웹 문서로부터 한-영 병렬 말뭉치 자동 구축과 문장 단위 정렬

양주일, 김선호, 송만석

연세대학교 컴퓨터과학과, 서울시 서대문구 신촌동 134, 우:120-749
{zooil, pobi, mssong}@december.yonsei.ac.kr

Mining the Web for Korean-English Parallel Corpora and Sentence Alignment

Zooil Yang, Seonho Kim, Mansuk Song
Deptpartment of Computer Science, Yonsei University

요 약

다국어를 이용한 통계적 자연어 처리의 연구가 진행됨에 따라 병렬 말뭉치의 중요성이 대두되고 있다. 그러나 여러 가지 제약점으로 인하여 현재 이용 가능한 한국어 병렬 말뭉치가 드문 상황이다. 월드 와이드 웹 상에는 다양한 언어로 번역된 문서들이 있으며 이를 병렬 말뭉치로 구축, 활용한다면 말뭉치의 희소성으로 인한 문제를 해결할 수 있다. 본 논문에서는 웹 상에서 번역문서 후보를 추출한 다음 HTML 문서 구조를 비교하여 번역문서인지를 판별하고 문장 단위 정렬을 이용하여 병렬 말뭉치로 구축하는 방법을 제시한다.

1. 서 론

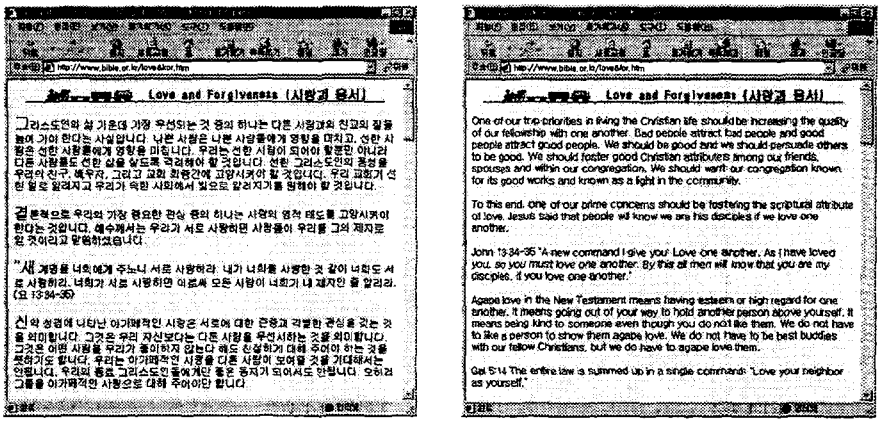
자동 기계 통역, 통계적 기계번역, 정렬을 이용한 대역어 자동 추출, 다국어 정보 검색(cross-lingual information retrieval) 등의 연구가 활발해짐에 따라 병렬 말뭉치의 중요성이 대두되고 있다. 그러나 현재 한국어를 포함하는 병렬 말뭉치의 양이 적고, 특정 역영을 중심으로 구성되었거나, 저작권 문제와 지적 소유권으로 인해 병렬 말뭉치를 구하기 힘든 실정이다.

월드 와이드 웹(World Wide Web) 상에는 여러 종류의 다양한 문서들이 존재하고 있으며, 직접 컴퓨터에서 이용 가능한 텍스트 파일들로 구성되어 있다. 이러한 웹 문서들은 말뭉치를 구축하는 방법으로서 효용 가치가 높으며, 웹 상에 존재하는 번역 문서들은 병렬 말뭉치로서 중요성이 평가되고 있다. 이러한 웹 문서들은 2개 이상의 언어로 번역된 경우도 상당하며 다국어 병렬 말뭉치(multi-lingual bilingual corpus)로 구축이 가능하다.

Resnik은 병렬 웹 문서 후보쌍을 추출하여 HTML 태그를 비교하는 방법과 문자(character)의 개수를 이용한 통계정보를 사용하여 단순히 두 문서의 구조만으로 병렬 문서인지를 결정하는 방법을 제시하였다.[4,5] 그리고 이러한 방식은 동족의 언어에는 어느 정도 유용함이 밝혀졌다. 그러나 한국어의 경우, 다른 언어로 번역된 웹 문서들보다 태그들의 수가 많은 것이 일반적이고, 한국어와 영어는 다른 어족에 속한 언어로서 어순이 다르므로 웹 문서에 나타낸 태그 패턴의 순서가 상이하다. 따라서 위와 같은 방법으로 병렬 문서를 판별할 수 없으며, 병렬 문서라 하더라도 한국어로 구성된 웹 문서의 특성으로 인해 문장 정렬된 말뭉치를 얻으려면 추가적인 정렬 과정이 필요하다.

두 언어 사이에는 의미를 가지는 단어 사이의 출현 빈도가 유사하다. 이러한 특징은 상이한 언어군 사이에서도 나타나며, 형식어를 제외한 의미있는 단어들은 선형적 상관관계를 가진다. 이러한 단어들의 출현 빈도를 이용하여 상이한 어족에 속한 언어 사이에 문장 정렬을 수행 할 수 있다.

본 논문에서는 웹 상에 존재하는 번역 문서들의 후보를 순서쌍으로 생성한 다음, HTML 문서 구조를 비교하여 두 문서가 같은 내용을 포함하는 번역 문서인지를 판별한다. 그리고 번역 문서로 인식된 웹 문서로부터 태그



[그림 1] 한국어, 영어 번역 웹 문서

들 사이의 관계를 이용해 세그먼트 단위로 문장을 추출하고 병렬 말뭉치를 구축한다. 마지막으로 병렬 말뭉치에 존재하는 문장들의 의미 있는 단어와 어휘들의 통계 정보를 이용해 대응 문장간의 score를 계산하고, 이를 바탕으로 문장들간의 최대 유사 정렬(maximum likelihood alignment)을 발견하기 위해 dynamic programming을 사용한다.

2. 웹 상의 번역 문서

웹 상에는 2개국어 이상의 언어로 번역된 다수의 문서들이 존재한다. 이러한 번역 웹 문서들이 같은 내용을 포함하는 경우, 문단 구분, 들여쓰기, 그리고 동일하게 사용된 이미지 배치 등 문서의 구조가 비슷한 경우가 많다. 따라서 문장의 구조 병렬성을 판단하는 기준으로 삼았다.[4,5] 그러나 동일한 내용을 기술하더라도 대부분의 웹 문서를 제작하는데 있어서 어느 한 문서는 상응하는 번역문서보다 더 많은 내용이 들어 있어 문서 내에서 서로 대응되지 않는 부분이 존재할 수 있고, 번역 웹 문서라고 하더라도 문서의 구조가 상이할 수 있다. 또한 같은 내용으로 구성된 페이지의 태그의 수가 일치하지 않는 경우가 있는데, 문장 장식이나 주석문 다른 문서로의 연결 링크의 개수 등으로 인해 어느 한 페이지가 더 많은 태그를 포함하게 된다. 이는 모국어로 구성된 페이지의 경우 일반적으로 나타나는 경향으로, 한국어 문서와 그 번역문의 경우에도 한국어로 구성된 웹 문서가 더 많은 태그을 포함하고 있다.[1]

그림 1은 한국어 웹 문서와 영어로 번역된 웹 문서이다.[2] 이 문서는 문단사이의 1:1 대응이 가능한 잘된 번역의 웹 문서로서 HTML 태그는 그림 2와 같다. 두 문서의 HTML 순서는 비슷한 분포를 가진다. 그러나 한국어 문서는 총 96개의 태그를 사용하였으며, 영어 문서는

<html> <head> <title>Love and Forgiveness (사랑과 용서)</title> </head> <img SRC="single1.gif" ALIGN="bottom" ALT WIDTH="560" HEIGHT="21"> <body> <body BGCOLOR="fff4do" TEXT="Navy"> <p> <p> 그 리스도인의 삶 가운데 가장 우선되는 것 중의 하나는 다른 사람과의 친교의 질을 높여 가야 한다는 사실입니다. ...	<html> <head> <title>Love and Forgiveness</title> </head> <img SRC="single1.gif" ALIGN="bottom" ALT WIDTH="560" HEIGHT="21"> <body> <body bgcolor="fff4do" text="navy"> <p> <p> One of our top priorities in living the Christian life should be increasing the quality of our fellowship with one another. ...

[그림 2] 웹 문서 번역쌍의 HTML 태그

1) 실험 자료에서 한국어 웹 문서가 영어 웹 문서보다 약 4.7% 더 많은 HTML 태그를 사용함.

2) http://www.bible.or.kr/love&kor.htm 와 http://www.bible.or.kr/love&for.htm

53개의 태그만을 사용하고 있다. 이러한 차이는 볼드체나 이탤릭체, 주석문 등의 태그들을 모국어로 작성된 웹 문서에 더욱 빈번하게 사용하며, 웹 문서 작성시 오류로서 문서내에 잘못된 태그의 사용을 들 수 있다. 예를 들어 그림 2에서는 한국어 웹문서에서 글자를 장식하기 위한 그 가 한국어 웹문서에만 쓰였고 두 문서 모두 <body>를 두 번 사용하는 문서 오류가 있다.

3. 웹 문서로부터 병렬 말뭉치 구축과 문장 단위 정렬

웹 문서로부터 병렬 말뭉치 구축과 문장 단위 정렬은 우선, 번역 웹 문서 후보쌍을 추출한다. 다음으로 문서내에 사용된 태그들을 전처리과정을 통하여 분리해내고, 문서 구조 분석을 통해 번역 문서를 판별한다. 이를 통하여 번역 문서로 판별된 두 병렬 웹 문서로부터 문장 세그먼트를 추출하고 문장 단위 정렬을 통해 병렬 말뭉치로 구축하게 된다.

3-1 후보쌍 생성 및 전처리

병렬 웹 문서로서 평가하기 위한 후보쌍을 추출하는 과정은 검색엔진을 이용하였다. 알타비스타[3] 검색엔진에서 조건 검색을 사용하여 Korean과 English에 관련 연결이 존재하는지 검색할 수 있다.[4]

```
anchor:"language1" AND anchor:"language2"
```

한글 알타비스타[4] 검색엔진에서 위의 형식으로 한국어 영어 문서에 대해서 검색한 결과 1081개의 웹 문서를 출력하였다. 그리고 문장 검색 옵션을 이용하여 "korean english" 검색어에 대한 결과 705개를 조사하였다.[5] 그러나 이들 웹 문서들 중 영어 페이지로 링크를 포함하고 있지만 빈 페이지인 경우와 또는 웹 문서가 현재 사라진 경우, 그리고 링크를 포함하고 있지 않으나 korean, english라는 단어만을 포함하는 것이 대부분이었다.

다음 과정으로 이렇게 추출한 후보쌍의 URL 리스트를 이용하여 웹 페이지를 가져온 후, gethtml이란 프로그램을 작성하여 웹 문서에 사용된 이미지나 JAVA Applet 등 다른 불필요한 것들은 제외하고 HTML 태그를 포함해 텍스트만을 저장한다.

병렬 웹문서에 사용된 HTML 문서 구조는 그림 2에서와 같이 비슷한 순서를 가지지만 특정 페이지가 더 많은 양의 태그를 포함하고 있어 이를 이용하여 번역 문서로 판별하기 위해서는 전처리 과정이 필요하다. 전처리 과정에서는 문서 구조 판별에 상관없는 태그들은 제거하고 문서 구조에 영향을 미치는 태그들만을 선택하였다. 표 1은 전처리과정에서 제거한 태그와 선택한 태그들을 나타내고 있다.[6]

문서 구조에 상관없는 HTML 태그	<A>, <B>, <BIG>, <EM>, <FONT>, <META>, <SCRIPT>, <SMALL>, <STRONG>, <U>, <!--COMMANT--> 등
문서 구조에 중요한 HTML 태그	<HTML>, <BODY>, <BLOCKQUOTE>, , <CENTER>, <DIV>, <Hn>, <HR>, <UL>, <OL>, <LI>, <TABLE>, <TR>, <TD> 등

[표 1] 전처리과정에서 제거, 선택된 태그

[START:html]	[START:html]
[START:head]	[START:head]
[START:title]	[START:title]
[WORDCNT:5]	[WORDCNT:3]
[END:title]	[END:title]
[END:head]	[END:head]
[START:center]	[START:center]
[START:img]	[START:img]
[END:center]	[END:center]
[START:body]	[START:body]
[START:body]	[START:body]
[START:p]	[START:p]
[START:p]	[START:p]
[WORDCNT:17]	[WORDCNT:15]
[WORDCNT:12]	[WORDCNT:11]

[그림 3] 전처리과정을 통한 태그의 정렬

3) http://www.altavista.com
4) http://www.altavista.co.kr
5) 한국어와 영어에 대한 웹 문서의 연결은 [Korean][English]와 같은 표현으로 나타나며 문장 검색어 "korean english"로 검색 가능
6) HTML 태그들은 http://www.w3.org/TR/html40/index/elements.html 문서의 목록에서 분류

전처리과정에서 선택된 태그는 다음과 같이 분류하였다.

시작태그 : [START:tag_name]
종료태그 : [END:tag_name]
단어개수 : [WORDCNT:단어개수]

시작과 종료태그는 태그들의 이름을 포함하고 있으며 단어개수는 태그를 제외한 문장의 단어수를 의미한다. 그림 3 은 전처리과정을 통해 문장 순서대로 분류된 태그들을 나타내고 있다.

3-2 문서구조를 이용한 번역 문서쌍 결정

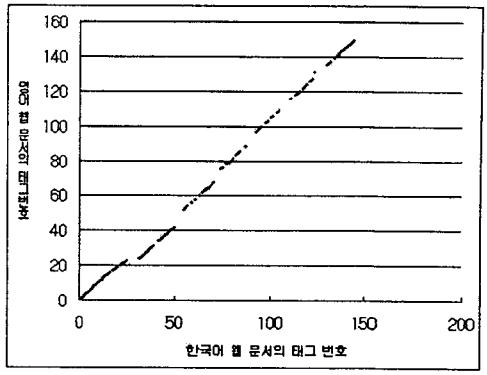
전처리 과정을 거친 웹 문서들을 정확한 번역문서인지 아닌지 평가하는 방법으로 한국어와 영어에 대해 2차원 순서쌍 매핑 방법을 이용하였다. 우선, 한국어 태그에 대해 모든 가능한 영어 태그를 매핑 시키고 선형적으로 증가하지 않는 태그쌍들을 제거하였다. 그림 4은 한국어와 영어 태그에 대한 매핑된 순서쌍을 그래프로 나타낸 것이다. 문서구조를 이용한 번역 문서쌍 결정과정에서 일치하지 않는 태그쌍이 전체의 20%를 넘으면 서로 번역문서가 아닌 것으로 판별하였다.

3-3 문장 단위 정렬 및 말뭉치 구축

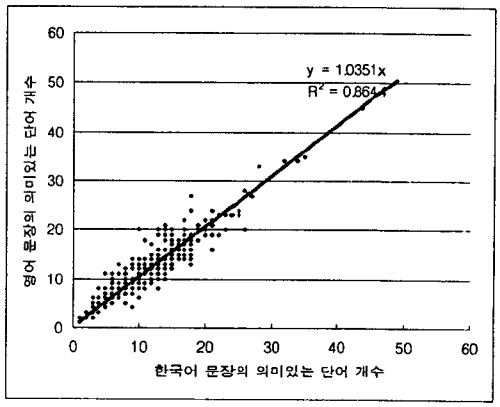
병렬 말뭉치로 구축하기 위한 마지막 단계는 번역 웹 문서로 평가된 후보쌍으로부터 태그들을 제외한 문장 세그먼트를 추출한 다음, 문장 단위 정렬을 통해 병렬 말뭉치로 구축하게 된다.

문장 단위 정렬은 의미 있는 단어들의 개수를 이용하여 dynamic programming 기법으로 정렬한다. 원문이 번역될 때, 그것에 대응하는 번역문의 길이 사이에는 상관관계가 있다.[3] 이러한 문장 간의 길이의 상관관계를 이용한 통계적 방법으로 문장 단위 정렬을 수행할 수 있다.[1] 그러나 영어와 프랑스어와 같이 동일한 어족에 속하는 언어간의 정렬에는 어느 정도 효용성이 있으나, 문장의 길이가 비슷하지 않고 상이하게 다른 언어쌍에 대해서 좋은 결과를 기대할 수 없다.[8]

의미를 지니지 않고 문장 내에서 형식적인 기능을 하는 영어에서 관사는 한국어에 존재하지 않으며, 전치사는 한국어에서 조사로서 나타난다. 한국어의 어절 수는 관



[그림 4] 한국어와 영어 태그의 2차원 매핑



[그림 5] 한국어와 영어의 의미있는 단어수 사이의 상관 관계 그래프

사와 전치사를 제거한 영어 문장의 단어수와 일정한 상관관계를 가진다. 그림 5는 CNN live[7] 에서 추출한 300개의 문장가운데 의미있는 단어들의 상관관계를 보여준다. 이러한 상호 관련성을 바탕으로 dynamic programming algorithm을 이용하여 번역문 사이의 문장 정렬을 수행한다.

- **Dynamic Programming Algorithm**

원본 문서의 문장을 s_i (i=1, ..., I), 번역 문서의 문장을 t_j (j=1,...J) 그리고 두 문장의 거리 함수를 d라 할 때, 두 문장의 거리를 최소화 하는 D(i,j)는 다음과 같다.[3]

7) 도서출판 다락원에서 출판하는 CNN 방송 청취용 정기간행물

$$D(i,j) = \min = \begin{pmatrix} D(i,j-1) & + & d(0,t_j;0,0) \\ D(i-1,j) & + & d(s_i,0;0,0 \\ D(i-1),(j-1) & + & d(s_i,t_j;0,0) \\ D(i-1),(j-2) & + & d(s_i,t_j;0,t_{j-1}) \\ D(i-2),(j-1) & + & d(s_i,t_j;s_{i-1},0) \\ D(i-2),(j-2) & + & d(s_i,t_j;s_{i-1},t_{j-1}) \end{pmatrix}$$

4. 실험 및 결과

후보추출 과정을 통해 얻은 210개(한국어 105개, 영어 105개)의 웹 문서쌍에 대해 실험을 수행하였다. 후보쌍들 중 정확한 번역인지 문서 구조 분석을 통해 결정된 웹 문서쌍은 47개였으며, 38개의 문서가 동일한 번역문이었다.

문서 구조 분석을 통한 세그먼트 추출 과정에서 [그림 5] 같은 어휘 사전을 얻을 수 있었다. 이것은 웹 문서내에서 링크로 사용된 단어로서 태그들로 나뉜 부분에서 정렬된 결과이다.

정치	Politics
외교	Diplomacy
경제	Economy
과학	Science
문화	Cultural
유산	Assets
스포츠	Sports
언어	Korean Language
한국소개	Introduction

[그림 5] 문서 구조 분석과정에서 추출된 어휘

[그림 6]은 정렬과정을 통해 얻은 문장을 나타낸다. 정렬된 문장은 대부분 문장 사이의 대응 관계가 1:1이었으며, 2:1, 1:2 등의 여러 문장 사이의 정렬은 나타나지 않았다. 이는 다수의 웹 문서에서 번역문이 각 문장들 사이의 1:1 번역으로 이루어졌고, 여러 개의 문장으로 번역된 문서는 문서 구조 결정 과정에서 문서가 일치하지 않는 것으로 판단, 제거되었기 때문이다.

5. 결론 및 향후 연구

본 논문은 웹 상에서 번역 문서 후보쌍을 추출한 다음 문서 구조를 이용하여, 번역문인지 결정하고 문장 정렬을 통한 한국어-영어 병렬 말뭉치 구축을 시도하였다. 실험을 통하여 웹 문서를 병렬 말뭉치 기초 자료로 활용과 의미있는 단어를 이용하여 문장 정렬을 수행할 수 있었다. 향후 연구 계획은 후보쌍 추출시 다양한 접근 방

그리스도인의 삶 가운데 가장 우선되는 것 중의 하나는 다른 사람과의 친교의 질을 높여 가야 한다는 사실입니다.
One of our top priorities in living the Christian life should be increasing
the quality of our fellowship with one another.

나쁜 사람은 나쁜 사람들에게 영향을 미치고 선한 사람은 선한 사람들에게 영향을 미칩니다.
Bad people attract bad people and good people attract good people.

한국인이 하나의 언어를 사용해 온 것은 민족적 동질성을 유지하는데 중요한 역할을 해왔다.
The Korean people speak one language, an exclusive language of their own, which plays an important role in keeping a compatriotic feeling among the people.

한국어문자인 한글은 1443년 조선시대 세종대왕이 창제하였으며 세계 문자사에서 가장 과학적인 문자체계를 가진 것으로 평가되고 있다.
Hangul is the Korean alphabet and it was invented in the 15thcentury under the reign of King Sejong the 4thKing of the Chosun Dynasty.

[그림 6] 병렬 말뭉치로 구축한 문장 예

법의 시도와 문장 정렬시 통계적 방법만을 이용하지 않고, 문장내의 어휘 정보를 이용한 정렬에 관한 연구 그리고 다수의 문장으로 대응되는 문서의 문장정렬에 관해 연구하는 것이다. 그리고 한국어-영어 웹 문서 이외에 다른 언어로 구성된 다국어 페이지에 대한 연구가 필요하다.

참고문헌

[1] Brown, P. F., Jenniger C. Lai, and Robert L. Mercer, Aligning Sentences in Parallel Corpora, *In Proceedings of the 29th Annual Meeting of the Association for Computational linguistics*, 1991

[2] Dekai Wu, Aligning A Parallel English-Chineses Corpus Statistically with Lexical Criteria, *32nd Annual Meeting of the Assoc. for Computational Linguistics*, ACL-94, 1994

[3] Gale, W. A., and Church, K. W., A Program for Aligning Sentences in Bilingual Corpora, Using Large Copora, MIT Press. 1994

[4] Phlip Resnik, Parallel strands: A preliminary investigation in to mining the web for bilingual text, *In Proceedings of the Third Conference of the Association for Machine Translation in the Americas,* AMTA-98, 1998

[5] Philip Resnik, Mining the Web for Bilingual Text, *37th Annual Meeting of the Association for Computational Linguistics,* ACL'99, 1999

[6] Philip Resnik, Evaluating Multilingual Gisting of Web Pages, *presentation at the AAAI Symposium on Natural Language Processing for the World Wide Web*, 1997

[7] Simard, M. and Plamondon, P., Bilingual Sentence Alignment: Balancing Robustness and Accuracy, Machine Translation vol. 13, no. 1, 1998

[8] 이주호, 자동 정렬을 통한 영한 복합어의 역어추출, 한국과학기술원 전산학과 석사학위 논문, 1999