웹용 다국어 기계번역을 위한 전처리기

안동언, 이영우[U], 서진원, 정성종
전북대학교 컴퓨터공학과

A Preprocessing System for Multi-Lingual Machine Translation of Web Pages

Dong un An, Young Woo Lee, Jin Won Seo, Sung Jong Chung
Department of Computer Engineering, Chonbuk National University

요 약

여러 언어들로 작성된 웹문서들을 다국어 기계번역기에서 번역하기 위해서는 우선 해당 웹문서가 어떠한 언어로 작성되었는지를 알아내야 한다. 코드 분석을 통하여 웹문서를 작성한 언어를 알게되면 해당 언어를 번역하는 기계번역기를 작동시킬 수 있다 또한, 웹문서에서 기계번역의 대상은 HTML 태그를 제외한 일반 문장이다 따라서, 웹용 기계번역의 전처리기에서 웹문서에서 HTML 태그를 분리하여야 하며 번역이 완료된 후 번역된 문서에 HTML 태그를 복원하여 웹브라우저에서 번역된 문서를 볼 수 있어야 한다. 본 논문에서는 웹용 다국어 기계번역을 위한 전처리기의 태그관리기와 코드인식기를 설명한다

1. 개 요

최근에 웹이 빠른 속도로 확산되어 많은 사람들이 인터넷에서 정보를 획득하고 있다. 그런데 일반 사용자들이 정보를 획득하는데 있어서 장애가 되는 것 중의 하나가 이해하지 못하는 언어로 작성된 웹문서이다 이를 극복하기 위하여 기계번역시스템을 사용하고자 하는 노력이 있어왔다

여러 언어들로 작성된 웹문서들을 다국어 기계번역기에서 번역하기 위해서는 우선 해당 웹문서가 어떠한 언어로 작성되었는지를 알아내야 한다. 코드 분석을 통하여 웹문서를 작성한 언어를 알게되면 해당 언어를 번역하는 기계번역기를 작동시킬 수 있다. 현재 상용화된 웹용 기계번역시스템들은 다국어가 아닌 영어 또는 일어의 하나의 언어만을 대상으로 하고 있다 [1, 2] 영일한 기계번역기라고 하더라도 사용자가 언어의 선택하도록 되어 있다.

"에서로-웹/EK"[3]를 다국어 기계번역 시스템인 "에서로-웹/다국어"로 확장하기 위해서는 어떤 문자로 웹문서가 작성이 되었는지 알아 낼 수 있는 코드 인식기가 필요하다. 1바이트 코드인 영어, 독어, 프랑스어와 2바이트 코드인 한국어, 일어, 중국어를 자동으로 인식하도록 한다.

본 연구는 한국진자통신연구원의 지원으로 수행된 "기계번역을 위한 웹문서 특성 처리 연구"의 연구결과중 일부분임니다.

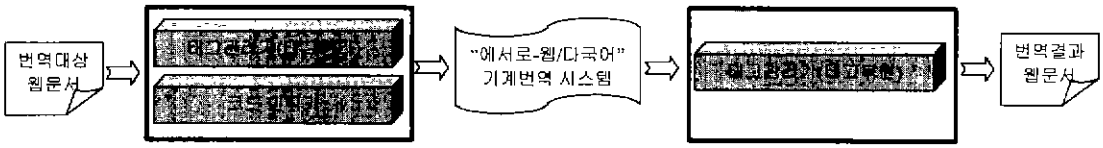
웹용 기계번역시스템은 웹문서를 처리할 수 있도록 기존의 일반 기계번역시스템을 확장한 것이다 실제 웹문서에서 기계번역의 대상이 되는 것은 HTML 태그를 제외한 일반 문장이다. 따라서, 웹용 기계번역의 전처리기에서 웹문서에 HTML 태그를 분리하여야 하며 번역이 완료된 후 번역된 문서에 HTML 태그를 복원하여 웹브라우저에서 번역된 문서를 볼 수 있어야 한다.

본 논문에서는 웹용 다국어 기계번역을 위한 전처리기의 태그관리기와 코드인식기를 설명한다 2장에서는 전처리기의 구성을 보여주고 3장과 4장에서는 태그관리기와 코드인식기를 각각 설명한다 5장에서는 실험결과를 기술한다

2. 전처리기의 구성

웹용 다국어 기계번역기의 전처리기는 <그림 1>과 같은 태그관리기와 코드인식기로 구성되어 있다

인터넷을 통해 얻어진 번역대상 웹문서를 먼저 태그관리기를 통해 기계번역의 대상이 일반 문장과 HTML 태그로 분리한다. 이때, 번역된 결과 문장과 태그를 결합할 수 있도록 태그가 분리된 위치를 기록하여 파일에 저장하여야 한다 HTML 태그의 형태 정보를 이용하여 LEX로 작성된 오토마타를 이용하여 웹문서에서 문장과 태그를 분리하였다



<그림 1> 웹용 다국어 기계번역기의 전처리기

이렇게 분석되어 얻어진 일반 문장에서 코드인식에 필요한 자질들을 추출하여 웹문서의 코드를 분석한다. 웹문서 편집기를 이용한 경우에 메타 태그에 사용한 코드 종류가 적혀 있는 경우가 있다. 이 경우에는 메타 태그만을 분석하면 된다. 그렇지 않은 경우에는 코드를 자동으로 인식할 수 있는 코드인식기가 필요하다. 기존의 전처리기에서는 코드인식기에 대한 고려는 없었다[4, 5]

3. 태그관리기

태그관리기는 웹문서의 특정인 태그들에 대한 분리 및 복원을 처리한다. 일반적인 문서와 다르게 웹문서는 HTML 태그들을 가지고 있기 때문에 번역 대상이 되는 문장만을 기계번역 시스템에 넘겨야 한다. 태그들은 일정한 패턴에 의해서 구성이 되므로 정규표현에 의해서 쉽게 표현할 수 있다. 따라서, 태그관리기는 LEX를 이용하여 구현한 오토마타이다

다음과 같이 몇 가지 태그를 LEX의 표현으로 나타낼 수 있다

- 일반적인 태그형태 \<[^\>]+\>
- 값을 가진 태그형태 \<{ALPHA_NUM}+({WC}+
  {ALPHA_NUM}+{WC}*\={WC}*[^]+)+\>
- 하이퍼링크 [Hh][Tt][Tt][Pp]" //"[^ \t\n]+

위와 같이 정의된 태그표현에 의해 태그의 분리가 이루어지며, 스크립트 언어까지 정규표현에 의한 인식과 분리가 가능하다.

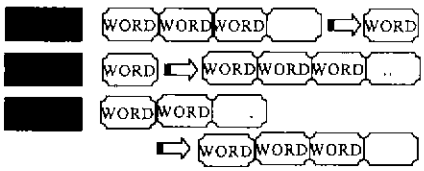
웹문서에서 단순히 태그를 제거하는 것은 매우 간단한 작업이다 그렇지만, 웹문서에서 일반문장을 추출한 후 태그를 복원하여 번역된 웹문서를 만들기 위해서 태그들은 4가지 종류로 분류하였다.

- 문장 시작 태그(sent_start_tag)
- 문장 끝 태그(sent_end_tag)
- 단어 시작 태그(start_tag)
- 단어 끝 태그(end_tag)

일반적으로 기계번역시스템은 한 문장 단위로 번역을 한다. 따라서, 문서를 문장 단위로 나누어 주어야 한다. ".", "!", "?"와 같은 종결부호는 당연히 문장의 분리 단서로 이용한다. 태그들 중에도 문장 분리의 단서를 제공하는 것들이 있다. 웹문서에 있는 표나 테이블의 경우에는 셀에 있는 내용이 한 단어라고 하더라도 독립적으로 번역이 되어야 한다 단순히 태그만을 분리한다면 이러한 부분은 연관성이 없는 단어들로 이루어진 문장이 되어서 파싱에서부터 문제가 생기게 된다. 따라서, "htm", "body", "title", "td", "r", "br" 등과 같은 태그들은 문장 시작 태그로 분류하여 문장을 기준으로 하여 태그들을 분리한다.

태그 복원에서는 분리된 문장이 기계번역 시스템을 거쳐 번역이 완료된 후에 기계번역 시스템이 넘겨주는 대역어와 위치 정보와 비교하여 태그들을 다시 복원시킨다. 이를 위하여 태그를 분리하면서 문장 번호와 단어 번호를 태그의 위치정보로 태그 파일에 기록한다

태그 복원 과정에서 나타나는 문제점들의 하나가 <그림 2>와 같은 경우이다. 번역대상 단어열과 번역결과 단어열이 1:1 대응이 되지 않는 경우이다. 이런 경우를 처리하기 위하여 단어 시작 태그를 분류한 것이다. 번역 결과에 의해서 단어열이 축소가 되거나 확장이 되어도 태그 파일의 단어 위치 정보를 이용하여 복원한다.



<그림 2> 태그 복원의 모호성

4. 코드인식기

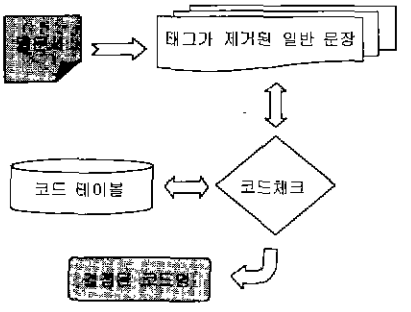
다국어 환경에 맞춘 능동적인 다국어 웹 기계번역을 위해서는 다양한 코드의 웹문서를 처리할 수 있어야 한다 본 연구에서는 현재 인터넷 상에서 웹문서에 쓰이는 다양한 언어들의 코드를 인식하여 다국어 기계번역 시스템이 자동적으로 번역 대상과 번역 결과에 해당하는 언어를 처리하는 기계번역 모듈을 동작하도록 하는 것이다.

1바이트 코드인 영어, 불어, 독어와 2바이트 코드인 한국어, 중국어, 일본어를 처리하도록 한다 현재 웹브라이저에서 지원하고 있는 일본어와 중국어 코드는 다음과 같다[6]

- Japanese (Shift-JIS)
- Japanese (EUC-JP)
- Traditional Chinese (Big5)
- Traditional Chinese (EUC-TW)
- Simplified Chinese (GB2312)

웹문서 편집기를 이용한 경우에 메타 태그에 사용한 코드의 종류가 적혀 있는 경우가 있다. 이 경우에는 메타 태그만을 분석하면 된다

그렇지 않은 경우에는 코드를 자동으로 알아내는 코드인식기가 필요하다. 전체적인 개요는 <그림 3>과 같다.

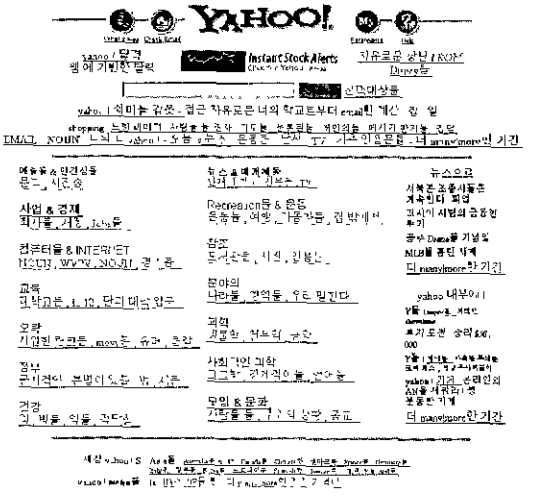


<그림 3> 코드 인식

코드인식은 대상 웹문서에서 태그들이 분리되고 기계번역 시스템으로 보내는 일반 문장을 첫 바이트에서부터 차례대로 코드데이블과 비교해 나간다. 여러 코드들 중에서 하나의 확실한 코드를 판단할 수 있을 때까지 코드 범위 체크를 반복한다

예를 들어 일본어 코드 중에서 Shift-JIS 코드의 형식은 다음과 같다[6].

< Shift-JIS Encoding Specification >
Two-Byte characters
First Byte 129-159, 224-239
Second byte 64-126, 128-252
Half-width katakana
Byte range 161-223
ASCII/JIS-Roman
Byte range 33-126



<그림 5> 영한 기계번역 후 복원된 한국어 웹문서

5. 실험 및 결과

태그관리기의 태그 분리는 LEX를 이용한 오토마타로 처리하고 있다. 태그 파일에 위치정보를 부여하는 부분과 태그를 복원하는 부분과 코드인식기는 C 언어로 구성하였다.

<그림 4>는 영문 웹문서에 대하여 코드를 인식하고 태그를 분리한 후 영한 기계번역이 완료된 문서에 대하여 태그를 복원한 결과이다. 웹문서의 태그가 완벽하게 복원된 것을 알 수 있다

6. 결론

본 연구에서는 웹상의 다국어 환경에 능동적으로 대처하기 위하여 웹용 다국어 기계번역 시스템의 전처리기인 태그관리기와 코드인식기를 개발하였다

태그관리기는 번역대상 웹문서에서 일반 문장과 HTML 태그를 분리하여 기존의 기계번역시스템을 이용할 수 있도록 한다. 번역된 결과 문장은 분리되었던 태그와 결합하여 번역결과 웹문서를 복원할 수 있도록 하였다. 이를 위하여 태그를 4가지 종류로 분류하였고 문장과 단어 위치 정보를 태그 파일에 기록하여 태그를 복원할 때 이용하였다. 또한, 웹문서가 어느 언어로 작성되었는지 자동으로 알아낼 수 있는 코드 인식기를 개발하여 다국어 기계번역 시스템 개발 환경에 대비하였다.

앞으로의 과제는 다음과 같다 단순히 종결부호와 태그만을 이용하여 문장을 분리하는 것은 정확하지 않다. 좀 더 정확한 문장 분리를 위해서는 독립된 처리 모듈이 필요하다 또한, 프레임을 사용한 웹문서를 번역하고 복원하기 위해서는 프레임을 구성하는 여러 웹문서를 반복적으로 호출하는 기능이 필요하다.

참고문헌

[1] 앙꼬르 3.0-IBM 영한 기계번역 시스템, 하이소프트, 1998
[2] Transle98-EJK, 서울대학교 언어공학연구소, 1998
[3] 심철민, 여상회, 정한민, 김태완, 박동인, 권혁철, "에서로-웹/EK :영한 웹 문서 번역 시스템," 제 9회 한글 및 한국어 정보처리 학술대회, 1997, pp 277-282
[4] 여상회, 정한민, 최영숙, 김태완, 박동인, "실용적인 영한 기계번역을 위한 전처리기의 설계 및 구현," 제 8회 한글 및 한국어 정보처리 학술대회, 1996, pp 313-319
[5] 안동인, 유홍진, 서진원, 이영우, 경성종, 여상화, 김태완, 박동인, "웹용 영한 기계번역을 위한 문서 전처리기의 설계 및 구현," 제 9회 한글 및 한국어 정보처리 학술대회, 1997, pp 249-254
[6] Ken Lunde, Understanding Japanese Information Processing, O'Reilly & Associates Inc, 1993, pp 59-100