# 에서로-웹/EK™: 영한 웹 문서 번역 시스템

심철민*, 여상화*, 정한민*, 김태완*, 박동인*, 권혁철**
*시스템공학연구소 자연어정보처리연구부 기계번역연구실
**부산대학교 전자계산학과

{cmsim | shyuh | twkim | dipark}@seri.re.kr, hckwon@hyowon.pusan.ac.kr

# FromTo-Web/EK ™: English-to-Korean Machine Translation System for HTML Documents

Chul-Min Sim*, SangWha Yuh*, Hanmin Jung*, TaeWan Kim*, Dong-In Park*, Hyuk-Chul Kwon**
*Machine Translation Laboratory, Natural Language Information Processing Department, SERI
**Computer Science Department, Pusan National University

## 요 약

최근 들어 웹 상의 문서를 번역해 주는 번역 시스템이 상용화되고 있다. 일반 문서와 달리 웹 문서는 HTML 태그를 포함하고 있어 번역 시스템에서 문장 단위로 분리하는데 어려움이 있다. 또한 그 대상 영역이 제한되지 않으므로 미등록어 및 구문 분석 실패에 대한 대처 기능이 필요하다. 따라서 웹 문서의 번역 품질이 일반 문서 번역에 비해 현저히 떨어지게 된다.

이 논문에서는 HTML 태그를 보유한 영어 웹 문서를 대상으로 하는 번역 시스템인 "에서로-웹/EK"에 대해 기술한다. 에서로-웹/EK 는 HTML 문서의 특성을 고려하여 태그를 분리, 복원하는 태그 관리자를 별도로 가진다. 또한 태그를 유지하면서 영어에서 한국어로 변환되는 과정에서 발생하는 어휘 분리, 어휘 통합, 어순 변환 등의 다양한 변환 현상을 처리한다. 이 시스템은 변환 방식에 기반한 번역 시스템으로서 영어 해석, 영한 변환, 한국어 생성의 단계를 거친다. 구현된 시스템은 Netscape 와 DDE(Dynamic Data Exchange) 방식으로 연동하여 HTML 문서를 번역한다.

## 1. 서론

최근 들어 웹 문서를 번역하는 번역 시스템들이 상용화되고 있다. 특히 일한 번역 시스템의 경우는 실용화 수준에 이르고 있다. 그러나 영한 번역 시스템은 영어와 한국어의 어휘적 차이와 구조적 차이로 인해 그 개발이 어려우므로 상대적으로 실용화가 늦어지고 있다. 현재 상용화된 영한 번역 시스템의 경우 정형화된 일반 문서에 대해서는 번역 품질이 높지만 웹 상에 존재하는 태그를 포함하는 HTML 문서의 경우 번역의 질이 실용화에 미치지 못하고 있다. 그 대표적 이유로 다음을 들 수 있다.

- HTML 태그로 인해 문장 분리가 어렵다.
- 웹의 특성상 대상 영역이 무제한이므로 미등록어가 많이 발생한다.
- 명사의 나열이나 구나 절만으로 구성된 문장이 많이 존재한다.

이 논문에서는 정형화된 일반 문서를 대상으로 하는 영한 번역 시스템인 "에서로/EK"를 웹 문서를 처리하도록 개량한 "에서로-웹/EK"에 대해 기술한다. 이 시스템은 HTML 태그를 분리/복원하는 태그 관리기, 영어 문장에 대한 구구조를 생성해 내는 영어 해석기, 영어 구구조로부터 영어 의존구조를 생성하고 이를 한국어 의존구조를 거쳐 한국어 구구조로 변환하는 영한 변환기, 한국어 구구조로부터 한국어 문장을 생성해 내는 한국어 생성기, 웹 브라우저와 연동하기 위한 사용자 인터페이스로 구성된다.

## 2. 시스템 구성

"에서로-웹/EK"는 전형적인 변환(transfer) 방식의 기계 번역 시스템이다. 즉, 입력된 영어 문장에 대해 형태소 해석, 구문 해석을 거쳐 구구조 트리를 생성해 내고, 이를 의존구조로 변형한 후, 변형된 영어 의존구조를 한국어 의존구조로 변환한다. 변환된 한국어 의존구조를 한국어 구구조로 변형하고 이로부터 최종적으로 한국어 문장을 생성한다.

한편, "에서로-웹/EK"는 태그를 보유한 HTML 문서를 처리 대상으로 하므로 HTML 문서로부터 번역 단위인 영어 문장을 추출하고, 번역 결과 생성된 한국어 문장에 태그를 복원해 넣는 역할을 담당하는 태그 관리기가 별도로 존재한다. 또한 웹 문서 번역을 위해 웹 브라우저와 실시간에 자료를 주고받는 기능을 하는 사용자 인터페이스도 존재한다. <그림 1>은 "에서로-웹

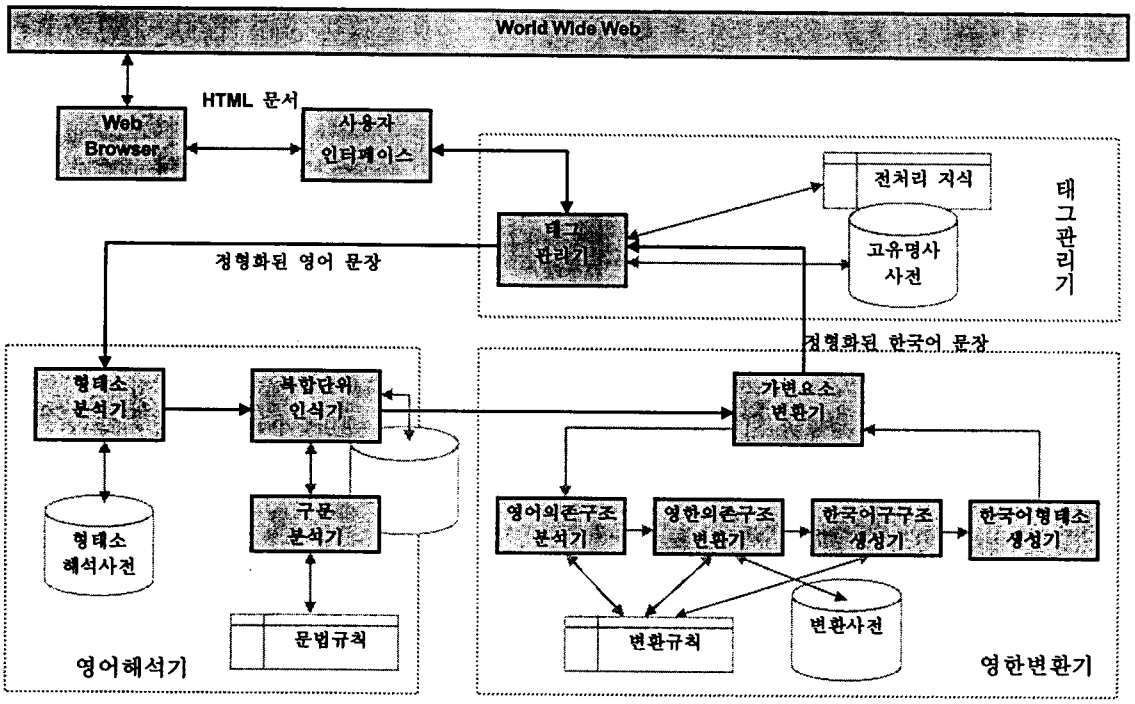



그림 1. "에서로-웹/EK"의 시스템 구성도

/EK"의 시스템 구성도이다.
"에서로-웹/EK"는 다음과 같은 단계를 거쳐 수행된다.

- 사용자 인터페이스 : 웹 브라우저를 실행하고 그로부터 HTML 문서를 전달 받아 번역한 후 번역된 HTML 문서를 다시 웹 브라우저를 통해 출력한다. 기타 Frame 처리, URL 복원 등의 역할을 수행한다.

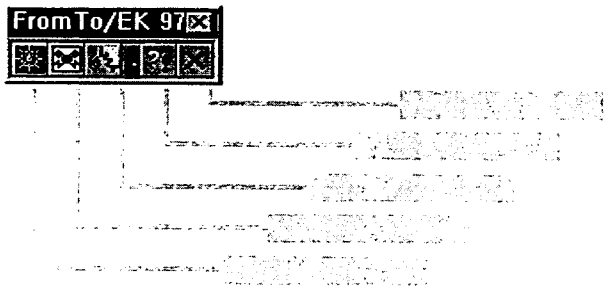


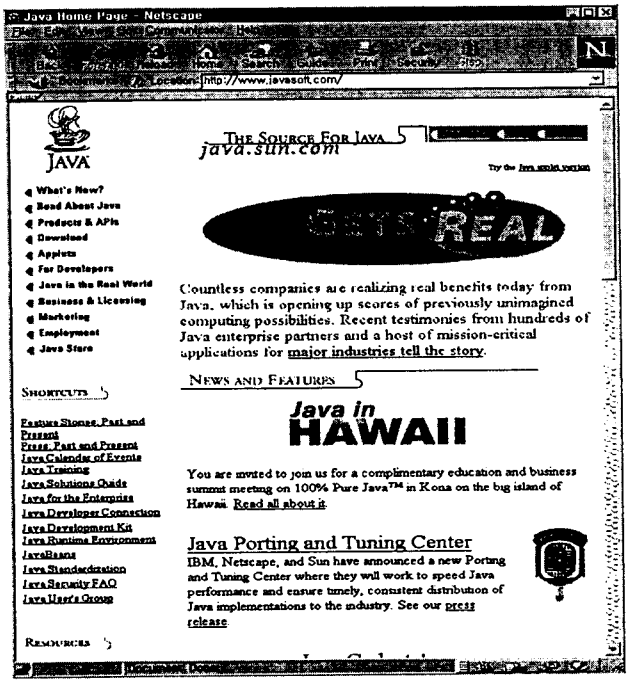

그림 2. "에서로-웹/EK"의 사용자 인터페이스

- 태그 관리기 : 사용자 인터페이스에서 넘겨준 HTML 문서를 분석하여 전처리 지식을 참조하여 날짜 등의 특수한 패턴을 처리하며, 고유명사 사전을 바탕으로 고유명사를 미리 인식하여 문서를 정규화한다. 정규화된 HTML 문서로부터 태그와 내용을 분리하여 문장 단위로 번역을 수행한다. 한 문장의 번역이 완료되면 태그를 복원하고, 문서 전체의 번역이 완료되면 번역된 HTML 문서를 생성하여 사용자 인터페이스에 반환한다.
- 영어 해석기 : 정규화되어 입력된 영어 문장에 대해 형태소 해석을 거쳐, 복합 단위 인식을 수행한다. 복합 단위 인식기는 복합 단위 사전에 존재하는 패턴에 대해 복합 단위와 가변 요소를 결정한다.

  다음으로 구문 분석기를 호출하여 가변 요소에 대한 구구조 서브 트리를 생성한 후 복합 단위에 대한 노드를 결합하여 문장 단위 구구조 트리를 완성한다. 구문 분석기에서 참조하는 문법 규칙은 ACFG(Augmented Context Free Grammar)에 기반한다.

- 영한 변환기 : 복합 단위 인식기의 인식 결과 생성된 영어 구구조 트리 및 가변 요소에 대한 서브 트리와 문장 단위의 구구조 트리를 입력받아 가변 요소 변환기에서는 하부 서브 시스템을 호출하여 영한 변환을 수행한다.

  복합 단위의 사이에 나타날 수 있는 가변 요소의 경우는 부분적인 영어 구구조 트리를 입력받아 GWL(Grammar Writing Language)로 기술된 변환 규칙 및 영한 변환 사전을 참조하여 영한 변환을 수행한다. 1 단계로 영어 구구조 트리를 영어 의존구조 트리로 변형하는 영어 의존구조 분석기가 실행되고, 다음으로 영어 의존구조를 변환 사전을 참조하여 한국어 의존구조로 변환하는 영한 의존구조 변환기, 한국어 의존구조를 한국어 구구조로 변형하는 한국어




```
;; Sent 6
  (("recent") ("testimonies") ("from") ("hundreds") ("of") ("java")
  ("enterprise") ("partners") ("and") ("a") ("host") ("of")
  ("mission-critical") ("applications") ("for")
  ("<a href=\"/features/1997/sept/real.html\">" :TAG HTML)
  ("major") ("industries") ("tell") ("the") ("story") ("</a>":TAG
  HTML) (".") ("</FONT>" :TAG HTML) ("</TD>" :TAG HTML))
;; Sent 7
  (.....
  ("<A HREF=\"/features/index.html\">" :TAG HTML)
  ("feature") ("Stories") (":") ("past") ("and") ("Present")
  ("</A>":TAG HTML) ("</FONT>" :TAG HTML))
;; Sent 8
  (("<BR>" :TAG HTML)
  ("<FONT SIZE=\"-1\">" :TAG HTML)
  ("<A HREF=\"/pr/archives.html\">" :TAG HTML)
  ("Press") (":") ("past") ("and") ("Present") ("</A>":TAG HTML)
  ("</FONT>" :TAG HTML))
;; Sent 9
  ( ("<BR>" :TAG HTML)
  ("<FONT SIZE=\"-1\">" :TAG HTML)
  ("<A HREF=\"/nav/new/calendar.html\">" :TAG HTML)
  ("java") ("calendar") ("of") ("Events") ("</A>":TAG HTML) )

6, ((recent, 1) (testimonies, 1) (from, 1) (hundreds, 1) (of, 1) (java, 1)
   (enterprise, 1) (partners, 1) (and, 1) (a, 1) (host, 1) (of, 1)
   (mission-critical, 1) (applications, 1) (for, 1) (major, 1) (industries, 1)
   (tell, 1) (the, 1) (story, 1) (., 6))
7, ((feature, 1) (Stories, 2) (:, 6) (past, 1) (and, 1) (Present, 2))
8, ((Press, 2) (:, 6) (past, 1) (and, 1) (Present, 2))
9, ((java, 1) (calendar, 2) (of, 1) (Events, 2))
```

그림 3. 태그 분리 및 문장 인식 예

구구조 생성기, 한국어 구구조를 깊이 우선으로 탐색하면서 한국어 형태소를 생성해 내는 한국어 형태소 생성기의 순서로 실행된다.

가변 요소 변환기에서는 복합 단위 사이의 여러 가변 요소들에 대해 한국어 형태소 생성을 수행한 후, 가변 요소를 내포하는 전체 문장에 대해 영어 의존구조 분석기, 영한 의존구조 변환기, 한국어 구구조 생성기, 한국어 형태소 생성기를 수행하여 번역문을 생성해 낸다.

## 3. HTML 태그 처리 방안

"에서로-웹/EK"는 HTML 문서를 처리 대상으로 하므로 HTML 태그를 고려해야 한다. 일한 번역의 경우 HTML 파서를 이용하여 문서 전체를 분석하는 방법과 태그와 태그 사이의 내용을 번역 단위로 하는 방법이 사용된다. 일본어와 한국어의 어순이 유사하므로 전자와 후자의 번역 품질이 큰 차이가 나지 않는다. 따라서 대부분의 시스템이 보다 간단하고 처리 속도가 빠른 후자의 방법을 사용한다. 그러나 영어와 한국어의 경우는 어휘적 차이 뿐만 아니라 구조적 차이가 존재하므로 일한 번역에서와 같이 후자의 방법을 사용할 경우 번역 품질이 실용화 수준에 미치지 못한다.

"에서로-웹/EK"의 태그 관리기는 HTML 문서 전체를 파싱하지 않고 휴리스틱을 이용한 분석 방법을 사용한다.

태그 분리

태그 분리 과정에서는 문서의 레이아웃 정보와 문장 부호 정보를 사용한다. 태그 관리기에서는 번역 엔진에서 처리할 수 있는 단위를 한 문장으로 본다. 태그 분리 및 문장 인식 전략은 다음과 같다.

- 순서가 중요한 태그(<A>, </A>, <TITLE>, </TITLE> 등)의 경우 쌍을 맞추어 문장을 분리한다.
- 태그와 태그 사이에 여러 문장이 존재할 경우 문장 부호 단위로 문장을 분리한다.
- 테이블의 내부 항목은 문장으로 간주한다.
- 리스트의 한 항목은 문장으로 간주한다.
- 복합 명사, 구나 절 이후 여백의 라인이 존재하면 제목으로 간주하여 문장으로 간주한다.

입력된 HTML 문서는 태그 분리 후 다음과 같은 정보가 추출되어 번역 엔진에 전달된다.

*<문장번호, 어절번호, 어절내용, 시작태그, 끝태그, 구별자>*

태그 분리 후 문장 인식 결과 생성된 문장 번호, 문장 내에서의 어절 순서, 실제 어절 내용, 어절의 앞 부분에 존재하는 태그, 뒷 부분에 존재하는 태그, 어절의 상태(숫자, 영어, 한글, 특수문자) 정보가 추출된다. <그림 3>은 javasoft 홈 페이지에 대해 태그 분리 및 문장 인식을 수행한 예이다.

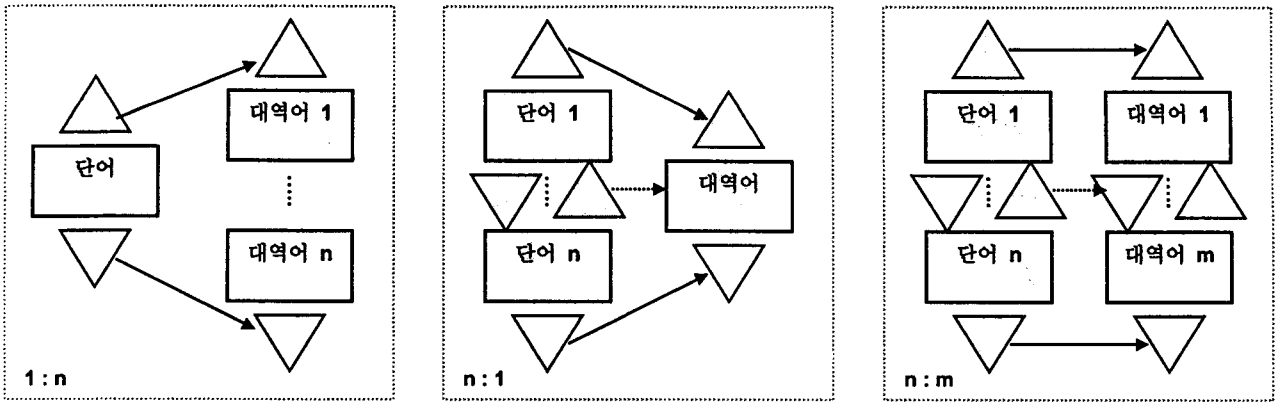



그림 4. 어휘 변환에 따른 태그 처리 전략

태그 복원

번역 과정에서는 어휘 변환은 복합 단위 인식기, 영한 의존구조 변환기에서 일어난다. 1 대 1로 어휘 변환이 일어날 경우 태그 처리는 문제가 발생하지 않는다. 그러나 1 대 n, n 대 1, 혹은 n 대 m으로 어휘 변환이 일어나는 경우 태그의 확장/축소가 필요해 진다. <그림 4>는 어휘 변환 형태에 따른 태그의 확장/축소 전략을 보여준다. 번역 실행 중 어휘 변환 단계에서는 어절 ID의 순서 정보만을 처리하며, 문장 전체의 번역 종료 후 태그 관리기에서 아래와 같은 전략을 이용하여 태그를 복원한다. 다음은 태그 복원기에서 어휘별로 태그를 처리하는 전략이다.

- 1 대 n : 원시 단어의 시작태그와 끝태그를 변환된 단어들의 시작과 끝으로 한다.
- n 대 1 : 원시 단어들의 시작태그와 끝태그를 분석하여 순서가 중요한 태그의 경우 대역어의 시작과 끝태그를 결정한다. 이 과정에서 내부 단어의 크기나 색깔 정보 등 레이아웃이나 앵커와 관련이 적은 태그들은 복원시키지 않는다.
- n 대 m : 이러한 어휘 변환은 주로 숙어나 복합단위가 존재할 경우에 발생한다. 이 경우 n 대 1의 경우와 유사하게 원시 단어들의 태그를 분석하여 변환된 단어들 전체의 시작과 끝태그를 결정한다. 이 과정에서 중요도가 떨어지는 태그들을 복원시키지 않는다.

번역 과정

영어와 한국어는 언어적인 차이로 인해 번역 과정에서 어순이 바뀐다. 태그 관리기에서 태그를 보다 정확히 복원하기 위해서는 어순의 변경에 관한 정보를 유지해야 한다. "에서로-웹/EK"에서는 어절별로 고유한 ID를 유지하여, 번역 과정에서 어휘 수나 어순의 변화가 발생했을 경우 이를 복제/통합하는 방식을 사용한다. 문장 전체의 번역이 종료되면 태그 관리기에서 어절 ID의 순서 정보를 이용하여 태그를 복원한다. 다음은 구조 변환의 처리 전략이다.

- 태그 관리기는 어절별로 고유한 어절 ID를 부여한다.
- 1 대 n의 경우, 어절 ID는 각각의 어절로 복제한다.
- n 대 1의 경우, 어절 ID도 서로 통합하며, 이 경우 ID의 순서를 유지한다.
- n 대 m의 경우, n개 어절의 ID를 통합한 후 이를 m개 어절에 복제한다.

다음 <그림 5>는 서브 시스템을 거치면서 어절 ID가 복제/통합되는 과정이다. 처리 단계는 다음과 같다.

① 입력 문장은 "Find out the latest Java technologies for the enterprise."이다.
② 태그 관리기에서 어절 단위로 고유한 어절 ID를 부여한다.
③ 숙어 및 복합 단위가 인식되면 해당하는 어절 ID를 통합한다.
④ 영어 구구조를 의존구조로 변형하면서 노드의 통합이 일어나고 이 경우에도 해당하는 어절 ID를 통합한다.
⑤ 앞 장의 어휘 변환에서 기술했던 변환 전략에 기반하여 영어 의존구조를 한국어 의존구조로 변형한다.
⑥ 한국어 의존구조로부터 한국어 구구조를 생성해 낸다. 이 단계에서는 조사나 어미 등의 노드가 새롭게 추가되며 이러한 노드는 앞의 명사나 용언의 어절 ID를 복제하여 자신의 어절 ID 값으로 한다.
⑦ 한국어 형태소 생성기에 의해 한국어 구구조로부터 한국어 어절 리스트가 생성된다. 번역이 종료한 후 동일한 어절 ID 열을 가지는 연속된 어절들을 통합한다.

## 4. 웹 브라우저와의 연동

이 논문에서 기술한 "에서로-웹/EK"는 대표적인 웹 브라우저인 Netscape와 Windows에서 제공하는 메시지 전달 방법인 DDE(Dynamic Data Exchange) 기법을 이용하여 동적으로 연동되는 번역 시스템이다. <그림 6>은 웹 브라우저인 Netscape와 "에서로-웹/EK"가 DDE API를 통해 연동되는 과정을 보여준다. "에서로-웹/EK"가 구동된 후 도구 버튼을 누름으로써 Netscape를 구동할 수 있다. 홈 페이지가 웹 브라우저로 다운로드된 후 번역 버튼을 누르면 ① 홈 페이지를 지역 디스크로 저장하고, ②

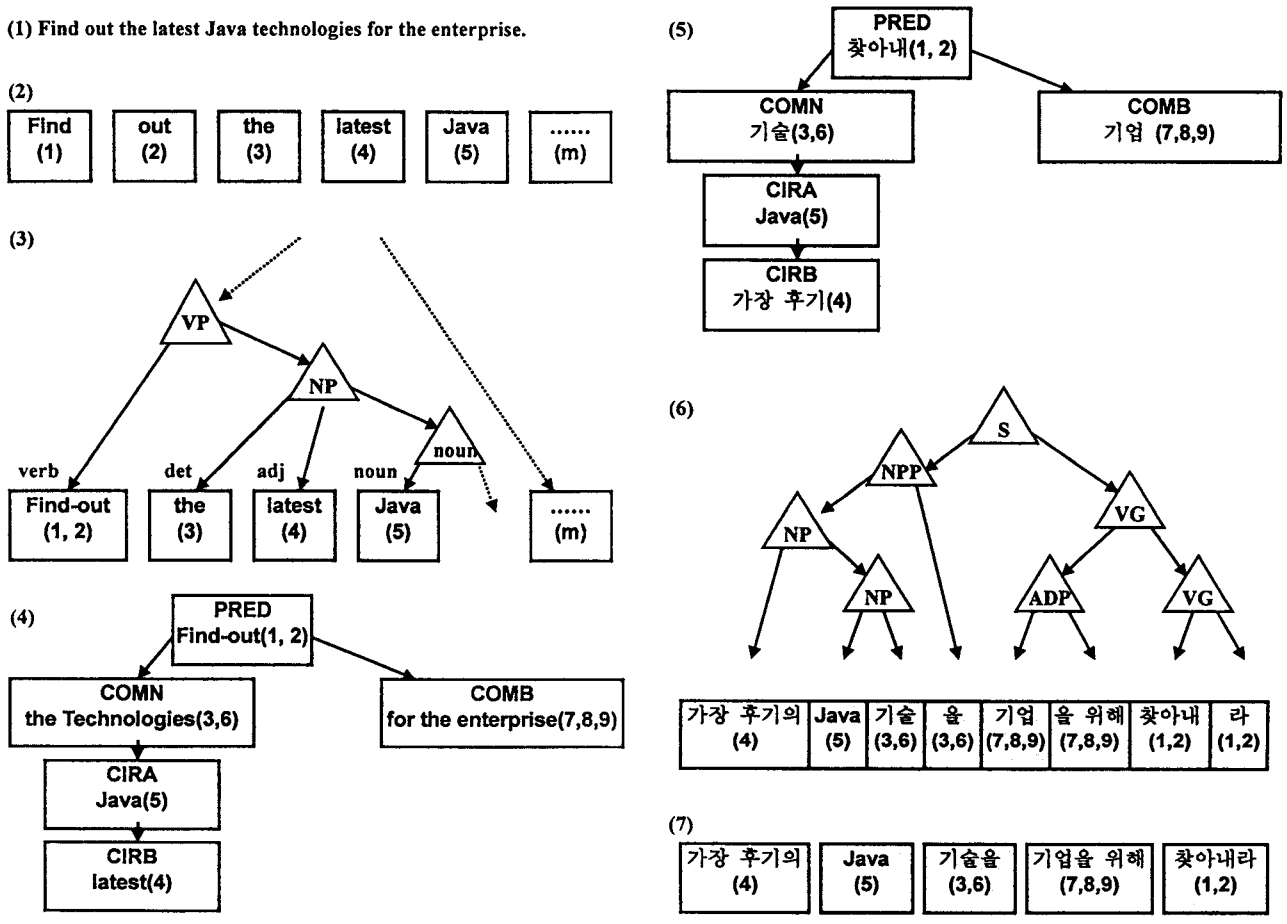



그림 5. 어절 ID의 처리 과정

태그 분리 및 문장 인식을 수행한 후, ③ 문장 단위로 번역을 실행한다. ④ 문장 번역 후 태그를 복원하고, ⑤③④ 과정을 반복하여 파일 전체가 번역되면 번역된 파일을 웹 브라우저로 업로드한다.

## 5. 결론

이 논문에서는 HTML 문서를 번역 대상으로 하는 영한 번역 시스템인 "에서로-웹/EK"에 대해 설명하였다. "에서로-웹/EK"는 웹 브라우저와 DDE 방식으로 연동하면서 HTML 문서를 번역하는 사용자 인터페이스와, HTML 문서로부터 태그를 분리하고 번역 단위인 문장을 인식하는 태그 관리기, 영어 형태소 해석 및 구문 해석을 수행하는 영어 해석기, 영어 해석 결과로부터 한국어를 생성해 내는 영한 변환기로 구성된다.

HTML 문서의 경우 번역 과정에서 태그 정보를 유지해야 한다.. "에서로-웹/EK"는 어절 ID를 통합/복제하는 방식을 사용하여 영한 변환 과정에서 발생하는 어휘 변환과 구조 변환에 따른 태그 정보 손실을 줄였다. 이에 따라 보다 정확한 번역 결과를 기대할 수 있게 된다.

향후 연구 방향은 다음과 같다.
첫째, 다양한 웹 문서 레이아웃을 분석하여 문장 인식의 정확도를 향상시켜야 한다.
둘째, 태그 복원의 정확도 향상을 위한 지식을 축적해야 한다.
셋째, 전문 분야 인식 및 전문 용어 사전을 이용하여 번역 품질을 향상시켜야 한다.
넷째, 번역 실패에 대한 대책(fail soften)이 마련되어야 한다.
다섯째, 실시간 온라인 번역을 제공하는 사용자 인터페이스를 제공해야 한다.

"에서로-웹/EK"는 보다 높은 번역 품질과 견고성을 제공하기 위하여 향후 PANGLOSS[Nirenburg, 1995]와 같은 다중 엔진을 사용할 계획이다. 사용자는 번역 시스템의 구동을 전혀 느끼지 못하면서 자신이 원하는 언어로 웹을 브라우징할 수 있게 될 것이다.

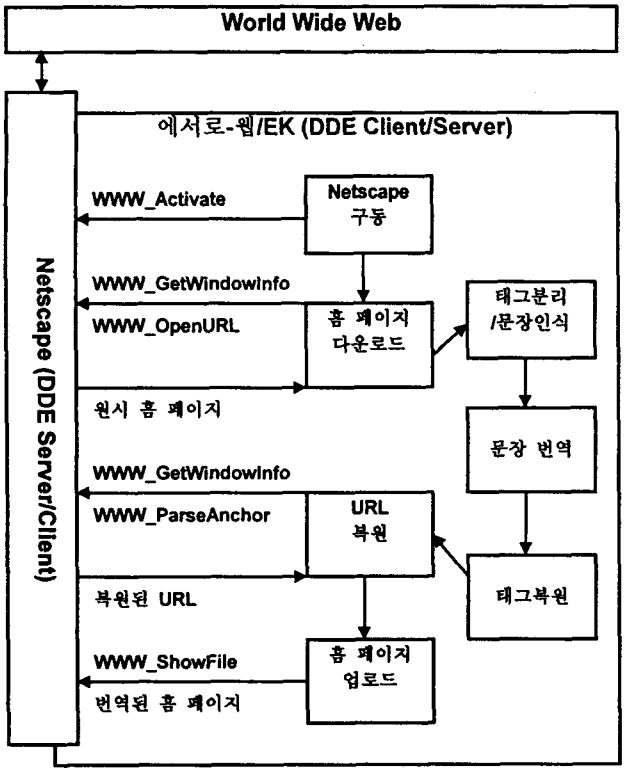



그림 6. Netscape 와 "에서로-웹/EK"의 DDE 연동

## 참고문헌

[김영식, 1989] 김영식, "영한 기계번역에 있어서 변환과 역어선택," 한국정보과학회 '89 가을 학술발표논문집, 한국정보과학회, pp567-570, 1989.

[이상엽, 1990] 이상엽, "영한 기계번역을 위한 영어 의미 해석 시스템의 설계 및 구현," 한국정보과학회 '90 가을 학술발표논문집, 한국정보과학회, pp237-240, 1990.

[최운천, 1990] 최운천, "영한 기계번역을 위한 한국어 생성기의 설계 및 구현," 한국정보과학회 '90 가을 학술발표논문집, 한국정보과학회, pp221-224, 1990.

[Choi, 1994] Key-Sun Choi, "An English-to-Korean Machine Translator : MATES/EK," Computational Linguistics. 14, 1994.

[Murata, 1995] T. Murata, "WWW machine translation system : W2-PENESE," Natural Language Processing 108-22, 1995.

[Nirenburg, 1995] S. Nirenburg, The PANGLOSS Mark III Machine Translation System, CMU-CMT-95-145, 1995.

[정한민 a, 1997] 정한민, "다국어 기계 번역 시스템을 위한 복합 단위 인식기," 한국정보처리학회 '97 춘계 학술발표논문집, 한국정보처리학회, pp415-418, 1997.

[정한민 b, 1997] 정한민, "에서로/EK : WWW 상에서의 번역 서버," 한국정보과학회 '97 봄 학술발표논문집, 한국정보과학회, pp475-478, 1997.