제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

웹 문서로부터 한영 병렬말뭉치의 자동 구축

서형원*, 김형철*, 조희영*, 김재훈*, 양성일**
*한국해양대학교 컴퓨터공학과
**한국전자통신연구소
e-mail:{hide90, yhdosu, serensis, jhoon}@bada.hhu.ac.kr, siyang@etri.re.kr

Automatically Constructing English-Korean Parallel Corpus from Web Documents

Hyung-Won Seo*, Hyung-Chul Kim, Hee-Young Cho*, Jae-Hoon Kim*, Sung-Il Yang**
*Dept. of Computer Engineering, Korea Maritime University
**Electronics and Telecommunications Research Institute

요 약

인터넷이 발전하면서 웹에는 같은 내용을 다양한 언어로 표현한 문서들이 많이 존재한다. 이와 같은 웹 문서의 성질을 이용하여, 이 논문은 웹으로부터 수집된 병렬문서(parallel document)를 이용하여 한영 병렬말뭉치 구축 시스템을 설계하고 구현한다. 이 논문에서 구축과정을 요약하면 다음과 같다. 첫째, 웹 문서수집기를 이용해서 웹으로부터 한영 웹문서(html 문서)를 각각 수집한다. 둘째, 수집된 각 언어의 웹 문서에서 불필요한 내용(태그와 광고 문구 등)을 제거하여 문장을 추출하고, 추출된 문장을 단락단위로 정렬한다. 셋째, 단락단위로 정렬된 문서를 문장정렬(sentence alignment) 방법을 이용해서 문장을 정렬한다. 끝으로 정렬된 병렬문장을 단어 단위로 분리하여 병렬말뭉치를 구축한다. 이와 같은 방법으로 이 논문에서는 약 42만 5천 문장의 한영 병렬말뭉치를 구축하였다.

1. 서론

1960년대 이후 많은 연구자들이 자동번역시스템을 개발하기 위한 연구를 진행하였으나(Hutchins and Somers, 1992), 아직은 일반 사용자들은 만족할 만한 시스템은 거의 없다고 해도 과언이 아니다. 미국 국립 표준 기술원(NIST)[1]의 보고에 따르면 가장 좋은 성능을 가진 자동번역시스템의 성능이 0.5137 BLEU[2]이다(NIST, 2005). 물론 이 측도가 인간의 이해도와 완전히 일치한다고는 생각되지 않는다. 다만 이 결과로 볼 때, 약 반세기 동안 꾸준히 연구되고 있지만, 아직도 번역의 질은 그다지 높지 않다는 것이다. 그러나, 자동번역시스템은 완전한 번역보다는 인간의 번역 작업을 도와주거나 외국문서를 이해하는 데는 많은 도움을 주고 있다. 더구나 최근 통신기술의 급속한 성장으로 다양한 언어로 의사를 전달할 필요성이 점점 더 늘어나고 있으며, 자동번역에 대한 수요도 급증하고 있다. 자동번역에 대한 응용 분야를 살펴보면 매우 다양하다. 예를 들면, 단순한 문서번역, 번역업체의 초벌 번역, 웹 문서 번역, 기술 문서 번역, 전자우편 번역, 방송자막 번역, (휴대폰/PDA) 자동 통역, 다국어 정보검색 등이 있다.

최근 인터넷의 급속한 성장으로 다양한 언어로 의사를 전달할 필요성이 점점 더 늘어나고 있으며, 인터넷에는 같은 내용을 다양한 언어로 표현한 문서들을 자주 발견할 수 있다. 예를 들면 제품을 소개하는 매뉴얼, 뉴스 기사[3] 등이 있다. 이 논문은 이와 같이 웹으로부터 공개된 병렬문서(parallel document)를 이용해서 통계기반의 기계번역에 필요하는 병렬 말뭉치(parallel corpus)를 구축하고자 한다. 이 논문에서 구축과정을 요약하면 다음과 같다. 1) 웹 문서수집기를 이용해서 웹으로부터 한영 웹문서(html 문서)를

1) NIST: National Institute of Standards and Technology
2) BLEU : BiLingual Evaluation Understudy의 약자로서 번역의 질을 자동으로 측정하기 위한 하나의 측도이다(Papineni et al., 2001). 이 측도는 N-gram의 공기빈도를 이용하여 번역의 질을 측정한다.
3) engligh.donga.co.kr
engligh.etnews.co.kr
joins.com/cnn

각각 수집한다. 2) 수집된 각 언어의 웹 문서에서 불필요한 내용(태그와 광고 문구 등)을 제거하여 문장을 추출하고, 추출된 문장을 단락단위로 정렬한다. 3) 단락단위로 정렬된 문서를 문장정렬(sentence alignment) 방법을 이용해서 문장을 정렬한다. 4) 정렬된 병렬문장을 단어 단위로 분리하여 병렬말뭉치를 구축한다. 이와 같은 방법으로 이 논문에서는 약 42만 5천 문장의 한영 병렬말뭉치를 구축하였다. 구축된 말뭉치는 통계기반 기계번역에 그대로 이용할 수 있으며 한영 양국어 번역 사전 구축이나 다국어 정보검색 시스템의 색인어 번역 등에 이용될 수 있을 것이다.

이 논문의 구성은 다음과 같다. 2장에서 병렬말뭉치 구축 방법과 문장정렬 방법에 대해서 간단히 소개한다. 3장과 4장에서 각각 병렬말뭉치 구축 시스템과 한영 병렬말뭉치 구축에 대해서 기술한다. 마지막으로 5장에서 결론을 맺고 향후 연구 과제를 기술한다.

2. 관련 연구

2.1 병렬말뭉치

병렬말뭉치는 같은 내용을 두 개 이상의 언어로 표현된 말뭉치를 말한다. 병렬말뭉치의 기원은 Rogetta Stone[4]으로 같은 내용은 이집트어와 그리스어로 표기되었다. 병렬말뭉치의 대표적인 예는 성경이다[5]. 성경은 같은 내용은 수십 가지의 언어로 표현되어 많은 사람들에게 읽혀지고 있다. 병렬말뭉치는 자연언어처리와 기계번역에서 언어정보 구축 분야에서 널리 사용되고 있으며 그 밖에도 언어 교육, 사전 편찬, 대조언어학 연구 등에서 널리 활용되고 있다. 국내에서 개발된 대표적인 병렬 말뭉치는 세종 병렬말뭉치[6], KAIST 병렬말뭉치[7] 등이 있다. 외국의 경우에는 유엔 병렬말뭉치[8], Europarl[9], Hansards[10] 등이 있다.

2.2 병렬말뭉치 구축

병렬말뭉치를 구축하는 방법은 말뭉치 구축 도구를 이용해서 수동으로 구축하는 방법과 자동으로 구축하는 방법이 있다. 수동으로 구축하는 방법은 정확하지만 많은 인력과 시간 그리고 경비가 지출되기 때문에 자주 사용하지 않는다.

병렬말뭉치 구축은 원시문서의 종류에 따라서 조금씩 다를 수 있다. 원시병렬문서는 일반적으로 원시문서와 원시문서의 번역본이다. 이와 같은 병렬문서의 출처는 대개 책이나 웹이다. 이 논문은 웹 문서로부터 병렬문서를 수집하는데 웹으로부터 수집된 병렬문서는 대량의 문서를 쉽게 구할 수는 있으나 두 문서가 완전히 같은 내용이 아닐 경우가 종종 발생한다. 그래서 웹을 통해 자동으로 구축하는 방법은 짧은 시간 동안에 많은 양의 말뭉치를 구축할 수 있으나, 구축된 말뭉치에는 항상 어느 정도의 오류가 포함되어 있을 수 있다.

2.3 문장정렬

문장정렬(sentence alignment or bead)은 웹이나 일반문서로부터 수집된 원시문서와 대역문서로부터 문장단위로 정렬하는 것이다. 일반적으로는 원시언어의 한 문장이 목적언어의 한 문장으로 대응되나(1:1, 약 90%), 그렇지 않은 경우들[11]이 종종 발생된다(Manning, and Schütze, 1999). 문장정렬방법에는 길이기반 정렬방법(length-based alignment method)(Gale and Church, 1993), 편차 정렬방법(offset alignment method), 어휘 정렬방법(lexical alignment method)이 있다. 길이기반 정렬방법은 "번역문의 길이는 원시문장의 길이에 비례한다"라는 가정에서 시작되고, 편차 정렬방법은 병렬말뭉치의 단어(혹은 문자 n-그램)가 나타난 위치의 차이를 이용해서 정렬하는 방식으로 문장정렬에는 그다지 많이 사용되지는 않았다. 어휘 정렬방법은 대역사전 등과 같은 어휘정보를 이용해서 문장을 정렬한다.

문장정렬의 결과는 병렬말뭉치이다. 앞에서 언급했듯이 병렬말뭉치는 수작업으로 구성되는 경우가 정확하지만 많은 인력과 시간 그리고 경비가 지출되기 때문에 자동으로 구축된다. 자동구축의 경우에도 대체로 좋은 성능을 보였다(Gale and Church 1993; Munteanu et al., 2004).

2.4 문장정렬 도구

Align_region(Gale and Church 1993)은 문장의 길이를 통계적으로 모델링하여 문장정렬에 이용한다.

4) http://en.wikipedia.org/wiki/Rosetta_Stone
5) http://www.kidok.info/BIBLE/
6) http://www.sejong.or.kr/
7) http://bola.or.kr/
8) http://www.ldc.upenn.edu/Catalog/readme_files/un.readme.html
9) http://people.csail.mit.edu/koehn/publications/europarl/
10) http://www.isi.edu/natural-language/download/hansard/

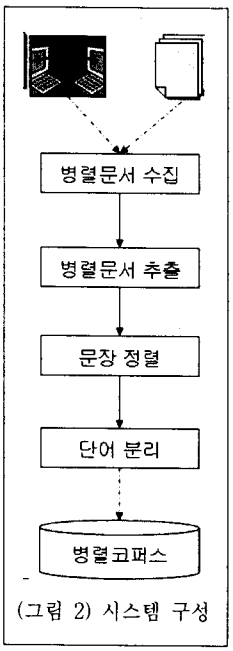
11) 대개 0:1, 1:0, 1:2, 2:1, 1:3, 3:1, 2:2, 2:3, 3:2의 문장정렬이 있다(Manning and Schütze, 1999).

이는 한 언어에서 문장이 길면 번역된 다른 언어의 문장도 길다라는 성질을 이용하여 모델링 하였다. 문장의 대응은 선형적(linear)이라는 성질을 이용하고 있다. 이 도구는 길이 이외 특별한 정보를 이용하지 않기 때문에 모든 언어에 쉽게 적용할 수 있다는 장점이 있으나 웹 문서와 같이 오류가 많이 포함된 병렬 문서를 정렬하기에는 적합하지 않다.

Champollion(Ma 2006)은 사전을 기반으로 문장을 정렬하며 align_region에 비해서 다음과 같은 두 가지의 특징을 가지고 있다. 첫째로, 입력 병렬문서는 일대일 번역을 가정하지 않는다. 따라서 삽입과 삭제와 같은 정렬도 자주 발생할 수 있다. 둘째는, 사전을 기반으로 하여 번역된 단어에 가중치를 주어 접근한다는 것이다. 이는 영어와 아랍어 그리고 중국어에 대해서 적용하여 좋은 결과를 보였다.

3. 한영 병렬말뭉치 구축 시스템의 설계 및 구현

이 절에서는 한영 병렬말뭉치 구축 시스템의 설계 및 구현에 관해서 기술한다. 제안된 시스템은 병렬문서 수집, 병렬문서 추출, 문장정렬, 단어분리 단계로 구성되어 병렬코퍼스를 구축하며, 이하의 절에서 이들 각 단계에 대해서 자세히 설명할 것이다.



(그림 2) 시스템 구성

3.1 병렬문서 수집

같은 뜻의 영어와 한글이 번역되어 문서화되어 있는 것에는 여러 가지 종류가 있지만, 문학 작품의 경우 의역이나 생략 등이 많아서 자동으로 병렬말뭉치를 구축하기에는 적합하지 않다. 따라서 이 논문에서는 비교적 대응관계가 정확한 신문기사를 중심으로 병렬말뭉치를 구축한다. 이 논문에서 병렬문서로 이용하는 인터넷 신문기사는 동아일보[12], 중앙일보[13], 전자신문[14], VOA News[15] 등이다. 이와 같은 신문기사는 웹 문서 수집기[16]를 이용하여 한영 병렬문서(html 문서)를 수집한다.

각 웹 사이트에서는 매우 다양한 형태로 신문기사를 제공하므로 각 사이트마다 서로 다른 문서 수집기를 사용해야 한다. 예를 들면, 동아일보의 경우에는 한국어 기사를 영어로 번역하여 제공하며, 모든 번역된 영어 기사에 한국어 기사의 링크가 있는 것은 아니다. 따라서 번역된 영어 기사를 중심으로 한국어 기사가 링크되어 있는 문서를 찾아야 한다. 한편 중앙일보의 경우에는 CNN 기사를 한국어로 번역하여 서비스를 제공한다. 따라서 번역된 한국어 기사를 중심으로 링크된 영어 문서를 찾아야 한다. 이와 같이 병렬문서를 제공하는 사이트에 따라서 매우 다양한 방법으로 서비스를 제공하므로 수집하고자 하는 사이트마다 서로 다른 문서수집기를 작성해야 한다.

3.2 병렬문서 추출

수집한 HTML문서에서 기사가 들어있는 부분의 특이한 패턴을 정의한 후 그 패턴을 이용한 추출 프로그램을 작성해서 기사만을 추출해 낸다. 하지만, 이 작업 역시 수집한 사이트마다 HTML문서의 양식이 달라서, 사이트 별로 각기 다른 문서추출기가 작성되어야 한다. 또한 하나의 사이트에서도 번역 서비스의 담당자의 변화로 인해 문서 형식의 차이, 그리고 불완전한 번역(표의 번역 유무, 문서 서술형식의 차이 등)으로 병렬문서 추출기의 기능을 매우 복잡하게 한다. 이와 같이 추출된 병렬문서에도 병렬문서로서 적합하지 않은 자료들이 발견될 수 있는데 이들은 병렬문서에 포함되지 않도록 한다. 실제로 문서에서 일부분이 번역이 되지 않은 오류가 있었는데, 이런 문서쌍에서 오류가 있는 문서는 경험적으로 정해진 오류규정[17]을 참조하여 버렸다. 마지막으로 한글과 영어의 번역시 뜻은 상통하지만 문장이 정렬되어지지 않는(1:1로 매칭되지 않는) 경우도 많이 발생하게 되므로, 추출된 영어 및 한글 기사의 문장정렬이 필요하게 된다.

3.3 문장 정렬

2.3절에서 언급했듯이 문장정렬 방법에는 크게

12) http://english.donga.com/
13) http://joins.com/cnn/
14) http://www.etnews.co.kr/
15) http://www.voanews.com/korean/
16) http://www.gnu.org/software/wget/
17) 문장의 개수 차이나 파일 크기가 일정 이상으로 너무 작음 등

사전기반의 문장정렬(champollion)과 길이기반의 문장정렬(align_region) 방법이 있다. 이 논문에서 이들 두 방법 모두를 이용하여 문장을 정렬한다. 이들 두 도구는 서로 장단점을 보완할 수 있기 때문에 두 도구의 결과를 분석함으로써 병렬말뭉치의 질을 더욱 높일 수 있다.

3.4 단어 분리

단어분리는 단어의 활용 등으로 인하여 변형된 단어의 원형을 찾는 과정이다. 단어의 원형을 찾기 위해서는 일반적으로 형태소분석과 품사 태깅을 수행한다. 이 논문에서는 영어와 한국어의 완전한 형태소 분석과 품사 태깅을 이용하지 않고 단순한 단어 분리 기능을 이용한다.

영어의 경우 단어 분리는 비교적 간단한다. 단순히 문장기호를 분리하거나 간단한 약어처리 등으로 단어를 분리할 수 있다. 그러나 한국어의 경우 단순한 문장기호의 분리만으로는 단어가 분리되었다고 말할 수 없다. 왜냐하면 용언의 활용이 다양하고 단어 뒤에 조사나 어미가 매우 다양한 형태로 결합하여 어절을 구성한다. 단어의 활용이 매우 다양하기 때문에 어절 자체를 그대로 사용할 수 없다. 이 논문에서는 단어를 분리하기 위해서 (김재훈&이공주, 2003)을 이용한다.

4. 한영 병렬말뭉치 구축

3장에서 구현된 한영 병렬말뭉치 구축 시스템을 이용하여 약 42만 5천 문장의 병렬말뭉치를 구축하였다(<표 1>). 구축된 말뭉치에 포함된 단어수는 한국어와 영어에 대해서 각각 17,888,617개와 15,581,136개이다.

<표 1> 구축된 병렬말뭉치의 크기

출 처	크 기
동아일보	210,455
전자신문	168,460
중앙일보	6,347
VOA	21,979
기타	17,745
합계	**424,986**

5. 결론 및 향후 연구

이 논문은 웹으로부터 수집된 병렬문서(parallel document)를 이용하여 한영 병렬말뭉치 구축 시스템을 설계하고 구현한다. 이 논문에서 구축과정을 요약하면 다음과 같다. 1) 웹 문서수집기를 이용해서 웹으로부터 한영 웹문서(html 문서)를 각각 수집한다. 2) 수집된 각 언어의 웹 문서에서 불필요한 내용(태그와 광고 문구 등)을 제거하여 문장을 추출하고, 추출된 문장을 단락단위로 정렬한다. 3) 단락단위로 정렬된 문서를 문장정렬(sentence alignment) 방법을 이용해서 문장을 정렬한다. 4) 정렬된 병렬문장을 단어 단위로 분리하여 병렬말뭉치를 구축한다. 이와 같은 방법으로 이 논문에서는 약 42만 5천 문장의 한영 병렬말뭉치를 구축하였다.

앞으로 구축된 말뭉치를 이용하여 통계기반 한영 기계번역 시스템과 한영 양국어 번역 사전 구축, 다국어 정보검색 시스템의 색인어 번역 등에 이용할 계획이다. 또한 정확하게 병렬되지 않은 병렬문서로부터 병렬말뭉치를 구축하는 방법에 대해서 좀더 체계적인 연구가 필요할 것이다.

참고문헌

[1] Gale, W.A. and Church, K.W. (1993). A program for aligning sentences in bilingual corpora, *Computational Linguistics*, vol. 19, no. 1, pp. 75-102.

[2] Hutchins, W. J. and Somers, H. L. (1992) *An Introduction to Machine Translation, Academic* Press, London. Academic Press.

[3] Ma, X. (2006) "Champollion: A Robust Parallel Text Sentence Aligner", *Proceeding of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

[4] Manning, C.D. & Schütze, H. (1999), *Foundations of statistical natural language processing*, Cambridge, MA: MIT Press.

[5] Munteanu, D.S., Fraser, A., and Marcu, D. (2004), "Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora", *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2004)*.

[6] NIST (2005), *Machine Translation Evaluation Official Results*, http://www.nist.gov/speech/tests/summaries/2005/mt05.htm

[7] 김재훈, 이공주 (2003) "사례기반 학습을 이용한 음절기반 한국어 단어 분리 및 범주 결정", 정보처리학회논문지 B, 제10-B권, 제1호, pp. 47-56.