. 99.1%의 정확도를 갖는 이 알고리듬은 은닉 마르코프 모델을 이용하는데, 저자는 이 알고리듬의 구현을 시도하지 않았다.

(1)

```
<paralleltext id='许三观卖血记'>
  <chapter id = '1'>
    .
    .
    .
  </chapter>
  .
  .
  .
  <chapter id = '29'>
    .
    .
    .
  </chapter>
</paralleltext>
```

각각의 요소(element)는 더 작은 요소들로 이루어져서 결국 가장 작은 요소는 문자열들을 담는다. 요소의 표시는 여는 태그(opening tag)와 닫는 태그로 이루어진다. 여는 태그는 각 진 괄호쌍과 요소 이름으로 이루어지며 그 괄호쌍 안에 속성과 속성값의 쌍들이 등장할 수 있다. 닫는 태그는 각 진 괄호쌍안에 사선(slash) 하나를 앞에 둔 요소 이름의 꼴을 갖는다.

<chapter> 요소는 한 개 이상의 <paragraph> 요소들로 이루어지고 후자의 이 요소는 다시 한 개 이상의 <s> 요소로 이루어진다. 즉, 장은 문단들로 이루어지고, 문단 속에는 문장들이 있다는 관찰이 이 XML 파일의 내용구조에 반영된 것이다.

중국어 원문과 한국어 번역문의 대응 관계는 몇 가지 다른 방식으로 표현될 수 있겠지만, 우리는 가장 단순한 방식으로 이것을 표현하기로 했다. 즉, <paragraph> 요소는 짝수개의 문장들로 이루어지며 2k + 1번째 문장은 중국어 원문이고 그 직후의 문장은 한국어 대역문이라는 전제를 설정한 것이다. 그래서 전형적인 <paragraph> 요소의 구성은 다음과 같다.

(2)

```
<paragraph>
  <s lang='chinese'>
    ...
  </s>
  <s lang='kor'>
    ...
  </s>
  .
```

```
  .
  .
  <s lang='chinese'>
    ...
  </s>
  <s lang='kor'>
    ...
  </s>
</paragraph>
```

<s> 요소에 lang이라는 속성을 갖게 만듦으로써 홀수(2k + 1) 번째 <s> 요소가 중국어 문장이고 그 직후의 짝수(2k + 2) 번째 요소가 대역 한국어 문장이라는 전제가 충족되는지를 응용 프로그램이 검사할 수 있게 했다.

이렇게 만들어진 병렬 코퍼스의 부호화 규약은 utf-8이고 이 포맷으로 이 파일의 크기는 1149640바이트이다. 이 파일 안의 <paragraph> 요소의 예를 하나 들자면 다음과 같다.

(3)

```
<paragraph>
  <s lang='chinese'>“不是方铁匠, ”许三观说,
  </s>
  <s lang= 'kor'>“아니야, 방 철장이 아니야.
  </s>
  <s lang='chinese'>是何小勇, 为什么是何小勇？何小勇瞒着我让你们妈怀上了一乐,
  </s>
  <s lang= 'kor'>하소용이다. 왜 하소용이냐? 하소용이가 날 욕보이고 너희 엄마를 임신시켜 일락이를 낳게 한데다,
  </s>
  <s lang='chinese'>一乐又把方铁匠儿子的脑袋砸破了,
  </s>
  <s lang= 'kor'>일락이가 방 철장의 아들 대갈통을 박살내 버렸으니…….
  </s>
  <s lang='chinese'>你们说是不是何小勇把我们害的？”
  </s>
  <s lang= 'kor'>하소용이 우리 집을 이렇게 만든 거냐 아니냐?”
  </s>
</paragraph>
```

2.2. 소규모 코퍼스의 효율을 높이기 위한 낱말 분리

이렇게 문장 단위의 정렬체들이 글 전체에 걸쳐서 확립되면 그 다음에는 자동으로 문장보다 더 작은 단위의 정렬체들을 얻어내기 위한 제이단계 준비 작업이 뒤따른다. 그것은 문장의

단위들을 분석 또는 해석하는 일이다. 중국어 문장은 낱말들을 공란자들로 분리된 상태로 담지 않는다는 특징이 있다. 한국어식의 또는 익숙한 인구어식의 띄어쓰기가 중국어 표기의 관례가 아닌 것이다. 이런 문장들을 낱말별로 띄어쓰기를 해 주는 일이 필요하다.

한국어의 경우에는 각각의 문장이 소위 어절 단위로 띄어쓰기를 하지만, 이 어절들은 다양한 종류의 중의성을 드러낸다. 동일한 글자들의 연속체가 의미와 문법의 측면에서 상이한 요소들의 결합체인 예들이 많다는 것이다.[3)]

"쓰고"라는 어절의 의미는 적어도 세 가지로 확연히 구분된다. "소설을 쓰고"에서 그 중 하나가, "돈을 쓰고"에서 또 다른 하나가, 그리고 "모자를 쓰고"에서 제삼의 의미가 드러난다. "접시다"는 불규칙적인 다리의 움직임을 제의하는 청유문일 수도 있고, 특정한 종류의 그릇으로서의 정체를 서술하는 서술문일 수 있다. 이러한 중의성이 남아 있는 상태로는 문장의 하위단위들 사이의 자동정렬은 거의 불가능하다.[4)] 정렬체의 수가 수만에 이르지 못하는 소규모 코퍼스의 경우에는 전단계 정렬체들에서의 중의성 해소가 성공적인 대역표현쌍 추출의 필수 요건이다.

우리는 중국어 문장의 낱말 분리 (segmentation) 작업을 Stanford 대학교의 자연언어처리 연구실이 개발해서 제공하는 chinesesegmenter-2006-05-11이라는 프로그램군에 의존한다. 이 시스템에 관한 자세한 사항은 Tseng 등 (2005)에 기술되어 있다. 그리고 한국어 어절 해석에는 저자가 개발해서 제공하는 KWGInterpreter라는 프로그램을 사용한다. No (2007)이 이 시스템을 기술한다. 전자는 띄어쓰기가 없는 중국어 문장을 받아서 낱말 분리가 끼여 있는 공란자들로 표시되는 문장을 내어 놓는 프로그램이다. 후자는 한국어 문장을 입력물로 받아서 각 어절이 무슨 어휘소들로 이루어져 있는지를 밝혀 주는 프로그램이다.

이 두 프로그램으로 위 (3)과 같은 형식의 정렬체들을 포함하는 파일을 처리해서, 상당한 정도로 중의성이 해소된 정렬체들의 쌍들을 얻는다. (3)의 세 번째 문장쌍은 이제 다음과 같은 꼴을 띤다.

(4)

```
<s lang='chinese'>一乐 又 把 方 铁匠 儿子 的 脑袋 砸破 了 ,
</s>
<s lang='kor'>          {일락이 (ProperNoun NA)} {가 (PN GRAMMAR)}
        {방 (ProperNoun Pang)}
        {철장 (CountNoun smith)} {의 (PG GRAMMAR)}
        {아들 (CountNoun son)}
        {대갈 (CountNoun head)} {통 (CountNoun container)} {을 (PA GRAMMAR)}
```

3) 중의성 현상은 물론 중국어 문장들에서도 다반사로 나타난다. 중의성은 모든 자연언어의 대다수의 문장들에 나타나는 흔한 의미현상이다.

4) "불가능"은 과장된 표현이다. 정렬체의 수가 아주 많다면 모든 중의성 현상은 다의성 현상으로 대우 받아서 소단위 정렬체들 안에 이 다의성이 모두 반영될 수 있다. 그러나 대역표현쌍의 자동 추출은 유한한 기억공간을 갖는 컴퓨터에 의해 수행되므로 이용할 수 있는 정렬체의 수에 한계가 있음에 유념해야 한다.

{박살 (MassNoun NA)} {내 (cause_to_move_out)[GovernedForm]}
{버리 (GRAMMAR)[Past AdverbialClauseForm]} {… (FPunct ldot)} {… (FPunct ldot)} {. (FPunct period)}
</s>

이렇게 중국어 문장들은 모두 낱말로 분리된 형태로, 한국어 문장들은 모두 어휘소들의 연쇄체 형태로 표시된 파일이 우리가 구축하는 병렬 코퍼스의 최종 형태다.

2.3. 코퍼스의 규모

낱말들로 분리된 문장들이 문단을 이루고 문단들이 장을 이루며 장들이 작품을 이루는 것이 이 코퍼스의 계층구조이다. 이 코퍼스의 규모에 대한 개략적 정보는 위에 제시되었다. 낱말 수로 따졌을 때에 이 코퍼스가 얼마나 큰가 하는 의문에는 이제야 비로소 답할 수 있게 되었다.

이 코퍼스에 담긴 중국어 낱말 수는 82,071개이다. 이들 낱말들 중 다수는 두 번 이상 등장한다. 이 8만2천여 개의 낱말은 실은 낱말 ``토큰''들이다. 등장사례의 수가 아니라 추상적 존재로서의 낱말 수를 따지면 물론 8만여보다 훨씬 적은 6,670이다. 즉 이 수만큼의 상이한 낱말들이 한 번 이상씩 이 코퍼스에 등장한다는 것이다. 이 코퍼스에 들어 있는 중국어 낱말들 중에서 3,886개는 작품 전체를 통틀어 단 한번밖에 등장하지 않는다. 코퍼스 언어학의 용어 ``헤이팩스 레고메나''가 이들을 가리키는 표현이다.

코퍼스 언어학에서 낱말 두 개의 연속체들을 고려할 경우가 많다. 이것은 더 일반적으로 낱말 n개의 연쇄체들을 고려하는 경우들의 한 특수예라 할 수 있는데, 후자의 경우에 우리는 n-그램들을 다룬다고 하고 전자의 경우에 바이그램(bigram)들을 다룬다고 한다.

이 코퍼스에 들어있는 문장들에는 낱말 토큰의 수보다 하나 적은 수의 바이그램들이 들어있다. 우리는 의의 있는 바이그램으로 여겨지지 않는 바이그램들은 제거했다. (예를 들어서 ``구두점 -- 낱말'', ''수사 -- 낱말'') 그래서 바이그램 토큰의 수는 중국어의 경우에 76,718이고 한국어의 경우에 86,312이다.

이 코퍼스에 들어 있는 낱말 타잎의 수는 중국어 부문에서 6,670개이고 한국어 부문에서 4,292개이다. 중국어 부문의 헤이팩스 레고메나의 수가 3,886개이고 한국어 부문의 그것이 1,679개인 점을 감안하면, 한국어 번역본이 중국어 원본에 비해 더 단순한 어휘 선택을 한다고 생각할 수 있다.[5)]

	중국어 부문	한국어 부문
토큰 수 모노그램	82,071	91,569

5) 그러나 이 결론은 낱말 분리 시스템이 정확하다고 가정할 경우에만 온당한 결론이다. Tseng 등 (2005)의 chinesesegmenter-2006-05-11에 의한 중국어 원문의 낱말 분리는 일관성을 보이지 않는 부분들을 포함하는 결과를 내어 놓는다. 중국어 부문의 낱말 분리의 결과에 대해서는 최소한의 교정만 행했다.

	바이그램	76,718	86,312
	계	158,789	177,881
타잎 수	모노그램	6,670	4,292
	바이그램	29,382	30,527
	계	36,052	34,819
hapax legomena	모노그램	3,886	1,679
	바이그램	21,579	20,905
	계	25,013	22,553

이 표에서 보다시피, 이 코퍼스는 대략 8만 내지 9만 개의 낱말로 이루어지고 낱말 타잎의 수는 대략 4천 내지 6천여 개이며 한국어 부문이 중국어 부문에 비해 낱말의 평균 빈도가 높다. (중국어 어휘별 12.3회 대비 한국어 어휘별 21.3회)

3. Dice 계수와 더 작은 단위들의 대역관계 자동 추출

대역 문장쌍들을 정렬하는 데까지는 상당한 노력이 든다. 비록 컴퓨터의 도움을 받기는 하지만, 이 작업은 언어 처리 기술의 현상황에 비추어 볼 때 자동화될 수 없는 과정이다.

그러나 문장쌍 내부의 문장보다 더 작은 단위의 정렬체들을 얻어 내는 일은 다행히 대부분 자동화 될 수 있다는 것이 수십 년에 걸친 언어처리 관련 연구의 발견이다. 물론 전제 조건은 다수의 대단위 정렬체들이 이미 확립되어 있다는 것이다. 이 절에서는 비교적 큰 단위의 정렬체들이 주어져 있는 상황에서 대역 낱말쌍들을 포함하는 소단위 대응체들의 쌍들을 자동으로 추출하는 원리와 절차를 기술하고자 한다.

3.1. 대역표현쌍의 분포

어떤 중국어 표현 e_c가 들어 있는 문장들의 번역문들에 한국어 표현들의 집합 {e_k1, e_k2, ..., e_kn}가 들어 있다고 하자. 그런데 이들 한국어 표현들은 e_c가 들어 있지 않은 중국어 문장들의 번역문에도 등장할 가능성이 있다. 전체 집합에서 e_c를 포함하는 문장들의 수를 $f(e_c)$라 하고 e_{ki}를 포함하는 문장들의 수를 $f(e_{ki})$라 하자. 그리고 우리가 2절에서 누차 언급한 "문장쌍"을 잠시 단일한 문장이라고 가정한다면, e_c와 e_{ki}를 함께 포함하는 문장들이 있을 것이다. 이런 문장들의 수를 $f(e_c, e_{ki})$라고 하자. 그러면 이 세 수치를 바탕으로 중국어 낱말 e_c와 한국어 낱말 e_{ki}의 대역표현쌍 형성 가능성이 산출될 수 있다는 것이 우리의 기본 가정이다.

만약 e_c의 뚜렷한 대역표현이 e_{ki}라면, e_c가 등장하는 문장의 번역문에는 e_{ki}가 등장할 것이다. 즉, $f(e_c, e_{ki})$는 $f(e_c)$와 비슷할 것이다. 만약 "许三观"이라는 표현의 대역 표현이 "허삼관"이라면 "许三观"이 등장하는 문장의 수와 "허삼관"이 등장하는 문장의 수가 비슷할 것이다. 그 뿐 아니라, 한 문장과 그 대역문을 동일한 문장으로 간주한다면, "许三观"과 "허삼관"이 함께 등장하는 문장의 수가 그 둘 중의 하나만 등장하는 문장들의 수에 비추어 볼 때, 꽤 많을 것임을 직관으로 알 수 있다. 만약 "승리반점"이 "许三观"의 대역표현이 아니라면, "승리반점"이 등장하지만 "许三观"이 등장하지 않는 문장의 수, 그리고 "许三观"이 등장하지만 "승리반점"이 등장하지 않는 문장의 수가 꽤 많을 것이다.

이 느슨한 서술을 좀 더 엄밀화하자면, 우리는 "许三观"이 등장하는 문장의 수, "허삼관"이 등장하는 문장의 수, "승리반점"이 등장하는 문장의 수와 더불어 "许三观"과 "허삼관"이 다 등장하는 문장의 수와 "许三观"과 "승리반점"이 다 등장하는 문장의 수를 비교함으로써 이 두 한국어 표현이 "许三观"의 대역표현일 확률을 알 수 있다는 얘기다.

3.2. Dice 계수
일찌기 통계학자 Dice에 의해 도입된 이 상관계수가 우리의 대역표현쌍 수립 확률이 된다. Dice (1945), Smadja (1992) 및 Smadja 등 (1996) 참고.

(6) dice(x,y) = 2 * f(x,y) / (f(x) + f(y))

두 표현 x와 y의 Dice계수는 각 표현의 빈도들의 합으로 두 표현이 동시에 등장하는 사례의 빈도를 나눈 것의 두 배이다. 만약 중국어 표현 x가 이 코퍼스의 문장 다섯 개에 등장하고 한국어 표현 y가 문장 세 개에 나타나며, x와 y가 함께 나타나는 문장의 수가 두 개라면, x와 y의 Dice 계수는 2 * 2 / (5 + 3) = 0.5가 된다. 만약 중국어 표현 x가 이 코퍼스의 문장 다섯 개에 들어있고 한국어 표현 y가 문장 한 개에 들어있으며, x와 y를 다 담고 있는 문장의 수가 한 개라면 x와 y의 Dice 계수는 2 * 1 / (5 + 1) = 0.334가 된다. x가 들어 있는 모든 문장에 y가 들어 있고 x와 y가 함께 들어 있는 문장의 수가 x가 들어 있는 문장의 수와 같고 그것이 또 y가 들어있는 문장의 수와 같다면, 즉 f(x) = f(y)= f(x,y)라면 x와 y의 Dice 계수는 1.0이다.

우리는 이제 중국어 표현 x의 대역어가 한국어 표현 y일 확률을 구할 수 있다. 다름 아닌 x와 y의 Dice 계수이다. 이 값이 1.0이면 y는 x의 대역표현일 확률이 1.0이다. 이 값이 0.003이면 x와 y는 그야말로 우연히 동일한 문장에 들어간 표현들일 것이다. 즉, x와 y의 Dice 계수가 곧 대역표현쌍을 이룰 확률이다.

3.3. 추출 절차
통사구조와 기본 어휘의 목록이 완전히 다른 중국어와 한국어 사이의 표현들간의 대응관계는 다음과 같다고 추정할 수 있다. 첫째, 중국어의 낱말 하나가 한국어의 낱말 하나에 깨끗하게 대응하는 경우의 수는 아주 많지는 않을 것이다. 둘째, 출발점 언어의 어떤 낱말이든 목표점 언어의 낱말연쇄체에 대응될 수 있을 것이다. 출발점 언어의 낱말 하나와 대역관계에 놓이는 목표점 언어의 낱말 연쇄체의 길이는 원칙적으로는 꽤 클 수도 있을 것이지만, 대역확률을 계산하는 알고리듬의 서술을 단순화하기 위해 이 길이는 2로 제한하고자 한다.

중국어 낱말 V^c_i는, 따라서, V^k_j라는 한국어 낱말 하나로 그 의미가 깔끔하게 표현될 가능성이 있는가 하면, V^k_j와 V^k_k의 낱말 두 개의 연쇄체에 의해서밖에는 그 의미가 충분히 표현될 수 없을 가능성도 있다. 이와 반대 방향에서 고려하면, 한국어 낱말 V^k_h가 단일한 중국어 낱말 V^c_j에 의해 번역될 가능성과 중국어 낱말 두 개의 연쇄체 V^c_j V^c_k에 의해서밖에는 제대로 번역될 수 없을 가능성도 있다.

이 두 경우와 함께 중국어 낱말 두 개의 연쇄체 V^c_h V^c_i의 각 원소 낱말은 한국어 낱말 어느 것에도 단독으로 대응하지 않지만, 한국어 낱말 두 개의 연쇄체 V^k_k V^k_l에는 잘 대응하는 경우도 상상할 수 있다.

이런 이해하에서, 낱말들로 분리된 텍스트를 읽고 두 언어의 낱말 타잎들 모두와 바이그램 타잎들 모두를 얻어 낸다. 위 코퍼스에서 얻어 낸 바이그램들 중에는 구둣점과 같은 비언어적 표현들을 갖는 것들을 제외시키고, 또한 어떤 구성성분의 오른 쪽 끝에 놓이지 않는 것들을 제외시킨다. 의미합성성의 원리가 적용될 수밖에 없다고 판단되는 구성들의 내부의 바이그램들을 배제하려는 의도에서다. 이 과정에서 떨려 나가는 바이그램들의 예로는 "재빨리 일락", "를 모두", "떠나고 언제"등이 있다.

다음 단계에는 각 언어표현이 각 문장에 등장하는지 등장하지 않는지의 정보를 담는 표를 구축한다. 이 표는 중국어 표현의 수와 한국어 표현의 수를 단의 수로 갖고, 코퍼스 안의 최소 정렬체인 문장쌍의 수를 열의 수로 갖기 때문에 칸의 수가 아주 많을 수 있고, 따라서 컴퓨터의 주기억 용량을 많이 소요하는 주된 자료구조이다. 프로그램 구현시에는 이 표는 이차원배열대 (two-dimensional array)로서 기본 유형이 boolean 유형이다. 이 배열대의 구성을 도시하면 아래와 같다.

(7) 표현별 출현빈도 산출을 위한 배열대

	중국어표현1	중국어표현2	...	중국어표현m	한국어표현1	한국어표현2	...	한국어표현n
문장1								
문장2								
.								
.								
.								
문장l								

이 표가 담는 칸의 수는 (중국어 표현의 수 + 한국어 표현의 수) * 문장쌍의 수이다. 위 코퍼스에 대해서는 이 값이 (36,052 + 34,819) * 5353 = 약 3억8천만이다.

표 (7)이 옳게 만들어진 후에는 각 표현별 코퍼스내에서의 빈도가 쉽게 획득될 수 있다. 임의의 표현 x에 대해 이 표현의 상대적 위치를 나타내는 정수 d를 구할 수 있다는 것을 가정하자. 이 정수 d를 표현 x의 색인치(index)라 부른다. 각 표현의 색인치는 모든 표현들을 어떤 순서로 정돈함으로써 부여된다. 표현의 표기형을 기재하는 글자들의 Unicode 값에 의한 순서부여가 가장 단순한 정돈의 예가 될 것이다.

표현 x의 색인치가 d라고 알려지고 문장의 총계 totalspairs이 알려진 마당에 x의 이 코퍼스에서의 빈도는 다음 루프에 의해 산출된다.

(8)

```
int f = 0;
for (int i=0; i<totalspairs; i++)
{
  if (freq[d][i])
    f++;
}
```

모든 고려대상 표현에 대해 코퍼스상의 빈도가 이 방식으로 획득될 수 있다. Dice 계수를 산출하자면, 두 표현 x와 y의 각각의 빈도 외에 x와 y의 공동출현의 빈도를 얻어야 한다. 이것의 산출식은 (8)의 루프 안의 조건 부분을 'x가 이 문장에 등장하면'으로부터 'x가 이 문장에 등장하고 y가 이 문장에 등장하면'으로 바꾼 것과 같다. 즉,

(9)

```
int f = 0;
for (int i=0; i<totalspairs; i++)
{
  if (freq[d][i] && freq[e][i])
    f++;
}
```

물론, 여기에서 e는 표현 y의 색인치이다.

이제 f(x), f(y)와 f(x,y)를 다 구했으므로, dice(x,y)를 구하는 일은 간단한 덧셈 한 번과 곱셈 한 번, 그리고 나눗셈 한 번만 해 주는 일로 환원된다.

4. 결과 및 논의

이 과정을 통해 <중국어 표현, 한국어 표현, 계수>의 삼개조를 약 12억5천만 개 얻는데, 이들 중 대다수는 계수가 0.0이다. 계수가 0을 초과하는 삼개조의 수는 약 198만개인데 이 중에서 다수는 계수의 값이 매우 낮은 것들이다. 잠재적 대역표현쌍의 Dice 계수가 낮으면 이 쌍의 원소들이 대역관계에 놓일 확률이 낮다는 뜻임을 앞 절에서 확인하였다. 우리는 이 계수의 임계치를 몇 가지로 설정해서 각 임계치별로 정확도를 측정했다.

정확도는 복합적인 개념이다. 추출하고 싶어하는 대역표현쌍들을 얼마나 많이 추출하게 해 주는가 하는 물음과 추출되지 않아야 할 비대역쌍들이 얼마나 많이 추출되어 나왔는가 하는 물음에 함께 답해야 비로소 정확도를 계산할 수 있다. 최고로 정확한 시스템이라면 추출하고 싶어하는 대역쌍들을 100% 추출하면서 추출되지 않아야 할 쌍들은 하나도 추출하지 않는 시스템이다.

정확도의 이 두 측면은 언어처리 분야에서 각각 precision과 recall로 알려져 있다.

(10) recall 제대로 추출된 쌍들의 수 / 추출되어야 할 쌍들의 총수

precision 제대로 추출된 쌍들의 수 / 시스템에 의해 추출되는 쌍들의 총수

recall이 높다는 것은 추출이 실패한 경우의 수가 적다는 것을 뜻한다. 따라서 recall은 "수용정확도"라고 옮길 만하다. (이 경우에 precision은 "배척정확도"라고 할 만하다.) 수용정확도는 1.0인데 배척정확도는 0.1인 체계의 예를 든다면 옳은 대역표현쌍의 수가 열개인 상황에서 이 옳은 대역표현쌍 열 개를 다 추출하고, (추출한 쌍들의 총수가 백개인 상황에서) 옳지 않은 쌍을 아흔 개 더 추출하는 체계가 될 것이다.

대역표현쌍 추출시스템의 출력물은 그 시스템이 어떤 임계치를 이용하느냐에 따라 다른 수용정확도와 배척정확도를 보인다. Dice 계수의 임계치를 낮게 책정할수록 수용정확도가 높아지고 배척정확도는 낮아지는 반면, 임계치가 높아질수록 수용정확도는 낮아지고 배척정확도가 높아진다.

대역표현쌍 추출의 정확도에 관여하는 또 하나의 요인은 각 쌍의 원소들의 코퍼스내의 빈도이다. 중국어 표현 i의 빈도 f(i)와 이것의 잠재적 대역 표현인 한국어 표현 j의 빈도 f(j)가 모두 작은 수이면 비록 f(i,j)를 f(i) + f(j)로 나눈 값이 크더라도 i와 j 사이에 대역관계가 성립하지 않을 수 있다. 그 이유는 i와 j가 우연히 동일한 문장쌍 안에 들어 있지만 실제로 i의 대역 표현으로서 적절한 것은 j^prime이지 j가 아닌 경우가 있다는 것이다.

빈도가 낮은 표현들의 쌍들을 포함 시키는 경우에는 수용정확도를 높이는 반면에 배척정확도를 낮추는 결과가 생긴다. 만약 두 표현의 합계빈도가 k 미만인 쌍들을 모두 고려대상에서 제외해 버리면, 그러지 않는 경우에 비해 수용정확도가 떨어지는 반면에 배척 정확도는 올라 간다. k의 값이 커질수록 배척정확도는 1.0에 가까워진다. 그렇지만, 점점 더 많은 옳은 대역표현쌍이 포착되지 못하고 넘어갈 것이다.

두 표현의 빈도의 합이 3이고 Dice 계수의 값이 0.667인 쌍 마흔 개의 예를 [표 1]에 제시한다. 이들 중에서 옳은 대역쌍은 아홉 개로서 배척정확도 0.2의 낮은 성적을 보인다.

[표 1]

옳음 빈도(중) 표현(중) 빈도(한) 표현(한)

2 瘦得 1 나무 CountNoun tree_wood@처럼 PO like
2 瘦得 1 바람 MassNoun wind@좀 Adverb a_bit
2 瘦得 1 빼빼 Adverb NA
O 2 瘦得 1 빼빼 Adverb NA@마르 get_dry
2 瘦得 1 세 be_strong@불 blow
2 瘦得 1 자네 Pronoun you@피골 MassNoun skin_and_bones
2 瘦得 1 좀 Adverb a_bit@세 be_strong
2 瘦得 1 처럼 PO like@빼빼 Adverb NA
2 瘦得 1 하 GRAMMARN@바람 MassNoun wind
1 瘦得@皮包 2 상접 ProcessNoun NA
1 瘦得@皮包 2 상접 ProcessNoun NA@하 GRAMMARN

1 瘦得@皮包 2 피골 MassNoun skin_and_bones
O 1 瘦得@皮包 2 피골 MassNoun skin_and_bones@상접 ProcessNoun NA
1 瘦胳 2 촌 Measurer unit_of_geneological_distance@그 Pronoun that_thing_person
1 瘦胳@膊抬 2 촌 Measurer unit_of_geneological_distance@그 Pronoun that_thing_person
1 瘦都 2 탓 MassNoun blame@이 COPULA
1 瘦都@已经 2 탓 MassNoun blame@이 COPULA
1 瘫在 2 되 GRAMMARN@사람 CountNoun person
O 1 瘫在 2 마비 ProcessNoun NA
1 瘫在 2 마비 ProcessNoun NA@되 GRAMMARN
1 瘫在@了 2 되 GRAMMARN@사람 CountNoun person
1 瘫在@了 2 마비 ProcessNoun NA
O 1 瘫在@了 2 마비 ProcessNoun NA@되 GRAMMARN
1 瘫痪 2 걸리 be_hung@같 seem
O 1 瘫痪@了 2 걸리 be_hung@같 seem
2 白@了 1 넘기 hand_over@나이 MassNoun age
2 白@了 1 때 VAdjunctNoun time@이 ProperNoun Lee
O 2 白@了 1 백발 MassNoun NA@성성 StateNoun not_sparse
2 白@了 1 벌써 Adverb NA@예순 Numeral sixty
2 白@了 1 성성 StateNoun not_sparse
2 白@了 1 성성 StateNoun not_sparse@하 GRAMMARS
2 白@了 1 얼굴 CountNoun face@하얗 be_white
2 白@了 1 예순 Numeral sixty@훨씬 Adverb NA
2 白@了 1 이 COPULA@백발 MassNoun NA
O 2 白@了 1 질리 get_pale
2 白@了 1 질리 get_pale@허 ProperNoun Huh
2 白@了 1 하 GRAMMARS@다 PO at_onto
O 2 白@了 1 하얗 be_white@질리 get_pale
2 白@了 1 혈두 CountNoun blood_technologist@벌써 Adverb NA
2 白@了 1 훨씬 Adverb NA@넘기 hand_over

반면 두 표현의 빈도의 합이 6이고 Dice 계수의 값이 위와 동일한 쌍들의 예를 [표 2]에 제시하는데, 이들 마흔 개 중에서 옳은 대역쌍은 열네 개로서 배척정확도 0.35의 성적을 보인다.

[표 2]
옳음 빈도(중) 표현(중) 빈도(한) 표현(한)
O 3 浸到 3 속 MassNoun inner_part@들 get_in
3 浸到 3 피 CountNoun blood@속 MassNoun inner_part
3 涌泉 3 솟 soar@갚 pay_back

O 3 涌泉@相报 3 솟 soar@갚 pay_back
2 深蓝 4 바탕 MassNoun NA
2 深蓝@碎花 4 바탕 MassNoun NA
2 清炖 4 붕어 CountNoun NA
2 清炖@鲫鱼 4 붕어 CountNoun NA
3 滴 3 몇 NumeralAdjective several@방울 Classifier unit_of_liquid
O 4 漆匠 2 칠장이 CountNoun one_who_paints_furniture
4 漆匠@会 2 칠장이 CountNoun one_who_paints_furniture
O 3 灯笼 3 빨갛 be_red@달 get_hot
O 3 灾年 3 재난 MassNoun NA
4 灾荒 2 가뭄 MassNoun NA@지나 go_past
4 点@力气 2 고프 be_hungry@힘 MassNoun strength
3 烂肝 3 시커멓 be_black
O 3 烟囱@上 3 굴뚝 CountNoun chimney@앉 sit
2 照着@自己 4 보 look_at@다시 Adverb NA
3 熟悉 3 곳 CountNoun place@사람 CountNoun person
O 3 爹妈 3 어머니 CountNoun mother@아버지 CountNoun father
2 爹长 4 빼 disengage@. FPunct period
2 爹长@得 4 빼 disengage@. FPunct period
3 牙 3 갈 grind
4 牙根 2 아직 Adverb NA@시리 be_chilly
O 3 猪肝@和 3 간 CountNoun liver@과 PC with
3 猪肝@和 3 과 PC with@황주 CountNoun a_kind_of_beverage
3 猪肝@喝 3 EMPTYSTEM COPULA@돼지 CountNoun NA
4 玉米 2 머얼건 Unknown NX
4 玉米 2 머얼건 Unknown NX@죽 CountNoun grain_soup
2 王八蛋@的 4 개 CountNoun dog@자식 CountNoun child
O 2 现眼@了 4 망신 ProcessNoun NA
2 瓶@黄酒 4 한 NumeralAdjective one@병 CountNoun bottle
2 用@在刀刃 4 쇠 CountNoun steel
O 4 电影院 2 극장 CountNoun theater@, COMMA comma
4 电影院 2 성내 MassNoun city_area@소학교 CountNoun elementary_school
O 2 男孩@的 4 사내 CountNoun man@아이 CountNoun child
O 2 疼@吗 4 아프 be_sick_be_in_pain@? FPunct question_mark
O 2 疼@啊 4 아이야 Interjection NA@! FPunct exclamation
3 疼死 3 나 Pronoun I@아프 be_sick_be_in_pain
O 2 疼死@啦 4 아프 be_sick_be_in_pain@죽 die

두 표현의 빈도의 합이 9이면서 Dice 계수가 0.667인 쌍들 열 한 개를 뽑아서 검사하니 그에서 다섯 개가 옳은 쌍이다. 이는 배척 정확도 0.45를 뜻한다.

[표 3]

옳음	빈도(중)	표현(중)	빈도(한)	표현(한)
?	6	条@被子	3	이불 CountNoun NA@넉 NumeralAdjective four
X	3	次@血要	6	석 Numeral three
O	6	每@个	3	매 Adjective every
?	6	每@个	3	매 Adjective every@달 GridNoun month
O	4	没有动	5	꼼짝 Adverb NA@않 GRAMMARN
O	3	油条@了	6	꽈배기 CountNoun NA@튀기 fry
O	3	滴	6	방울 Classifier unit_of_liquid
O	3	烟囱@上	6	굴뚝 CountNoun chimney
X	5	生日	4	내 Pronoun I@생일 MassNoun birthday
?	5	生日	4	생일 MassNoun birthday@이 COPULA
X	4	用@钱	5	돈 CountNoun money@필요 StateNoun requisite

표현들의 빈도 수준을 3, 6, 9로 변화시키면서 동일한 상관계수를 갖는 쌍들의 비율을 산출해 보면 이렇게 빈도 수준이 올라 감에 따라서 배척정확도 역시 상승함을 알 수 있다.

[표 4] 상관계수 0.667에서의 표현들의 빈도합계에 따른 배척정확도

빈도 합계	정확도
3	0.2
6	0.35
9	0.45

빈도가 낮은 쌍들을 추출 대상에서 제외하면 추출의 배척정확도가 높아진다. 그러나 당연히 추출의 수용정확도는 낮아진다. 예를 들어 빈도 임계치를 4로 책정하면 합계빈도가 3 이하인 쌍들의 약 22%가 옳은데도 빠져 버리는 것이다.

두 표현의 빈도 합계가 3 이하인 쌍들을 모두 버리고 합계가 4 이상인 쌍들 중에서 상대적으로 저빈도로 등장하는 쌍들에 대해서는 상대적으로 높은 상관계수 임계치를 적용하고 고빈도로 등장하는 쌍들에 대해서는 상대적으로 낮은 상관계수 임계치를 적용하는 전략이 수용정확도와 배척정확도 둘 다를 합리적인 수준으로 유지할 수 있음을 보았다. 몇 차례의 임계치 수정을 거쳐 다음과 같은 책정이 가장 낫다는 결론을 내렸다.

[표 5] 빈도에 민감한 합리적인 Dice 계수 임계치

빈도	계수
4 -- 6	0.66
7 -- 11	0.51
12 -- 15	0.39
16 -- 19	0.34
20 -- 23	0.28
24 이상	0.24

이 임계치들을 적용했을 때에 추출되는 삼개조는 모두 구천오백여 개로 추정된다. (가능한 삼개조들의 10분의 1을 실제로 추출했는데 그 개수가 955이었다.) 중국어와 한국어에 모두 능통한 양어 전문가의 검사에 의하면 955개의 대역 표현쌍 후보 중에서 625개는 그릇된 것들이었다. 이것은 배척 정확도가 0.313임을 뜻한다.

이 코퍼스에 존재하는 대역쌍의 수는 위 2.3절의 통계치를 바탕으로 할 때에, 중국어 낱말 타잎 수 + 헤이팩스 레고메나 아닌 중국어 낱말 바이그램의 수의 10분의 1 = 6,670 + 780 = 7,450개로 추정된다. 이 추정치가 옳다면 앞 절들에서 기술된 이 대역쌍 추출시스템의 수용정확도는 약 0.442이다.

5. 결론

대역표현쌍을 자동으로 추출하는 과정을 기술하였다. 실제 대역표현 추출의 결과물은 두 가지 측면에서 불완전함을 보았다. 수용정확도를 높이자면 빈도 임계치나 상관계수 임계치를 낮춰야 한다. 배척정확도를 높이자면 두 임계치를 높여야 한다. 두 임계치 중 하나가 극소값을 갖게 되면 배척 정확도가 매우 낮아져서 원하는 쌍들만 선별하는 데에 너무 큰 노력이 든다. 따라서 배척 정확도를 높이자면 두 임계치 모두를 높여야 한다. 그러나 두 임계치 중 하나가 극대값을 갖게 되면 수용 정확도가 0이 된다. (예를 들어 Dice 계수가 1.0을 초과한다는 조건을 주거나 빈도 수 합계가 일백만을 초과한다는 조건을 주면 그렇다.)

상관계수 임계치는 코퍼스의 크기에 무관하게 두 표현이 대역표현일 가능성을 규제한다. 그러나 빈도수 합계는 코퍼스의 크기와 밀접히 관련을 맺는다. 빈도수 합계가 높은 쌍일수록 더 낮은 상관계수를 가짐에도 불구하고 유효한 쌍일 확률이 더 높다.

따라서 빈도 임계치를 높이 책정하고도 수용 정확도가 이 시스템에 견줄 만하려면 코퍼스의 규모를 늘여야 한다.

옳은 대역 표현쌍의 상관계수가 1.0이 아닌 경우가 왜 생기는가? 그 첫 번째 이유는 유의어의 존재다. 한국어의 "가장"과 "제일"은 모두 중국어 낱말 "最"의 훌륭한 대역 표현이다. 동일한 작가가 유의어 집단에서 하나만을 사용해서 글을 쓰지 않으며, 이 경향은 이 코퍼스의 한국어 부문의 저자, 즉 원본 소설의 번역가한테서도 나타난다. 둘째 이유는 중국어 낱말 분리자의 오류나 한국어 어절 해석 시스템의 오류이다. "躺"라는 글자가 있는 모든 중국어 문장에 "눕다"의 어떤 꼴이 쓰인 한국어 문장이 대역문으로 정렬되어 있다고 해도, 이 글자가 일관되게 낱말로 분리되어 있지 않다면 "躺"과 "눕"의 상관계수는 1.0에 미치지 못한다. 우리가 사용한 중국어 낱말 분리기는 다음과 같은 결과를 내어 놓는다.

(11) 在 病房 里 不声 不 响躺 了 二十多 个 小时

여기에서 "响躺"이라는 낱말은 없다고 보아서 "响 躺"으로 분리된 결과가 나왔다면 "躺"과 "눕"으로 이루어진 대역표현쌍의 상관계수가 더 높게 나타났을 것이다.

옳은 대역표현쌍의 상관계수를 떨어뜨리는 세 번째 요인은 옳지 않은 정렬이다. 코퍼스에 대역문으로 들어 있는 중국어 문장과 한국어 문장의 쌍이 엄밀한 의미에서의 상호대역 관계에 놓이지 않는 경우들이 그것이다. 오정렬은 정렬작업에서의 단순한 실수에 기인할 수 있고 번역오류에서 오기도 한다. 무엇이 옳은 번역이냐를 엄격하게 규정하지 못하는 만큼, 번역의 오류를 찾아 내는 일은 어렵다. 우리 코퍼스의 한국어 부문에는 중국어 부문에 존재하는 적지 않은 문장들이 단순 삭제되어 있음이 발견되었다. 가치 중립적인 평가를 한다면, 중국어 원본이 갖는 중복성과 지나치게 자세한 설명이 간결성을 추구하는 번역가의 문체적 특징으로 인해 사라졌다. 이 것은 Brown, Lai and Mercer (1991)과 손현정 (2007)의 여러 대응 유형 중에서 1대 0 유형의 대응이다.

이렇게 규모가 큰 병렬 코퍼스를 구축하는 데에는 많은 노력이 든다. 이 노력 중 상당한 부분은 (1) 문자인식 시스템의 오인식, (2) 중국어 낱말 분리시스템의 오분리, (3) 한국어 레마 해석시스템의 오해석으로 인한 교정의 필요에 경주되었다. 약 일천 시간에 달하는 것으로 보이는 이 작업의 대부분이 각 단계의 결과물에 대한 교정에 바쳐진 만큼, 이 세 개의 도구의 질을 개선함으로써만 소설 하나 분량의 병렬 코퍼스 구축에 드는 시간을 100 일꾼 시간 (man hours)이하로 떨어뜨리는 것이 가능할 것이다.

참고문헌

손현정 2007. 불-한 병렬 말뭉치의 문장 단위 정렬 방식에 대한 연구. 언어연구. 24권 제2호. 경희대학교 언어연구소. 서울.

Dice, Lee R. 1945. Measures of the amount of ecologic association between species. Journal of Ecology, 26.

Brown, Peter F., Jennifer C. Lai and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. Proceedings of the 29th Annual Meeting of Association for Computational Linguistics. Pp. 169--176.

Gale, William A. and Church, Kenneth W. 1991. Identifying word correspondences in parallel texts. In "Proceedings, DARPA Speech and Natural Language Workshop," Morgan Kaufmann, San Mateo, California.

No, Yongkyoon. 2007. KWGInterpreter: A lemmatizing POS tagger for the Korean language. In Proceedings of the 2007 Join Conference of LAK, MLSK and KSLI.

Tseng, Huihsin , Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. 2005. A Conditional Random Field Word Segmenter. In Fourth SIGHAN Workshop on Chinese Language Processing.

Smadja, Frank 1992. How to compile a bilingual collocational lexicon automatically. In "Proceedings, AAAI-92 Workshop on Statistically Based NLP Techniques," American Association for Artificial Intelligence.

Smadja, Frank, Kathleen R. McKeown, and Vasileios Hatzivassiloglou 1996. Translating collocations for Bilingual lexicons: A statistical approach. Computational Linguistics, 22.1.

[부록]

**7 晚 9 늦 be_late 6.250e-01
*4 晚@了 9 늦 be_late 6.154e-01
?.4 晚@了 2 이미 Adverb already@늦 be_late 6.667e-01
***28 晚上 31 밤 MassNoun night 4.407e-01
***28 晚上 12 저녁 MassNoun dinner 3.000e-01
X *7 暖和 8 온기 MassNoun hotness 4.000e-01
X 4 暖瓶 2 과 PC with@세수 ProcessNoun NA 6.667e-01
X 4 暖瓶 2 들 lift@붉 be_red 6.667e-01
X**4 暖瓶 15 병 CountNoun bottle 4.211e-01
X*4 暖瓶 3 병 CountNoun bottle@과 PC with 5.714e-01
X *4 暖瓶 4 보온 MassNoun NA 1.000e+00
*4 暖瓶 4 보온 MassNoun NA@병 CountNoun bottle 1.000e+00
X *4 暖瓶 3 와 PC with@이불 CountNoun NA 5.714e-01
X *4 暖瓶 3 요 CountNoun blanket@와 PC with 5.714e-01
X 2 暖瓶@和 2 과 PC with@세수 ProcessNoun NA 1.000e+00
X 2 暖瓶@和 4 깃발 CountNoun NA 6.667e-01
X 2 暖瓶@和 2 들 lift@붉 be_red 1.000e+00
X 2 暖瓶@和 3 병 CountNoun bottle@과 PC with 8.000e-01
X 2 暖瓶@和 4 보온 MassNoun NA 6.667e-01
X 2 暖瓶@和 4 보온 MassNoun NA@병 CountNoun bottle 6.667e-01
X 2 暖瓶@和 4 붉 be_red@깃발 CountNoun NA 6.667e-01
X 2 暖瓶@和 3 와 PC with@이불 CountNoun NA 8.000e-01
X 2 暖瓶@和 3 요 CountNoun blanket@와 PC with 8.000e-01
X 2 暖瓶@里 4 보온 MassNoun NA 6.667e-01
X 2 暖瓶@里 4 보온 MassNoun NA@병 CountNoun bottle 6.667e-01
X **17替4내Pronoun I@대신 VAdjunctNoun contrary_exchange_substitution_replacement 2.857e-01
***17替12대신 VAdjunctNoun contrary_exchange_substitution_replacement 5.517e-01
*4替@我4내PronounI@대신 VAdjunctNoun contrary_exchange_substitution_replacement 7.500e-01
X **4替@我12대신VAdjunctNoun contrary_exchange_substitution_replacement 3.750e-01
***40 最 9 가장 Adverb the_most 3.265e-01
***40 最 35 제일 VAdjunctNoun the_most 5.867e-01
X ***5 最@大 35 제일 VAdjunctNoun the_most 2.500e-01
*5 最@大 4 제일 VAdjunctNoun the_most@크 be_big 8.889e-01
X *6 最@好 9 가장 Adverb the_most 4.000e-01
X 2 最@怕 2 여자 CountNoun woman@제일 VAdjunctNoun the_most 1.000e+00
X ,2 最@爱 2 사랑 ProcessNoun love 1.000e+00
X 2 最@爱 2 사랑 ProcessNoun love@하 GRAMMARN 1.000e+00
X 2 最@爱 2 제 Pronoun I@제일 VAdjunctNoun the_most 1.000e+00

2 最@爱 2 제일 VAdjunctNoun the_most@사랑 ProcessNoun love 1.000e+00
X *2 最@粗 5 주사 MassNoun injection@바늘 CountNoun needle 5.714e-01
***21 最后 11 마지막 MassNoun NA 6.250e-01
***21 最后 8 마지막 MassNoun NA@으로 PO with_toward_as 4.828e-01
*3 最后@那 11 마지막 MassNoun NA 4.286e-01
X 3 最后@那 2 마지막 MassNoun NA@한 NumeralAdjective one 8.000e-01
***38 月 14 달 GridNoun month 3.077e-01
***38 月 24 달 Measurer month 7.419e-01
X ***38 月 6 석 Numeral three 2.727e-01
X ***38 月 10 한 NumeralAdjective one@달 Measurer month 4.167e-01
X *2 月@了 5 두 NumeralAdjective two@달 Measurer month 5.714e-01
*4 月@以前 4 달 Measurer month@전 VAdjunctNoun time_prior 7.500e-01
X *4 月@以前 3 전 VAdjunctNoun time_prior@오 come 5.714e-01
X 3 月光 2 달 CountNoun the_moon 8.000e-01
3 月光 2 달 CountNoun the_moon@빛 CountNoun light 8.000e-01
2 有@一半 2 절반 MassNoun NA@내 Pronoun I 1.000e+00
X ***27 有@两 25 둘 Numeral two 3.077e-01
X 2 有@二十多 4 여 NM or_more@년 Measurer year 6.667e-01
X *9 有@几 4 몇몇 Numeral several 4.615e-01
X ***9 有@四 33 네 NumeralAdjective four 2.857e-01
X *9 有@四 3 네 NumeralAdjective four@댓 NumeralAdjective four_to_five 5.000e-01
X *9 有@四 3 댓 NumeralAdjective four_to_five 5.000e-01
X 4 有@多 2 얼마 Pronoun how_much@하 GRAMMARS 6.667e-01
***14 有@多少 44 얼마 Pronoun how_much 2.759e-01
X **14 有@多少 6 얼마 Pronoun how_much@마시 drink 4.000e-01
X 3 有@良心 3 사람 CountNoun person@양심 MassNoun NA 6.667e-01
*3 有@良心 10 양심 MassNoun NA 4.615e-01
3 有@良心 3 양심 MassNoun NA@있 exist 6.667e-01
*5 有@道理 8 일리 MassNoun NA 7.692e-01
*5 有@道理 8 일리 MassNoun NA@있 exist 7.692e-01
2 有名 2 유명 StateNoun famous 1.000e+00
X 2 有名 2 유명 StateNoun famous@하 GRAMMARS 1.000e+00
2 有名@的 2 유명 StateNoun famous 1.000e+00
X 2 有名@的 2 유명 StateNoun famous@하 GRAMMARS 1.000e+00
X *9 朋友 5 소용 ProperNoun NA@친구 CountNoun friend 5.714e-01
***9 朋友 18 친구 CountNoun friend 5.185e-01
*9 朋友 3 친구 CountNoun friend@EMPTYSTEM COPULA 5.000e-01
3 朋友@了 3 친구 CountNoun friend@EMPTYSTEM COPULA 6.667e-01
X 3 服装店 3 시계포 CountNoun shop_carrying_clocks_and_watches 1.000e+00
X 3 服装店 2 시계포 CountNoun shop_carrying_clocks_and_watches@, COMMA

comma 8.000e-01
X 3 服装店 2 열 open@옷 CountNoun clothes 8.000e-01
X 3 服装店 3 옷 CountNoun clothes@, COMMA comma 6.667e-01
X 3 服装店 2 옷 CountNoun clothes@가 go 8.000e-01
X 3 服装店 3 정육점 CountNoun butcher's_shop 1.000e+00
X 3 服装店 2 정육점 CountNoun butcher's_shop@, COMMA comma 8.000e-01
X 3 服装店 3 천녕사 ProperNoun NA 6.667e-01
X 3 服装店 2 천녕사 ProperNoun NA@, COMMA comma 8.000e-01
X **11 木板 11 걸 hang 4.545e-01
*11 木板 3 널판지 CountNoun NA 4.286e-01
*11 木板 4 목 CountNoun neck@판 CountNoun board 4.000e-01
*5 木桥 2 목교 CountNoun wooden_bridge 5.714e-01
X *5 木桥 2 목교 CountNoun wooden_bridge@위 MassNoun upper_surface 5.714e-01
X 4 木桥@上 2 목교 CountNoun wooden_bridge 6.667e-01
4 木桥@上 2 목교 CountNoun wooden_bridge@위 MassNoun upper_surface 6.667e-01
2 木桩 2 말뚝 CountNoun NA 1.000e+00
2 木桩 2 말뚝 CountNoun NA@묶 bind 1.000e+00
X 2 木桩@上 2 말뚝 CountNoun NA 1.000e+00
X 2 木桩@上 2 말뚝 CountNoun NA@묶 bind 1.000e+00