제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

영한 기계번역을 위한 영어 복합명사 자동 수집

조재호, 김성동
한성대학교 컴퓨터공학과
e-mail:jjhnautilus@hotmail.com, sdkim@hansung.ac.kr

Automatic Collection of English Composition Nouns for English-Korean Machine Translation

Jae-Ho Cho, Sung-Dong Kim
Dept of Computer Engineering, Hansung University

요 약

영한 기계번역에서 복합명사는 어휘분석, 구문분석을 복잡하게 하고 사전에 의미가 등록되어 있지 않을 경우에는 올바르게 번역하기 어렵다. 또한 복합명사는 계속하여 새로 나타나고 있어, 정확하고 자연스러운 번역을 위해서는 복합명사를 독립적으로 처리하는 모듈이 필요하다. 본 논문에서는 복합명사를 구성하지 못하는 불용어를 파악하고 빈도수를 이용하여 복합명사를 자동으로 수집하는 방법을 제안한다. 문서를 번역하기 전에 복합명사를 파악하면, 복합명사에 대한 정보를 활용하여 어휘분석과 구문분석의 복잡도를 줄이고 복합명사를 포함한 문장을 보다 자연스럽게 번역할 수 있어 영한 기계번역 시스템의 성능 개선에 기여할 것이다.

1. 서론

영한 기계번역에서 복합명사를 포함하는 문장을 번역할 때, 이를 먼저 인식하여 하나의 단위로 분석하고 번역하면 보다 정확하고 자연스러운 번역을 할 수 있다. 현재 복합명사는 일반적으로 숙어로 처리하고 있다. 즉, 사전에 미리 복합명사에 해당하는 단어를 저장한 후 이를 인식, 번역한다. 그러나, 신조어가 계속해서 발생하고 있으며 사전에 미리 등록되지 않은 복합명사들에 대해서는 번역을 제공하지 못하고 복합어를 이루는 영어 단어가 그대로 출력되기도 한다. 따라서 번역 전에 복합명사를 미리 찾아, 하나의 단위로 취급하게 하여 분석하고, 대역어를 지정하여 이를 이용하여 번역을 수행하면 보다 정확한 번역을 생성할 수 있다.

본 논문에서는 문서번역 이전에 문서에서 나타나는 복합명사를 자동으로 수집하는 방법을 제안한다. 불용어를 이용하여 복합명사를 구성할 수 없는 단어를 제외하고, 문서에 나타나는 빈도수를 기준으로 복합명사를 확인하도록 하였다. 즉, 규칙에 의해 불용어를 확인하고 빈도수에 의해 복합명사로 판단하는, 규칙과 통계적인 방법을 혼용한 수집 방법을 제안한다. 논문에서 제안하는 복합명사 수집 방법은 일종의 복합어구 인식 즉 구 묶음 문제에 해당한다. 복합어구의 종류는 명사구, 동사구, 전치사구, 접속사구 등 다양하지만 본 논문에서는 명사구, 즉 복합명사만을 대상으로 하였다. 복합명사 인식 방법은 다른 복합어구의 인식 방법과 다를 뿐만 아니라 실제 문서에서 나타나는 빈도도 다른 복합어구에 비해 많으므로 복합명사의 인식만으로도 영한 기계번역 시스템의 성능 개선에 유효하다고 판단하였기 때문이다.

본 논문은 다음과 같이 구성된다. 2장에서는 복합어구 인식을 위한 기존 연구를 살펴보고 3장에서 복합명사 자동 수집 방법을 설명한다. 4장에서 실험결과를 제시하고 5장에서 제안한 복합명사 자동 수집 방법의 기대효과와 앞으로의 과제를 제시하며 논문을 마무리한다.

2. 관련 연구

영어에서의 복합명사 인식 문제는 명사구 추출(noun phrase extraction) 문제의 부분이라 할 수 있다. 명사구 추출은 문장에 존재하는 모든 명사구를 인식하는 문제인데, 본 논문에서의 복합명사 인식은 고유명사 역할을 하는 여러 단어를 인식하는 문제이다. 따라서, 개체명 인식(named-entity recognition: NER) 문제의 한 부류라고 할 수 있다. 개체명 인식은 자연언어처리의 다양한 분야에서 활용되고 있으며 활발한 연구가 수행되는 분야이다. 은닉 마코프 모델(Hidden Markov Model)[1], 최대 엔트로피모델(maximum entropy 모델)[2], CRF(Conditional Random Fields)[3] 등의 감독 학습(supervised learning) 방법이 개체명 인식을 위해서 가장 많이 사용되고 있다. 그리고, bootstrapping 기법을 적용하는 반감독 학습(semi-supervised learning) 방법에 의한 인식 방법도 다양하게 제시되었다[4]. 또한 군집화(clustering) 같은 비감독 학습(unsupervised learning) 방법에 의한 개체명 인식 방법에 대한 연구도 활발한데, [5]에서는 일반 명사와는 달리 개체명은 몇몇 기사에 동시에 나타나는 현상을 이용하여 드물게 나타나는 개체명을 확인하는 방법을 제시하

였다.

국내에서는 한국어에서의 복합어구 인식에 대한 연구는 활발하게 진행되고 있으나 영어에서의 복합어구 인식에 대한 연구는 그다지 많지 않다. 다만 1990년대 초에 영한 기계번역 시스템 개발에 대한 연구가 많이 수행될 때 관련연구가 있었으며, 대표적으로 [6]에서는 영어에서의 복합어의 유형을 분류하고 사전과 규칙에 의한 복합어 인식 방법을 제안하였다.

본 논문에서의 복합명사 자동 수집은 영한 기계번역의 어휘 및 구문분석을 용이하게 하고 대역어에 의한 번역을 통해 보다 자연스러운 번역문 생성을 위해 필요한 것으로서, 기존의 개체명 인식 방법들에 비해 간단한 방식을 적용하여 적은 노력을 통해 영한 기계번역 시스템의 성능 개선에 기여할 수 있도록 하는 것을 목적으로 한다.

3. 복합명사 자동 수집 방법

복합명사 수집을 위해서 규칙에 의해 불용어를 확인하고 불용어를 제외한 연속적인 단어들을 결합하여 복합명사 후보를 추출한다. 그리고 추출된 복합명사 후보들이 문서에서 나타난 횟수를 기준으로 하여 복합명사로 간주하는 방법을 택하였다. 간단한 방법이지만 자주 나타나는 복합명사를 수집함으로써 영한 기계번역 시스템의 성능 개선에 기여할 수 있다.

그림 1은 문서에서 복합명사를 추출하기 위한 알고리즘을 보여준다. 복합명사 추출 후에 문서에서 나타나는 빈도수가 일정한 기준 이상일 경우 복합명사로 간주한다.

```
Extract_CompositionNoun(Document, n, CNTable)
{
  Read a sentence from Document;
  Convert into word list;
  foreach word_w
  {
    if (word_w != uselessWord)
    {
      new_cWord = combine_with_next_words(n);
      if (new_cWord doesn't exist)
          add new_cWord into CNTable;
      else increment new_cWord's frequency;
    }
  }
}
```

(그림 1) 복합명사 추출 알고리즘.

위의 알고리즘에서 Document는 복합명사를 추출할 문서이며, n은 복합명사의 길이(2이면 두 단어 복합명사 추출, 3이면 세단어 복합명사 추출), CNTable은 알고리즘의 수행결과 생성되는 복합명사 테이블이다. 문장의 단어가 불용어가 아닌 경우에 인자로 주어진 복합명사 길이(n)에 만큼의 단어를 결합한 후 복합명사 테이블에 등록하여 문서에서 나타나는 빈도수(frequency)를 기록한다. 이 빈도수가 후에 복합명사로 간주하는데 이용된다. 불용어의 확인을 위해서 영어 어휘 분석기를 이용하여 다음과 같은 정보를 확인한다: 품사, 하이픈(hyphen) 포함여부, 아포스트로피(apostrophe) 포함여부, ~ly로 끝나는 형용사 여부, ~ing로 끝나는 형용사 여부, 비교급-최상급 형용사 여부, wh-단어 여부, 알파벳 첫 문자 여부 등.

그림 1 알고리즘에서 가장 중요한 부분은 복합명사를 구성할 수 없는 불용어를 확인하는 것이며, 본 논문에서는 불용어 판단을 위해 9가지 기준을 정하였다.

3.1 동사, 부사, 전치사, 관사, 접속사, 대명사

이들 품사의 단어는 불용어로 간주한다. 동사, 부사로 사용되는 단어가 복합명사를 구성할 가능성이 매우 낮기 때문에 이를 불용어로 간주하였다. 전치사, 관사, 접속사의 경우는 기관명이나 고유한 복합명사를 구성할 수 있지만 (Department of States, The Bill and Melinda Gates Foundation, ...) 대부분의 경우에는 복합명사를 구성할 수 없다고 판단하였다. 물론 이들 품사에 대해서는 예외적인 사항을 검사하여 불용어 여부를 판단한다. 대명사는 명사를 대표하는 역할을 하므로 대명사를 포함하여 복합명사를 형성하기 어렵다고 판단하여 불용어로 간주하였다.

3.2 하이픈(-)을 가지는 단어

하이픈이 있는 단어는 불용어가 아니다. 이들 단어는 형용사, 명사, 숫자 등이 하이픈에 의해 붙어있는 형태인데, 전체적으로 명사를 꾸며주는 형용사의 역할을 할 수 있으므로 불용어로 간주하지 않는다. 예외적으로 4-year, 900-student, 3-to-3 등은 하이픈이 있지만 이들 단어는 고유한 명사를 수식하여 복합명사를 형성하기 보다는 일반명사를 수식하는 역할을 하는 단어로 간주하여 불용어로 취급한다. 즉 숫자가 하이픈에 의해 다른 단어와 결합한 형태의 단어인 경우에는 하이픈을 포함한다 하더라도 불용어로 간주하였다.

3.3 아포스트로피(')가 있는 단어

아포스트로피가 있는 단어는 불용어이다. 이들 단어는 주로 동사와 명사, 형용사가 축약된 형태인데, 물론 소유격도 가능하지만 빈도수가 적으므로 무시하였다. 예를 들어, isn't, there's, today's 등은 복합명사를 형성하지 못한다. Research's, Cray's 등 고유명사가 아포스트로피와 's'와 결합된 단어의 경우에는 복합명사를 형성할 가능성도 있다. 그러나 그 빈도수가 매우 적기 때문에 후에 빈도수를 기준으로 복합명사를 확인하는 단계에서 제외될 가능성이 높아 미리 불용어로 간주하였다.

3.4 ly로 끝나는 형용사, 부사

ly로 끝나는 형용사와 부사가 자주 사용되는데 이러한 단어는 복합명사를 구성하지 못한다. 따라서 품사에 관계없이 ly로 끝나는 단어인가를 확인하여 불용어를 확인한다. 예를 들어, beautifully, stringently, particularly, relatively, formerly, separately, presumably 등이 ly로 끝나는 형용사와 부사인데, 이들을 불용어로 간주함으로써 복합명사 확인을 보다 용이하게 할 수 있다.

3.5 ing로 끝나는 단어

ing로 끝나는 단어는 불용어이다. 이들 단어는 주로 동사의 현재 진행형이거나 형용사, 부사의 품사를 가진다. 동사의 경우에는 3.1에서 이미 불용어로 간주하였다. 형용사의 경우 복합명사를 형성할 수 있는 품사인데, ing로 끝나는 형용사의 경우에는 동사의 현재 진행형처럼 움직이는 동작을 표현하므로 복합명사를 형성하지 못한다고 판단하였다. 예를 들어, aborning, absorbing, accommodating, accompanying 등의 ing로 끝나는 형용사, 부사는 불용어로 간주한다. 이와는 달리 spring, string, ceiling등은 ing로 끝나지만 명사이므로 불용어가 아니기 때문에 이들을 불용어로 간주하지 않는다.

3.6 er, est로 끝나는 비교급, 최상급 형용사

“er”, “est”로 끝나는 형용사 단어는 불용어이다. er, est로 끝나는 형용사는 불용어이다. 이들 단어는 주로 than과 같이 쓰여 “~보다”란 뜻으로 쓰이는데, 비록 형용사지만 주로 명사를 직접 수식하지 않는 서술적 용법으로 사용된다. 예른 들어, former, later, modest, higher 등이 있다. 따라서 복합명사를 형성하지 않는다고 판단하였다. 예외적으로 worker, teacher, keeper, engineer, singer, soldier등은 “er”로 끝나지만 형용사가 아닌 명사이므로 복합명사를 형성할 수 있어 “er”로 끝나는 단어라 할지라도 품사 확인이 필요하다.

3.7 육하원칙 단어: WH-word

육하원칙을 표현하는 단어인 when, where, who, how, what, why 등의 단어는 복합명사를 형성할 수 없는 확실한 불용어이다.

3.8 연속적인 형용사 또는 부사+형용사

연속적으로 나타나는 형용사들이나 부사와 결합된 형용사는 불용어로 간주한다. 예를 들어, very good, much more 등 연속적으로 부사, 형용사가 나타나는 경우 연속적인 단어 모두가 불용어에 해당한다.

3.9 첫 문자가 알파벳이 아닌 단어

첫 문자가 알파벳 문자가 아닌 경우는 숫자로 시작하는 단어이거나 기호의 경우에 해당한다. 숫자로 시작하는 단어는 뒤에 나오는 명사를 수식하는 경우가 대부분이므로 복합명사를 형성하지 않는다고 판단하였다. 그러나 일반 기호의 경우에는 Nerco Oil & Gas Inc., Telephone & Telegraph Co. 등과 같이 복합명사를 형성할 수 있어 전-후 단어를 고려하여 불용어의 여부를 판단하였다.

4. 실험결과

복합명사를 수집하기 위해 Wall Street Journal, IBM 문서, Brown 영역의 문장을 이용하였으며 데이터에 대한 통계는 표 1에 제시하였다.

<표 1> 복합명사 수집을 위한 데이터

	단어 개수	문장 개수
WSJ	1,105,510	53,838
IBM	59,749	4,404
Brown	956,898	50,440

3장에서 설명한 불용어 확인 방법을 적용하여 2, 3, 4 단어로 구성된 복합명사를 추출하였다. 5 단어 이상의 복합명사도 존재하겠지만 개수가 적을 것이라는 판단에 4 단어 이하의 복합명사만을 추출하도록 하였다.

<표 2> Wall Street Journal에서 추출한 복합명사

	2	3	5	7	10
2단어	7,312	3,394	1,387	811	449
3단어	1,241	497	169	82	45
4단어	242	85	20	14	9
합	8,795	3,976	1,576	907	503

<표 3> IBM 문서에서 추출한 복합명사

	2	3	5	7	10
2단어	1,606	832	384	236	141
3단어	476	195	71	42	26
4단어	125	58	21	10	6
합	2,207	1,085	476	288	173

<표 4> Brown 영역 문장에서 추출한 복합명사

	2	3	5	7	10
2단어	14,024	4,947	1,736	951	520
3단어	1,943	335	79	36	19
4단어	354	31	4	1	0
합	16,321	5,313	1,819	988	539

위의 표는 기준 빈도수를 2, 3, 5, 7, 10으로 정하고 기준 빈도수 이상 나타나는 복합명사의 개수를 계산하여 보여주고 있다. 기준 빈도수가 높을수록 추출되는 복합명사의 개수가 줄어듬은 당연하나, 10번 이상 나타나는 복합명사의 총 수는 1,215로서 논문에서 제시한 간단한 방법에 의해 많은 복합명사를 추출할 수 있음을 알 수 있다.

5. 결론

본 논문에서는 영한 기계번역의 성능 개선을 위해 번역 이전에 번역 대상 문서에서 복합명사를 자동으로 수집하는 방법을 제안하였다. 복합명사를 형성할 수 없는 불용어를 정의하여 문서에서 복합명사 후보를 추출한 후, 빈도수를 기준으로 복합명사로 간주하는 간단한 방법을 적용하였다. 이는 영한 기계번역을 위해 최대한 많은 복합명사를 인식하여 최대한으로 성능을 개선하기 보다는 의미 있는 성능 개선을 약간의 노력으로 달성하려는 목적에 부합한다고 할 수 있다. 따라서 제안한 방법이 실제로 적용될 수 있는 유용성이 있다고 판단한다.

제안한 방법은 번역 대상이 되는 문서에 대해서 의미있는 성능 개선 효과를 기대할 수 다른 문서에 대해서는 효과가 적을 수 있다. 그러나, 제안한 방법을 번역 대상이 아닌 많은 말뭉치에 대해서 적용한다면 많은 복합명사를 수집할 수 있으며, 이를 통해 영한 기계번역 시스템의 성능을 잠재적으로 향상시킬 수 있을 것으로 기대한다.

앞으로 추출한 복합명사에 대한 분석을 통해 불용어를 확장하고, 보다 정교한 불용어 확인 방법을 고안하여 대용량의 데이터로부터 의미있는 복합명사를 추출하는 연구가 필요하다. 또한 추출한 복합명사를 영한 기계번역 시스템이 이용할 수 있도록 구조화 하고, 대역어를 추가한 복합명사 사전으로 확대해야 하며 이는 영한 기계번역 시스템의 성능에도 긍정적인 영향을 미칠 것이다.

참고문헌

[1] Jansche, Martin. "Named Entity Extraction with Conditional Markov Models and Classifiers," Proceedings of Conference on Computational Natural Language Learning, 2002.

[2] H. L. Chieu and H. T. Ng, "Named Entity Recognition with a Maximum Entropy Approach," Proceedings of Conference on Computational Natural Language Learning, 2003.

[3] A. McCallum and W. Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-Enhanced Lexicons," Proceedings of Conference on Computational Natural Language Learning, 2003.

[4] M. Pasca, D. Lin, J. Bigham, A. Lifchits and A. Jain, "Organizing and Searching the World Wide Web of Facts-Step One: The One-Million Fact Extraction Challenge," Proceedings of National Conference on Artificial Intelligence, 2006.

[5] Y. Shinyama and S. Sekine, "Named Entity Discovery Using Comparable News Articles," Proceedings of the International Conference on Computational Linguistics, 2004.

[6] 장두성, 김덕봉, 최기선, "영한 기계번역에서의 복합어구 인식", 제4회 한글 및 한국어 정보처리 학술대회 발표논문집, 503-510, 1992.