(제 10회 한글 및 한국어 정보처리 학술대회)

영어 웹문서 기계번역을 위한 태그 관리기

안동언, 서진원*, 이영우, 정성종

전북대학교 컴퓨터공학과
전북 전주시 덕진구 덕진동1가 664-14
duan@moak.chonbuk.ac.kr, {jin, apollo}@calhp1.chonbuk.ac.kr, sjchung@moak.chonbuk.ac.kr

Tag Manager for Machine Translation of English Web Page

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

요 약

영어 웹문서를 한국어로 기계번역을 하기 위해서는 웹문서에 있는 HTML 태그들을 처리하여야 한다. 본 논문에서는 웹문서의 태그들을 처리해 주는 태그 관리기를 제안한다. 태그 관리기는 영한기계번역의 대상이 되는 영어 웹문서에서 태그를 분리하고, 번역이 완료된 후에는 분리된 태그들을 올바른 위치에 복원시키는 기능을 갖는다. 태그 관리기는 태그들의 위치정보에 따른 태그들의 분류와 이를 분리하고 복원하는 기능을 가지고 태그의 내용에 따른 문장 분리기능도 가진다.

1. 서론

인터넷에서 웹의 확산은 대단한 일이다. 불과 몇 년 사이에 인터넷은 대단히 빠른 속도로 발전해왔다. 몇몇 전문사용자의 전유물인 인터넷이 일반인들이 접하게 되면서 언어의 장벽은 새로운 문제로 떠오르고 있다. 인터넷에서 접할 수 있는 대부분의 웹문서들이 영어로 되어 있기 때문에 일반 사용자들이 정보를 획득하는데 걸림돌이 되고 있다.

따라서 인터넷을 이용하는데 있어서 영한기계번역시스템을 사용하고자 하는 요구가 생기고 있다. 웹문서의 기계번역시스템은 일반 문장의 기계번역시스템에 HTML 태그를 처리하는 태그 관리기를 추가한 것이다[1].

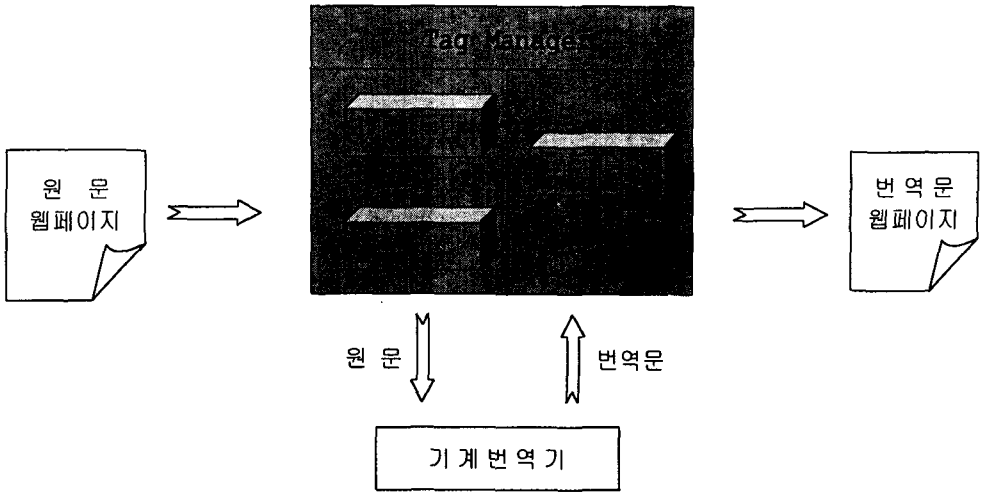
문장을 번역 단위로 하는 번역기에서 태그는 번역 대상이 아니다. 따라서, 태그 관리기에서는 영어 웹문서에서 태그들을 만나게 되면 태그들을 파일에 기록하고 번역이 완료되면 파일의 내용을 가져와서 한국어 웹문서를 복원시키는 방법을 사용한다.

일한기계번역과는 다르게 영한기계번역에서는 영어와 한국어의 어순이 다르기 때문에 번역된 문서와 태그를 결합하여 한국어 웹문서를 만들 때 태그를 올바른 위치에 복원하기 위한 방안이 마련되어야 한다.

본 연구는 한국전자통신연구소의 지원으로 수행된 "기계번역을 위한 웹문서 특성처리 연구"의 결과중 일부분입니다.

2. 태그 관리기의 구성

태그 관리기(Tag Manager)의 구성은 <그



<그림 1> 태그 관리기의 구성

림1>과 같다. 영한기계번역기는 한국전자통신연구원에서 개발중인 에서로-웹/EK를 사용한다[2].

태그 관리기는 영한기계번역기의 전처리로 태그분리기와 문장분리기가 있다. 후처리로는 태그복원기가 있다. 이 문장분리기는 [3]과는 다르게 문장 부호와 태그 정보만을 이용하여 문장을 분리해주는 단순한 것이다.

3. 태그 분리기

3.1 태그 분류

태그 분리기에서는 다음과 같이 웹문서에서 사용되는 태그들을 네 종류로 분류한다.

- Sentence_Start Tag
- Sentence_End Tag
- Word_Start Tag
- Word_End Tag

Sentence_Start 태그와 Sentence_End 태그는 문장을 꾸미는데 사용하는 태그이고, Word_ Start 태그와 Word_End 태그는 단어를 꾸미는데 사용되는 태그이다. 이러한 태그들은 실제 웹문서를 조사하여 태그가 나타나는 위치에 따라서 분류하였다.

분류된 태그들은 <표 1>과 같다.

Sentence_ Start 태그	<html> <body> <title> <head> <img> <table> <td> <option> <select> <hr> <map> <area> <script> <input> <p> <meta> <dl> <dt> <dd> <form>
Sentence_ End 태그	</html> </body> </title> </head> </table> </td> </option> </select> <hr> </map> </script> </dl> </dt> </form> 등
Word_Start 태그	 <address> <blink> <h1-6> 등
Word_End 태그	 </address> </blink> </h1-6>등.

<표 1> 웹문서 태그 분류

번역기는 문장 단위로 번역하기 때문에 번역 단위인 문장을 결정하는 것은 중요한 일

이다. Sentence_Start 태그와 Sentence_End 태그를 문장을 분리하는데 이용한다.

3.2 태그 파일

태그 관리기는 태그를 만나면 네 가지 분류에 따라서 태그에 정보를 부여하고 임시 파일에 태그 정보를 기록한다. 임시 파일의 형식은 다음과 같다.

```
16 0 Sstart_tag : <td width=160>
16 0 Wstart_tag : <a href="find.htm">
16 0 Wstart_tag : <b>
17 1 Wend_tag : </b>
17 1 Wend_tag : </a>
18 0 Sstart_tag : <br>
18 3 Send_tag : </td>
```

첫 번째 칼럼의 숫자는 문장 번호이고, 두 번째 칼럼의 숫자는 어절 번호이다. 세 번째는 태그의 분류이고 네 번째는 태그의 내용이다. 그러므로 위의 예에서 첫 번째 라인의 내용은 '16번째 문장의 첫 번째 단어 앞에 Sentence_Start 태그인 <td width=160>이라는 태그가 있다는 정보를 나타낸다. 문장 번호와 어절 번호를 같이 저장하는 이유는 복원시 위치를 참조하기 위함이다.

3.3 문장 분리

문장의 종결부호가 없는 문장은 태그 정보를 가지고 문장을 분리한다. 종결부호가 없고 태그가 Sentence_Start 태그이면 태그 관리기는 새로운 문장을 만들면서 문장번호를 증가시킨다.

문장을 분리할 때 Sentence_Start 태그와 Sentence_End 태그를 모두 이용할 수 있다. 그렇지만 Sentence_End 태그에서 새로운 문장을 분리할 경우 cgi의 입력 폼에서 부정확한 결과를 보이는 경우가 있다. 따라서 Sentence_Start 태그가 나오면 새로운 문장을 분리한다.

3.4 스크립트 및 이미지 맵 처리

스크립트 태그(<script>, </script>)와 이미지 맵 태그(<map>, </map>)는 시작 태그와 끝 태그 안의 내용을 번역할 필요가 없다. 따라서, 태그와 그 안의 내용 모두를 Sentence_Start 태그로 인식하여 번역 대상에서 제외한다.

4. 태그 복원기

태그 복원은 번역기에서 넘어오는 문장과 임시 파일의 태그를 결합하여 번역이 완료된 한국어 웹문서를 만든다. 태그를 파일로 저장할 때 얻었던 문장번호와 어절 번호의 정보로서 복원이 가능하다. 번역기에서 번역되어 전달되는 형식은 아래와 같다.

```
원문 :
AltaVista Main Page
번역문 :
  (("알타비스타" " 0")
  ("주요 페이지" "1 2 "))
```

번역문 뒤의 숫자는 그 번호의 영어 어절이 번역되어서 한국어 번역어가 생성되었다는 것을 나타낸다. 이 정보와 임시파일의 문장과 어절 번호를 비교하여 복원을 하게 된다.

4.1 태그의 복원

태그를 복원하면서 영어와 한국어간의 구조적인 차이와 어휘적인 차이를 고려해야 한다.

4.1.1 구조적 차이

영어는 SVO 언어이고 한국어는 SOV 언어이므로 번역되었을 때 어순이 다른 것은 당연한 것이다. 이 경우에 단어와 관련된 태그들은 변경된 어절의 위치를 찾아서 태그를 복원해 주어야 한다. 번역문에서 번

역어와 같이 넘어오는 어절 번호를 비교하여 복원한다.

4.1.2 어휘적 차이

한글과 영어의 어휘적인 차이로 인하여 번역된 결과가 원문의 어절 수와 다른 경우가 있다.

1 : N → manageable : 조작하기 쉬운
N : 1 → give up : 포기하다

1 : N의 경우에는 번역된 한국어 단어가 두 어절 이상이라고 하더라도 영어 어절은 하나이기 때문에 태그를 복원하는데 별다른 문제가 없다. 영어 한 어절의 단어 시작 태그와 끝 태그를 번역된 한국어 단어들의 시작과 끝 태그로 복원하면 된다.

N : 1의 경우에는 영어 어절들을 찾아서 단어 시작 태그들과 단어 끝 태그들을 모은다. 모아진 시작 태그와 끝 태그를 번역된 한국어 어절에 복원한다.

4.2 태그의 검증

웹문서에서 사용되는 태그는 반드시 < >로 열어서 </ >로 마치게 되어 있다. 이러한 순서가 지켜지지 않으면 태그로서의 기능은 상실된다. 그러므로 태그 관리기에서는 복원시 태그의 순서쌍을 검증한다. 태그는 중첩 구조를 가지므로 스택을 사용하여 태그를 복원하고 검증한다.

5. 구현 및 실험

본 논문에서 제안한 태그 관리기는 PC 윈도우 환경에서 작동하며 C언어로 구현하였다.

<그림 2>는 번역대상 문서로서 AltaVista의 메인 화면이며, <그림 3>은 태그 관리기에서 분리하고 복원까지 수행한 한국어 웹문서이다.

본 시스템의 수행 결과를 분석해 보면 Sentence_Start 태그와 Sentence_End 태그를 이용한 문서의 레이아웃은 만족할 만한 복원 결과를 보여준다. 하지만 문장의 가운데에 있는 링크 정보의 복원에서는 아직 미흡하다. 이것은 링크 정보가 하나의 독립된 문장으로 쓰이는 경우도 있고, 단어를 꾸며주는 경우도 있기 때문이다.

6. 결론

본 연구에서는 효과적인 웹문서 번역을 위한 태그 관리기를 제안하였다.

태그 관리기는 문장의 효율적인 분리, 태그의 종류별 분류, 태그의 분리 및 복원을 수행한다. 태그 관리기에서는 태그를 분리하여 임시파일에 저장한 후에 번역이 완료된 후에 번역문과 결합하여 동일한 웹문서를 구성한다.

복원시 태그의 순서를 체크하기 위하여 스택을 사용하여 순서가 바뀌는 경우를 검증하였다.

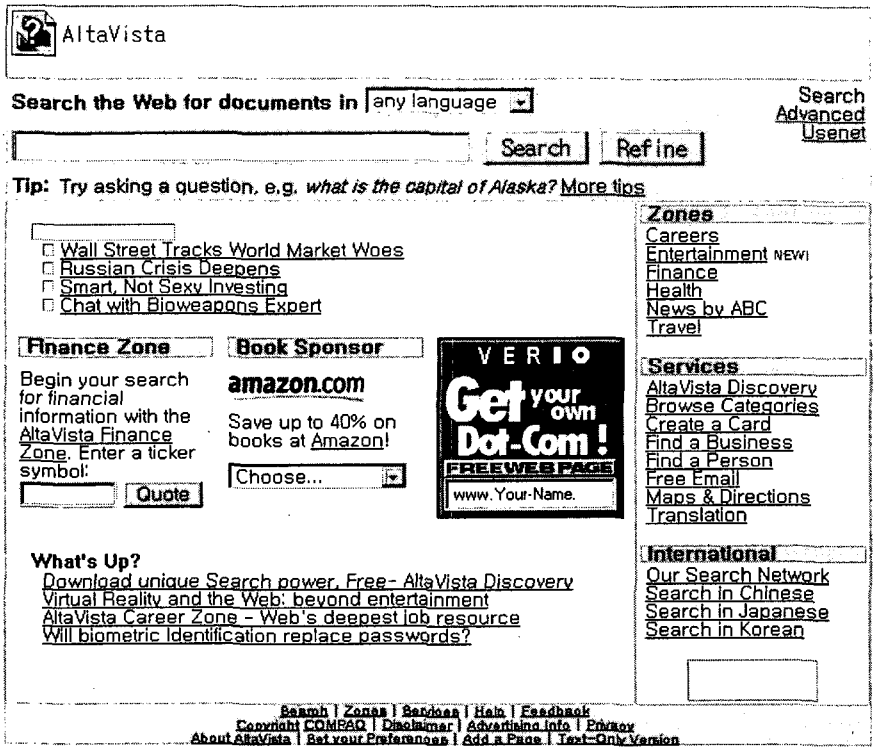
수행결과를 평가해 본 결과 전체적인 문장의 레이아웃은 만족할만하게 복원하였지만, 문장의 분리와 단어를 수식하는 Word_Start 태그와 Word_End 태그의 처리를 좀 더 보완하여야 한다.

참고문헌

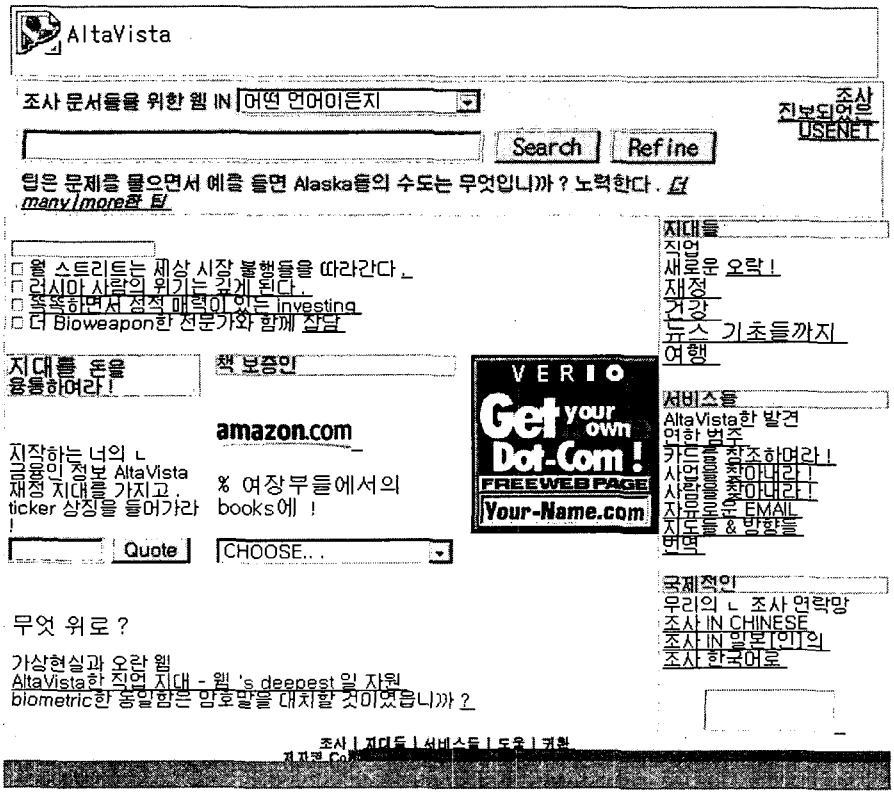
[1] 안동언, 유홍진, 서진원, 이영우, 정성종, 여상화, 김태완, 박동인, "웹용 영한 기계번역을 위한 문서 전처리기의 설계 및 구현", 1997년도 제 9회 한글 및 한국어 정보처리 학술대회, 1997, pp.249-254

[2] 심철민, 여상화, 정한민, 김태완, 박동인, 권혁철, "에서로-웹/EK : 영한 웹문서 번역 시스템," 1997년도 제 9회 한글 및 한국어 정보처리 학술대회, 1997, pp.277-282

[3] 여상화, 정한민, 채영숙, 김태완, 박동인, "실용적인 영한 기계번역을 위한 전처리기의 설계 및 구현," 1996년도 제 8회 한글 및 한국어 정보처리 학술대회, 1996, pp.313-319



<그림 2> 영어 웹 문서



<그림 3> Tag Manager 복원 문서