정렬 워크벤치의 설계 및 구현°

이재성*, 강정구*, 이주호*, Hung Le**, 최기선*
*한국과학기술원 전산학과, ** California Polytechnic State Univ.

The Design and Implementation of Alignment Workbench

Jae Sung Lee*, Jung-Goo Kang*, Ju Ho Lee*, Hung Le**, Key-Sun Choi*
*KAIST CS Dept., **California Polytechnic State Univ.

요약

통계적인 방법으로 병렬 코퍼스(parallel corpus)로부터 사전정보를 추출해 내는 정렬 시스템에 대한 연구가 세계 여러곳에서 진행되고 있다(신중호 1996; Dagan 1996; Fung 1995; Kupiec 1993). 그 결과로 만들어진 사전정보는 유용한 대역어와 대역 확률을 포함하고 있지만, 불필요하거나 잘못된 요소들도 많이 포함되어 있어 재조정 작업이 필요하다. 이는 사전정보를 직관적으로 확인함으로써 조정을 할 수도 있지만, 좀 더 정확한 조정을 위해 각각의 사전정보(정렬의 결과)가 코퍼스의 어떤 문장에서 나온 것인가 등을 확인할 필요가 있다. 정렬 워크벤치는 이와 같은 작업을 효율적으로 처리할 수 있도록 만들어졌으며, 현재 구현되어 작동되고 있다. 본 논문에서는 정렬 워크벤치를 위해 필요한 정렬시스템의 변형과 사전작업의 편의를 위해 제공되어져야 하는 기능 등에 관하여 설명하고, 간단한 평가 결과를 설명한다.

1. 서론

대량의 코퍼스가 구축되면서, 통계적 방법으로 새로운 어휘정보를 자동 추출하는 연구들이 진행되고 있다. 특히, 원어와 번역어가 동시에 구축되어 있는 병렬 코퍼스(parallel corpus; 또는 양국언어 코퍼스)는 통계적 정렬방법을 통해 통계적인 기계번역과 번역지식의 구축에 이용되고 있다(신중호 1996; Dagan 1996; Fung 1995; Kupiec 1993).

이러한 통계적인 정렬 방법은 대개 많은 양의 계산에 의해 저품질로된 대량의 지식을 구축하는데 편리하지만, 좀더 고품질의 지식을 구축하기에는 부족한 점이 있어서, 아직 좀더 연구가 필요하다. 현재로서는 저품질의 지식 정보를 우선 수동적인 방법을 통해 고품질화시킴으로서, 대량의 지식구축을 위한 방법으로 사용할 수 있을 것이다. 수동의 작업을 위해서는 우선 그 어휘정보를 바로 살펴보아 직관적인 판단을 할 수도 있겠지만, 좀더 정확한 판단을 위해서는 그 지식이 추출된 원본 코퍼스의 문장을 확인하는 것이 필요하다. 또한 이러한 판단을 근거로 사전을 쉽게 편집 및 관리할 수 있는 시스템이 필요하다.

본 논문에서는 정렬 결과로 만들어진 사전(이하 정렬사전)을 편집하고 출처 정보를 확인하기 편리한 정렬 워크벤치에 대해 소개한다. 정렬 워크벤치를 사용하기 위해서 우선 필요한 출처정보를 만들기 위해 정렬 시스템을 일부 수정하였고, 그 결과로 만들어진 정렬사전에 대해 수정 작업을 할 수 있도록 설계하였다. 우선, 정렬시스템이 만들어낸 정렬사전의 문제점을 2 절에서 분석하고, 이러한 문제를 해결하기 위한 연구나 이와 관련된 연구를 3 절에서 설명한다. 4 절에서는 정렬 워크벤치의 전체적인 작동환경을 소개하고, 5, 6, 7 절에서 그 기본 동작과 정렬사전의 구조, 실제 수행하고 평가한 결과에 대해 차례로 설명한다.

2. 정렬 사전의 문제점

정렬사전은 유용한 정보와 함께, 불필요한 정보나 잘못된 정보가 포함되어 있다. 그림 1 은 실제 신중호(1995)의 정렬시스템에 의해 만들어진 사전의 예이다. 이 그림의 (1)과 (2)는 "가격"이 "price"와 "cost"로 번역되며, 각각의 대역어로 정렬된 확률이 0.153801 와 0.12462 임을 나타낸다. 이 경우는 비교적 적절한 정렬의 예가 된다.

하지만 정렬 사전을 살펴보면 잘못된 정렬이 많이 포함되어 있다. 예를 들어 "국제어"의 경우, 원어가 두 개의 단어로 분리되어 정렬하기 보다는 하나로 정렬하는 것이 더 올바른 결과를 얻을 수 있을 것이다. 즉 (3)과 (4)의 대역어가 하나로 합쳐져서 "international language" 로 되어야 한다. "국제어"의 또 다른 정렬인 (5)와 (6)은 잘못된 정렬이다. 이와는 반대로 영어

° 이 논문은 1996 년도 한국학술진흥재단의 외국석학과의 공동연구과제 연구비에 의하여 연구되었음

"amphitheater"는 "원형 경기장"으로 번역되어져야 하지만, 원어를 한 단

원어	대역어	대역확률	
가격	price	0.153801	..... (1)
가격	cost	0.124624	..... (2)
국제어	international	0.273192	..... (3)
국제어	language	0.261181	..... (4)
국제어	an	0.201413	..... (5)
국제어	provided	0.135824	..... (6)
원형	amphitheater	0.156269	..... (7)
경기장	stadium	0.335752	..... (8)
ㄴ가	something	0.201688	..... (9)
ㄴ가	would	0.178573	..... (10)
ㄴ가	you	0.107707	..... (11)

그림 1. 정렬사전의 일부 (편의상 순서를 재조정함)

어로 국한시켜 정렬하였기 때문에 제대로 된 결과를 찾지 못하고 일부의 단어로만 정렬되었다.

정렬의 사전을 보고 대략 추정할 수 있는 것도 있지만, 실제 사용된 문장을 보는 것이 더 정확한 정렬을 할 수 있다. 예를 들면 "ㄴ가"와 같은 경우, 실제 문장 예를 보지 않고서는 정확한 정렬이 되었는지를 추측하기 어렵다. "ㄴ가"는 태거에 의해 분석되어 나온 종결어미(ef)로서 그림 1 의 (9), (10), (11)처럼 3 가지로 정렬되었다. (9)의 경우에 해당하는 문장을 하나 찾아 보면 다음과 같다. (이 문장예는 정렬의 전처리 과정으로 태거를 사용하여 태깅된 문장들이다. 여기에서 사용된 태그는 한국어의 경우 김재훈(1994)의 태그이고, 영어의 경우 Penn Treebank (Marcus et. al, 1993)에서 사용된 태그이다.)

호진/npd+은/jx 무엇/npd+이/jcp+ㄴ 가/ef 들/pv+어/ecs 올리/pv+었/efp+다/ef ./s.	Ho-chin/IN held/VBN up/IN something/NN ./S.

이 경우, "ㄴ가"는 "something"으로 정렬되어 있지만, 사실상 "무엇+이+ㄴ가" 가 모두 "something"으로 정렬되어야 함을 알 수 있다. (10)의 경우의 문장 예는 다음과 같으며 "ㄴ가"가 "would"와 적절하게 정렬된 것으로 보인다.

누가/npd	Who/WP
아테네/nq+까지/jca 뛰/pv+어/ecx+가 /px+서/ecs 그/npp+들/xn+에게 /jca 소식/nc+을/jc 전하/pv+고/ecx 싶/px+ㄴ가/ef ?"./sy	would/MD like/VB to/TO run/VB to/TO Athens/NNP and/CC tell/VB them/PRP the/DT news/NN ?/S.

(11)의 경우에 해당되는 문장 예는 다음과 같은 데, 이 경우 "ㄴ가"가 "you"로 정렬된다고 보기는 어렵다.

아름답/pa+ㄴ/exm 여인/nc+아/jcv ,/s, 그대/npp+는/jx 누구/npp+이/jcp+ㄴ 가/ef ?/sy	Who/WP are/VBP you/PRP ,/, beautiful/JJ lady/NN ?/S.

이 사전의 예에서 알 수 있듯이, 잘못된 항목의 유형은 대략 다음과 같은 것이 있다.

유형 1: 불필요한 원어 표제어가 포함되어 있다.
유형 2: 잘못된 역어가 포함되어 있다
유형 3: 합쳐져야 할 원어(역어)가 분리되어 있다.
유형 4: 분리되어야 할 원어(역어)가 합쳐져 있다.
유형 5: 위와 같은 잘못된 항목으로 인해 잘못된 번역확률이 포함되어 있다.

이와같이 정렬의 결과를 확인하고 분석하기 위해서는 정렬 결과와 관련 문장을 쉽게 연결시켜 볼 수 있어야 한다. 또한 필요한 경우, 정렬 사전의 원어 및 대역어를 직접 삭제, 수정, 추가하여 원하는 사전으로 만들 수 있으면 편리할 것이다.

3. 관련 연구

잘못된 정렬의 결과는 정렬 시스템을 수정함으로써 향상시킬 수도 있지만, 엄청난 양의 코퍼스가 필요하게 되고 현실적으로 그 한계가 있다. Wu(1994)는 기존의 사전을 정렬의 초기 사전으로 삼아 작은 코퍼스에서 정렬이 가능하도록 했다. 또, Fung(1995)과 Kupiec(1993)은 실용성있는 정렬사전을 만들 경우, 중요한 항목은 주로 명사나 명사구인 것에 착안하여 전처리단계로 태깅을 하여 명사나 명사구에 한하여 정렬을 하였다. 이러한 연구에서는 전처리 또는 후처리 단계에서 정렬사전을 수작업처리해야 할 필요가 있게 된다.

Termight (Ido 1994)는 기술 용어의 번역이 올바르게 되었는가를 확인하기 위해 만들어진 전문 번역가용 워크벤치이다. 이 시스템은 원어의 용어들을 우선 나열한 후, 그에 대응되는 역어들을 찾아낸다. 첫단계를 위해서는 우선 태거를 사용하여 대부분 기술용어의 대상이 되는 복합 명사구(multiword noun phrase)를 찾아낸다. 이 단계에서도 태거의 결과로 나온 대상명사구를 사람이 쉽게 편집할 수 있는 환경을 제공한다. 다음 단계로 정렬 프로그램을 이용하여 복합 명사구에 대응되는 역어들을 찾아 내고, 그 역어들 중 올바른 것들만을 골라 번역 용어집을 만들 수 있도록 했다.

Termight가 주로 명사구에 대해 작동할 수 있는데 반해 본 정렬시스템은 어떠한 품사에 대해서도 모두 정렬하고 그 정렬 결과를 확인해 볼 수 있도록 했다. 이 정렬 시스템은 Termight의 기능뿐만 아니라, 정렬 시스템에 대한 분석이나 다른 품사들의 정렬을 파악해 낼 수 있을 것이다.

4. 정렬 워크벤치의 작동 환경

여기에서 사용하고 있는 정렬시스템은 문장단위 정렬 코퍼스를 각각 형태소 해석 및 태깅을 한 다음, 이 결과를 가지고 한/영 정렬을 하였다(신중호 95). 정렬 결과는 여러가지 형태의 사전으로 생성되어 나오며, 그 사전에는 대역구 사전, 단어 대역 사전, 위치 및 기능어 대역 사전이 나온다. 각각의 사전은 모두 확률을 포함하고 있다. 그러나 이러한 사전에는 그 단어 또는 구(phrase) 나 기능어 등이 실제 코퍼스 문장의 어떤 부분에서 나온 것인지를 포함하고 있지 않다. 따라서 이러한 정보를 포함할 수 있도록 정렬 시스템이 수정되었다.

출처정보로는 원문 문장 번호, 원어 단어 위치, 역문 문장 번호, 역어 단어 위치를 사용한다. 정렬과정에서 이정보는 계속적으로 사용될 필요는 없으므로, 초기에 일단 값을 가지고 있다가, 최종 정렬쌍들이 정해지면, 이 정렬쌍에 추가한다. 실제 정렬가능한 모든 쌍에 대해 출처정보를 가질 경우, 많은 저장 공간이 필요하게 되므로 제대로 확인된 정렬쌍에 대한 정보만을 갖거나, 관심있는 정렬쌍에 대해서만 출처정보를 제공하는 것이 좀더 효율적일 것이다. 여기에서 관심있는 정렬쌍이란, 예를 들어, 명사들만의 정렬이거나, 동사, 조사 등 관심이 있는 품사 등을 제한하여 정렬시키는 것을 말한다. 출처정보를 처리하는 과정을 그림으로 나타낸 것이 그림 2 이다.

정렬 워크벤치가 다른 정렬시스템의 결과로 나온 정렬사전도 처리할 수 있도록 정렬사전의 표준형식을 지정하여 사용하고 있다. 이를 정렬워크벤치 사전형식으로 부르고 약어로서 AWD 형식으로 정하고 있다. (이에 대한 자세한 구조는 6 절을 참조하기 바람.) 이 AWD 형식으로 작성된 사전이 만들어 질 경우, 정렬 워크벤치는 대역언어의 사전 구축을 위한 도구로 사용될 수 있다. 또한 역으로 번역문이 올바르게 번역되었는지를 검사할 수 있는 시스템으로도 사용될 수 있다. 즉, 어떤 중요단어가 일관되게 대역어로 사용되고 있는지, 혹은 문맥에 따라 적절히 구사되고 있는지를 확인해 볼 수 있다.

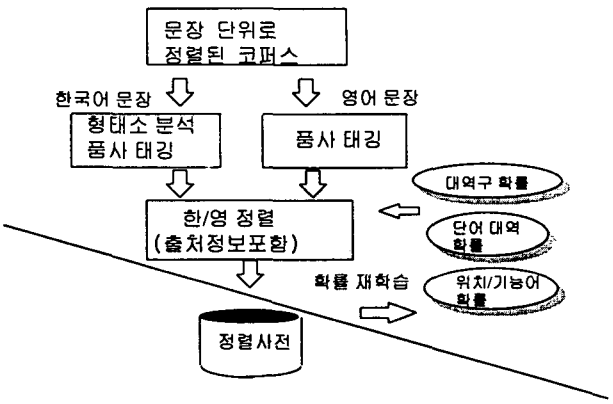

그림 2. 워크벤치 정보를 생성하는 정렬시스템

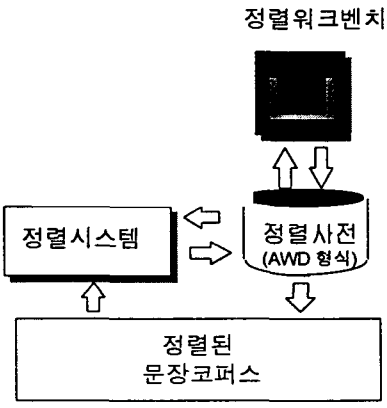

그림 3. 정렬워크벤치의 작동 환경

5. 기본 동작

정렬워크벤치(AWB)는 마이크로소프트 Visual C++로 구현되었고, 한글 윈도우 95 에서 작동된다. 사용자가 AWB를 수행시키면, 그림 4와 같이 초기 화면이 나타난다. 이 화면의 메뉴에서 화일-열기를 선택하여 AWD 형식으로 된 정렬사전을 연다. 화일이 열리면, 화면의 "원어" combo box에 원어 항목들이 나타난다. 사용자가 원하는 원어를 한 항목 클릭하여 선택하면, 대역어들이 대역확률과 함께 "대역어" combo box와 "확률" combo box에 나타난다. 사용자가 또 다시 대역어를 선택하면 원어와 대역어가 사용된 양국언어 번역예가 "번역문 대조" 상자에 나타난다. 각각의 문장 앞에는

그 문장의 번호가 표시되고, 해당되는 원어와 대역어는 밑줄과 붉은색으로 표시되어 구분을 쉽게 하도록 한다. 원어 "날씨"와 그의 한 대역어인 "weather" 에 대해 정렬된 문장예들을 확인하고 있는 화면이 그림 5 이다.

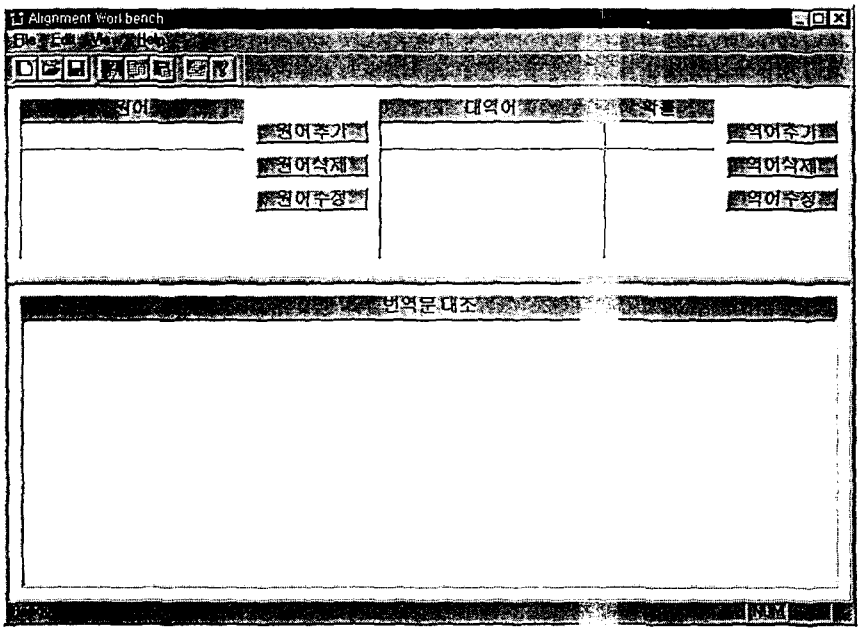

그림 4. 정렬워크벤치의 초기 화면

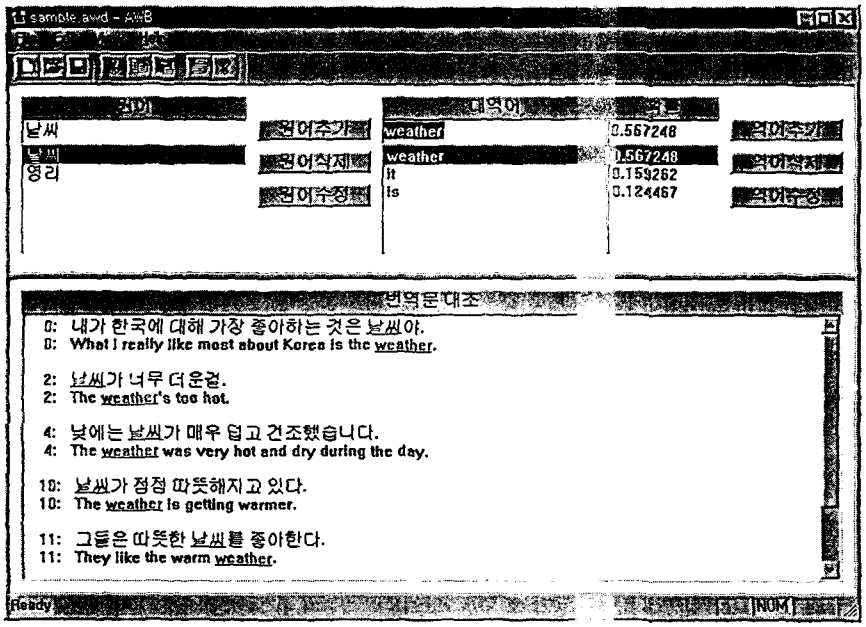

그림 5. "날씨" 와 "weather"로 정렬된 문장예를 확인하는 화면

새로운 항목을 추가하고자 할 경우는 바로 combo box의 편집 영역에 새로운 항목을 입력한 후, "원어(역어) 추가" 버튼을 누르면 된다. 원어 항목(또는 대역어 항목)중 삭제나 수정이 필요한 경우, 우선 그 항목을 선택한 후, 바로 "원어(역어) 삭제" 버튼을 눌러 삭제하거나, combo box의 편집 영역에서 내용을 수정한 후, "원어(역어) 수정" 버튼을 눌려 내용이 수정되도록 한다. 대역어 확률은 새로 추가되거나 수정된 항목에 대해서는 정의되지 않은 값으로 표시된다.

원어를 삭제할 경우, 관련된 역어가 모두 한꺼번에 삭제되며, 반대로 역어를 삭제할 경우에는 그 단어만 삭제된다. 만약 한 원어에 대한 마지막으로 남은 하나의 역어가 삭제될 경우에는 그 원어도 한꺼번에 삭제된다.

6. AWD 사전 구조

사전은 원어(<SWORD>)의 나열과 그에 대한 대역어(<TWORD>) 나열, 그리고 문장예(<SSENT> 및 <TSENT>)의 나열로 구성된다. 정렬사전의 구조를 도표로 표시하면 그림 6과 같다. 각각의 항목에 대한 설명은 다음과 같다. 우선 사전의 맨앞에는 사전의 헤더 정보로서 각각의 데이터가 얼마크기로 저장되었는지를 다음과 같은 파라미터에 저장한다.

SWORD
원어의 단어 갯수
TWORD
역어의 단어 갯수

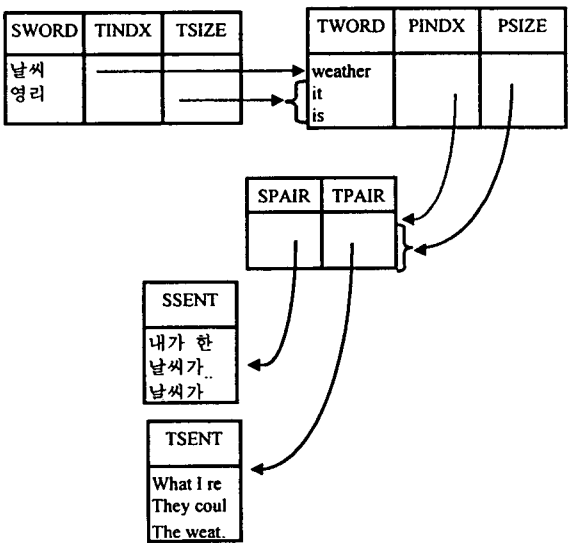

그림 6. 정렬사전의 연결 구조

SPAIR
정렬쌍 문장의 갯수
SSENT
원어 문장의 총 갯수
TSENT
역어 문장의 총 갯수

이 헤더 다음에 실제 그림에서 설명한 사전 구조를 표현하기 위한 데이터가 저장된다. 각 데이터의 구분은 태그로 구분되며, 그 태그 뒤에 실제 데이터가 저장된다. 각 태그에 대한 설명은 다음과 같다.

<SWORD>
원어 단어의 문자열이다.
<TINDX>
역어 단어의 시작 위치(인덱스)를 나타낸다.
<TSIZE>
대응되는 역어 단어의 갯수를 나타낸다.
<TWORD>
대역어의 문자열을 나타낸다.
<TPROB>
앞의 원어에 대한 대역어의 번역 확률을 나타낸다.
<PINDX>
정렬쌍에 대한 문장예의 시작 위치(인덱스)를 나타낸다.
<PSIZE>
정렬쌍에 대한 문장예의 갯수를 나타낸다.
<SPAIR>
원어 문장예의 인덱스 순서를 나타낸다.
<TPAIR>
역어 문장예의 인덱스 순서를 나타낸다.
<SPOS>
원어 문장예에서 단어의 위치를 나타낸다. 위치는 문장 첫글자로부터의 글자(1 바이트 단위) 갯수이며, 문장의 첫글자는 0 이 된다.
<TPOS>
역어 문장예에서 단어의 위치를 나타낸다. 위치는 문장 첫글자로부터의 글자(1 바이트 단위) 갯수이며, 문장의 첫글자는 0 이 된다.
<SSENT>
원어 문장의 문자열 또는 원어문장이 포함된 화일이름이다.
<TSENT>
역어 문장의 문자열 또는 역어문장이 포함된 화일이름이다.

7. 평가 및 맺음말

한국어 영어 정렬 시스템(신중호 1995)의 실험결과에 따르면, 각 단어에 대한 대역단어의 정확도는 학습 코퍼스에서 단어의 발생 빈도에 비례했으며, 모델에 따라 그 정확도에도 차이가 났다. 실제 추출된 단어들 중

에서 비교적 빈도가 많았던 실험용 단어들(15-16 번발생)도 정확도가 약 83%정도이다. 그러나 빈도가 적은 단어는 이보다도 훨씬 그 정확도가 떨어져서 약 60%정도 였고, 따라서 고품질 사전으로 변환하는데는 수동 작업이 많이 필요했다. 이런 문제는 현재 구축된 병렬 코퍼스의 양이 많지 않았기 때문에 발생되는 것이라고도 볼 수 있으며, 좀더 많은 병렬 코퍼스를 이용하여 빈도수가 많아 질 경우 그 정확도가 증가하여 수동 작업의 양을 줄일 수 있을 것이다. 이 과정에서 정렬 워크벤치는 번역된 단어가 쓰인 실제 문장의 예를 보여줌으로써, 정렬시스템이 어떻게 잘못된 역어를 생성하게 되었는가를 확인할 수 있게 해주었다.

현재 정렬 워크벤치는 원어나 역어의 삭제 기능만을 지원하고 있으며, 추가, 수정의 기능은 나중에 추가할 계획이다. 정렬시스템도 아직은 초기 단계이므로 복합어에 대한 처리나 구(phrase)처리 등이 미약하여 제대로 된 정렬사전을 구축하는데는 문제가 있다. 따라서, 정렬 워크벤치는 아직은 정렬 시스템의 성능을 분석하고 확인하는 수준으로 쓰이고 있다.

정렬 워크벤치를 다른 방향으로 이용하면, 양국언어의 용례를 찾아 보는 시스템으로 이용할 수 있고, 또는 번역문을 확인하는 시스템으로도 이용할 수 있을 것이다. 이를 위해서는 우선 미리 정의된 표준 양국언어 사전이 필요하다. 이를 이용하여 원문과 번역문에 대해 사전에서 제공된 단어만을 정렬시스템을 이용하여 정렬시키고, 그 결과를 정렬 워크벤치로 확인함으로써, 관심있는 단어들이 어떻게 쓰였는지를 실제 원문과 번역문에서 확인해 볼 수 있을 것이다.

현재 정렬 워크벤치는 화일을 중심으로 비교적 작은양의 데이터를 처리하고 있으나, 대용량을 처리할 수 있도록 데이터 베이스와 연결작업을 하고 있다. 또, 현재 정렬시스템이 워크벤치의 외부에서 동작하도록 되어있으나, 필요한 경우, 이를 워크벤치내에서 동작시킬 수 있도록 할 계획이다.

참고문헌

김재훈, 서정연(1994), "자연언어 처리를 위한 한국어 품사 태그," 한국과학기술원, 인공지능연구센타, CAIR-TR-94-55, 1994.

신중호 (1996), "한국어/영어 병렬 코퍼스에 대한 단어단위 및 구단위 정렬 모델," 석사학위논문, 한국과학기술원.

P. F. Brown, and et al. (1990), "A Statistical Approach to Machine Translation," Computational Linguistics, Vol 16, Num. 2, June.

P. F. Brown, and et al. (1993), "The Mathematics of Statistical Machine Translation: Parameter Estimation," Computational Linguistics, Vol 19, Num. 2.

I. Dagan, et al. (1994) "Termight: Identifying and Translation Technical Terminology," ACL ANLP.

I. Dagan. (1996), "Bilingual Word Alignment and Lexicon Construction," ACL96 Tutorial, June.

P. Fung (1995), "A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora," ACL95.

J. Kupiec (1993), "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora," ACL.

M. P. Marcus, and et al. (1993), "Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics.

D. Wu and X. Xia (1994), "Learning An English-Chinese Lexicon from a parallel corpus," in Proceedings of Association for Machine Translation in the America. 206-213.