2001 년 한국정보처리학회 춘계 학술발표논문집 제 8 권 제 1 호

한국어-일본어 정렬 기법 연구

김태완*
*인제대학교 정보컴퓨터공학부
e-mail : twkim@ijnc.inje.ac.kr

Aligning Word Correspondence in Korean-Japanese Parallel Texts

Taewan KIM*
*School of Information & Computer Engineering, Inje University

요 약

병렬 코퍼스의 확보가 과거에 비해 용이하게 됨에 따라 기계번역, 다국어 정보 검색 등 언어처리 시스템에 사용하기 위한 대역 사전 구축의 도구로서 정렬(Alignment) 기법에 대한 연구가 필요하다. 본 논문에서는 한국어-일본어 병렬 코퍼스를 이용한 정렬 기법에 관하여 제안한다.

1. 서론

기계번역, 다국어 정보 검색 등 각종 언어 처리 시스템은 그 기능을 수행하기 위한 어휘 지식의 집합체, 즉 사전이 필요하다. 사전의 개발에는 인간의 언어 능력이 필요하다. 따라서, 지금까지 전적으로 인간의 수작업에 의존하여 사전이 개발되어 왔으며, 막대한 인력, 비용, 시간이 필요한 전통적인 고비용 저효율 작업으로 인식되어 왔다. 또한, 수작업에 의해 개발된 사전은 실생활의 보편적인 언어 사용 현상을 반영하지 못한, 단순 사전적인 정보의 집합체에 지나지 않아, 실제 사용을 위해서는 상당한 기간의 수정 및 보수가 필요하다. 언어 처리 시스템이 실용화되지 못하고 늘 소규모 연구용 시스템 개발에 그쳐 온 가장 큰 이유가 여기에 있다. 따라서, 실생활의 보편적인 언어 사용 현상을 반영하면서 사전을 자동적으로 또는 반자동적으로 구축하여 주는 방법에 대한 연구가 필요하다.

근래에 주목받고 있는 언어 간 자동 정렬에 대한 연구([1],[2],[5],[6],[7])도 바로 이러한 사전 자동 생성의 필요성을 인식한 것으로부터 비롯되었다고 할 수 있다.

2. 관련 연구

정렬에 관한 기존의 연구에는 정렬의 단위에 따라 문장 단위 정렬, 명사구 정렬, 단어 정렬, 문자 정렬, 파싱 트리 정렬 등이 있다. 정렬 단위가 작을수록 대상 언어 간의 기본 단위의 상이성 문제가 대두된다. 예를 들어 문자 단위의 정렬(Church 1993)의 경우 영어와 불어와 같이 유사한 알파벳을 사용하고 많은 단어들이 어원이 같아 비슷한 형태를 갖는 언어들 사이에서만 가능하고, 한국어와 영어와 같이 문자 단위에서의 관련성이 없는 경우는 적용하기 어렵다. 파싱된 트리나 구와 같이 상위 단위의 경우에는 언어에 관계없이 일반적인 단위를 갖게 되므로 기본단위의 상이성 문제를 극복하기 쉬워진다. 그러나, 상위 단위로 갈수록 정확한 정렬 단위의 추출이 어려우므로 정렬 오류의 원인이 될 가능성도 증대된다. 예를 들어 파싱된 트리 간의 정렬을 수행할 경우 정확하게 파싱된 트리가 선택되어야 한다는 문제와 파싱된 트리 간의 호환성 문제를 안고 있다.

[6]에서는 영어와 불어 간에 문장 정렬된 2,600 문장에 대해 명사구의 정렬을 시도하였다. 각 언어에 대한 태거와 명사구 인식기를 사용하여 각각의 명사구 인덱스를 추출한 후, 추출된 각 언어별 명사구 인덱스와 정렬된 코퍼스를 대상으로 Mapping Algorithm 을 적용하여 명사구의 정렬을 수행한다. [3]에서는 영어와 불어 단어간의 정렬을 위해 영어 텍스트와 불어 텍스트를 K 개의 부분으로 나눈 후, 특정 영어 단어와 불어 단어가 각 부분에 나타나는가 아닌가를 분석하여 텍스트 내에서의 distribution 을 파악할 수 있게 한다. 이것은 K-dimensional binary vector 로 표시할 수 있다. 이로부터 contigency matrix 를 만든 후, Mutual Information 식을 사용하였다. [5]에서도 이와 유사한 방식을 취하고 있으나 Mutual Information 을 사용하지 않고 영어 단어와 불어 단어 간의 유사도 계산에 Dice coefficient 를 사용하여 정렬을 시도하였다.

3. 일본어와 한국어의 비교

일본어와 한국어는 한자어를 많이 사용한다. 특히, 명사의 경우는 거의가 한자어이다. 단, 일본어에서는 한자어를 한자로 표기하는데 반해 한국어에서는 한글 전용 관습으로 인해 한자어의 한국어 음독을 한글로 표기한다. 이것은 일본어와 한국어의 정렬에 있어서 훌륭한 단서가 된다. 따라서, 일본어 한자 코드를 동일한 한글 한자 코드로 변환하여 주는 기능과 해당 한자의 한국어 음독을 한글로 표기하여 주는 기능이 있다면 손쉽게 해당 한자어간의

정렬이 가능하다.

4. 일본어 한국어간 자동 정렬

일본어 한국어 자동 정렬 과정을 그림 1 에 보인다. 그림에서 보인 바와 같이 제안된 일한 자동 정렬 방법은 크게 다음과 같이 4 단계로 일한 자동 정렬을 수행한다.

① 문장 정렬된 일본어와 한국어 코퍼스 전처리 단계

② 일본어와 한국어는 개념어, 특히 명사의 경우 동일한 한자어를 사용한다는 유사성을 이용하여 1 차 정렬을 수행하고, 정렬된 일본어, 한국어 한자어를 anchor 로 하여 문장 정렬 상태의 parallel corpus 를 문장 보다 세분화된 segment 단위 정렬 corpus 로 변환하는 단계

③ segment 단위로 정렬되고 정규화된 각 언어 corpus 로부터 정렬 대상들을 자동 추출하는 단계

④ 자동 추출된 정렬 대상들의 segment 출현 유사도를 이용한 2 차 정렬 수행

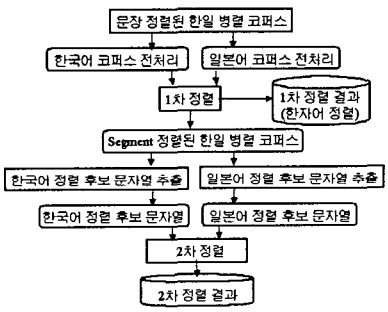


그림 2. 한국어-일본어 정렬 개념도

4.1. 전처리

4.1.1. 일본어 코퍼스 전처리

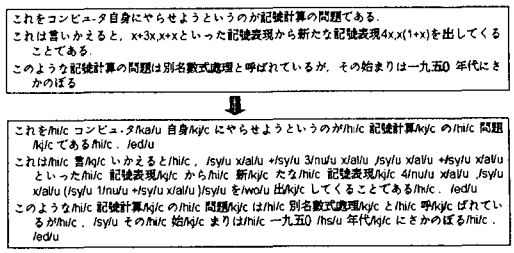
정렬의 효율성을 높이기 위해 일본어 전처리기에서는 다음과 같은 기능을 수행한다.

- 자종에 따라 일본어 코퍼스를 동일 자종 문자열 단위로 분리
- 분리된 문자열의 자종이 무엇인지를 표기
- 분리된 문자열이 1 개 단위로 취급될 수 있는지, 아니면 하위 문자열로 분리될 수 있는 것인지를 표기

1 개 단위로 취급할 수 있는 것에는 다음과 같은 것이 있다.

- 한자 표기 숫자로 구성된 문자열
- 알파벳으로 구성된 문자열
- 가다가나로 구성된 문자열
- 기호로 구성된 문자열

일본어 코퍼스를 전처리한 결과의 예를 그림 2 에 보인다.



hi : 히라가나, ka : 가타가나, kj : 한자, sy : 기호, c : 복합어, u : 단일어

그림 2. 일본처 코퍼스 전처리

4.1.2. 한국어 코퍼스 전처리

한국어에서는 한글 전용의 관례가 굳어져 일반적으로 한자가 사용되지 않는다. 따라서, 본 논문에서는 한국어에서의 한자 병용의 경우는 제외한다. 한국어 코퍼스의 전처리는 일본어 전처리와 달리 정렬 대상에서 제외할 수 있는 기호, 알파벳, 숫자열을 @로 대체함으로써 실제 한글로 표기된 문자열만을 원본과 동일하게 유지한다. 그림 3 은 한국어 코퍼스 전처리의 예이다.

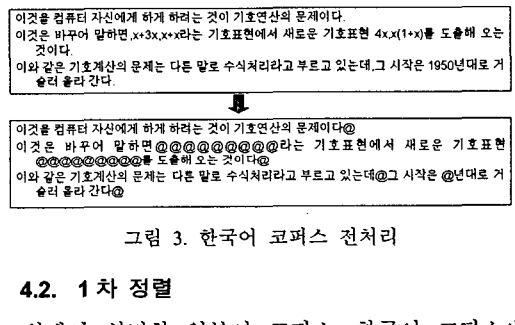


그림 3. 한국어 코퍼스 전처리

4.2. 1 차 정렬

위에서 설명한 일본어 코퍼스, 한국어 코퍼스에 대한 전처리 결과를 대상으로 1 차 정렬을 수행한다. 1 차 정렬은 일본어와 한국어가 한자어를 많이 사용하며 일본어 한자어와 한국어 한자어는 특히 명사의 경우 거의 동일한 한자어를 많이 사용한다는 언어적 유사성을 이용하여 한국어-일본어 사이에서 상호 일치하는 순수 한자어에 대한 정렬을 수행한다. 1 차 정렬 방법을 그림 4 에 예시한다.

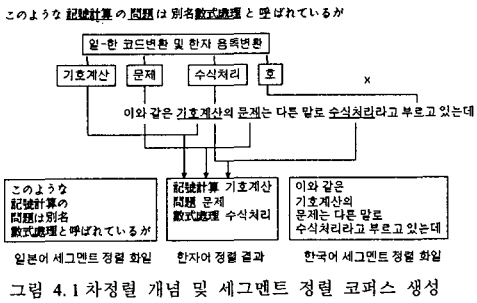


그림 4. 1 차정렬 개념 및 세그멘트 정렬 코퍼스 생성

1 차 정렬 결과 일본어와 한국어에서 사용된 한자어가 정렬된다. 물론 사용된 한자어 전부가 정렬

되지는 않는다. 그러나, 이렇게 정렬된 일본어, 한국어 한자어를 anchor 로 이용하여 기존의 문장 정렬된 일한 코퍼스로부터 그보다 더욱 세밀화된 세그멘트 단위로 정렬된 일한 코퍼스를 얻을 수 있다. 세그멘트 단위로 정렬된 코퍼스는 문장 단위 정렬 코퍼스보다 변별력이 높아지므로 추후 적용될 통계적 기법의 정렬에서 향상된 성능을 보이는 근거가 된다. 1 차 정렬 결과의 예를 그림 5 에 보인다.

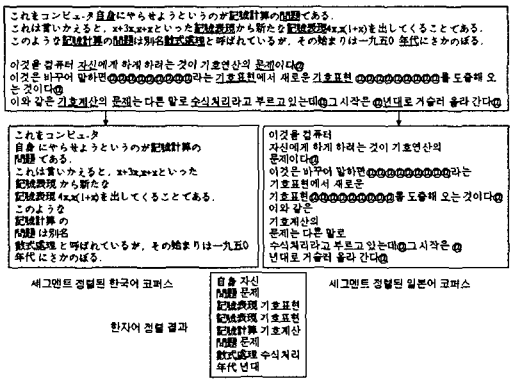


그림 5. 1 차 정렬 결과

4.3. 정렬 대상 후보의 자동 추출

일본어와 한국어 간의 정렬을 위해서는 각 언어마다 정렬할 대상을 추출하여야 한다. 정렬 대상을 구하기 위한 방법으로서 일본어와 한국어 각각에 대한 형태소 분석기 또는 구문 분석기를 이용하는 방법이 있을 수 있다. 그러나 이러한 방법은 다음과 같은 문제를 가지고 있다.

- 형태소 분석기, 구문 분석기 자체가 충분한 품질을 보증하기 어렵다.
- 형태소 분석기, 구문 분석기의 유지, 보수에도 사전의 등록, 갱신 등 적지 않은 노력이 필요하다.
- 형태소 분석기, 구문 분석기의 확보가 어렵다.

따라서, 본 논문에서는 형태소 분석기, 구문 분석기와 같은 자원을 사용하지 않고, 단지, 한일 병렬 코퍼스와 그 수가 제한되어 있어 구축이 용이한 양 언어의 기능어 테이블을 이용한 방법을 제안한다. 이는 또, 정렬에 의한 방법은 정확한 품질을 요구하기 보다는 지식 구축 도구로서의 성격이 강하다는 것에 착안한 것이다.

4.3.1. 한국어 정렬 대상 후보 문자열의 추출

한국어와 일본어의 정렬을 위해서는 각 한일 병렬 코퍼스로부터 올바른 단어를 정렬 대상 후보로 추출해 내어야 한다. 한국어 정렬 대상 후보 단어를 추출하기 위한 기본적인 방안은 한국어가 "개념어 + 기능어"로 구성되어 있다는 문법적 성질을 이용한다. 예를 들면

"본서는 인공지능이란 무엇인가 무엇을 추구하고 있는가 어디에 한계가 있는가를 밝히는 데에 주안점을 두고 있다"

라고 하는 한국어 문장을 분석하여 보면

"본서+는 인공지능+이란 무엇+인가 무엇+을 추구하+고 있+는가 어디+에 한계+가 있+는가를 밝히+는 데+에 주안점+을 두+고 있+다."

와 같이 "개념어부 + 기능어부"로 구성되어 있음을 알 수 있다. 따라서, 기능어 테이블을 이용하여 기능어부를 떼어내면

"본서 인공지능 무엇 무엇 추구하 있 어디 한계 있 밝히 데 주안점 두 있"

과 같이 본 논문에서 정렬 대상으로 후보 문자열이 추출된다. 한국어 기능어 데이터는 기능어(조사, 어미 및 상당구) 및 그 음운변화형을 문법적으로 출현 가능한 순서에 따라 기계적으로 생성해 낸 것으로 7,400 개의 단일/복합 기능어를 포함하고 있다.

4.3.2. 일본어 정렬 대상 후보 문자열의 추출

일본어는 한국어와 달리 띄어쓰기가 없으므로 일본어 표기의 특성을 이용하여 일본어 정렬 후보 문자열을 추출한다. 일본어 정렬 후보 문자열 추출에서는 모든 가능한 후보들을 추출하는 것으로 한다. 그 이유는 부적합한 정렬 후보 문자열들은 전체 텍스트를 대상 범위로 할 때 그 출현 빈도가 낮고 또한 동일한 세그멘트에서 공기할 확률이 낮기 때문에 한국어-일본어 정렬 단계에서 배제되기 때문이다. 이하 일본어 정렬 후보 문자열을 추출하는 방법을 설명한다.

1) 각 세그멘트를 대상으로 "한자로 구성된 문자열 + 히라가나로 구성된 문자열"을 하나의 단위로 하여 문장을 분리한다.

本書は 人工知能とは 何か, 何を 追究しているか

2) 일본어 기능어 테이블을 이용하여 기능어의 앞뒤를 분리한다. 일본어 기능어는 일본 마이니찌 신문 93 년판 품사 부착 코퍼스로부터 기능어에 상당하는 품사가 부착된 연속된 어휘들을 결합하여 얻었으며, 그 수는 8188 개이다.

本書は 人工知能 とは 何か, 何を 追究し ているか

3) 아직 "한자로 구성된 문자열 + 히라가나로 구성된 문자열" 형태로 남아 있는 구성단위는 원래의 형태 외에 자종이 한자에서 히라가나로 바뀌는 위치를 분리한 형태를 부가적으로 생성하여 준다.

本書 は 本書は 人工知能 とは 何 か 何か, 何 を 何を 追究 し 追究し ている か

4.4. 2 차 정렬

한국어 정렬 대상 문자열들과 일본어 정렬 대상 후보 문자열들을 이용 하여 한국어 정렬 대상 단어에 가장 적합한 일본어 정렬 후보 문자열을 찾아내어 대응시킴으로써 한국어-일본어 정렬을 행한다. 각 언어의 정렬 대상 문자열은 해당 문자열이 나타난 세그멘트 번호 정보를 가지고 있다. 정렬 확률의 계산에는 Dice Coefficient 를 사용한다.

$$T(w_k) = \arg\max_{w_j} \frac{2 \times \eta(w_j, w_k)}{n(w_j \in \phi(w_k)) + n(w_k)}$$

$T(w_k)$: 한국어 정렬 대상 w_k 에 대응하는 일본어 정렬 후보 문자열 w_j 를 생성

w_j : 일본어 정렬 대상

w_k : 한국어 정렬 대상

$\phi(w_k)$: 한국어 정렬 대상 w_k 가 나타난 세그멘트에 대응하는 일본어 세그멘트들에 출현하고 있는 일본어 정렬 후보 문자열의 집합

$n(w_k)$: 한국어 정렬 대상 w_k 의 출현 빈도

$n(w_j)$: 일본어 정렬 대상 w_j 의 출현 빈도

$\eta(w_j, w_k)$: w_j, w_k 가 같은 세그멘트에서 나타난 빈도

5. 실험

실험을 위하여 일본에서 간행된 단행본 "人工知能と人間", "モバイル革命"."SGML の理解" 3 권을 번역하여 병렬 코퍼스를 구축하였다. 전체 문장 수는 5,999 문장이며 한글 번역문은 한자를 사용하지 않은 한글 문장이다. 실험은 다음과 같은 세가지 사항을 고려하여 실험을 수행하였다.

첫째, 정렬 후보를 3 개까지 출력한다.
둘째, 정렬의 기준이 되는 언어를 한국어로 하였다.
셋째, 정렬된 대역어의 품질에 따라 2 등급으로 평가한다.

A : 한국어, 일본어 정렬 대상 후보 문자열이 모두 단어로 인정 가능
B : 한국어 정렬 대상 후보 문자열은 완전한 단어이며, 일본어는 완전한 단어가 아닌 대역어가 후보의 부분 문자열에 포함되어 있음

B 등급의 경우까지를 포함한 이유는 정렬 기법에 의한 사전 자동 생성 기술은 100% 완전한 대역 사전의 구축을 보장할 수 없으며, 사전 개발시의 인적, 물적, 시간적 비용을 낮추기 위한 도구로 사용되는 것이므로 사용자가 어느 정도 가공하여 대역어를 구할 수 있는 경우이기 때문이다.

표 1. 한국어-일본어 정렬 실험 결과

구분	Rank / 등급	1 개수	1 비율(%)	2 개수	2 비율(%)	3 개수	3 비율(%)	총 단어수	등급별 비율(%)
비활용어	A	2450	58.9	104	2.5	11	0.2	4162	61.6
	B	186	4.5	5	0.1	4	0.1		4.7
활용어	A	376	41.9	24	2.6	4	0.4	896	44.9
	B	45	5.0	4	0.4	3	0.3		5.7

이와 같은 방법으로 실험한 결과를 표. 1 에 보인다. 한국어 세그멘트 코퍼스로부터 얻어진 고유한 정렬 대상 단어는 총 5,058 개에 달하였다. 이 중에서 비활용어가 4162 개, 활용어가 896 개였다. 정렬 성공된 것들 중 A 등급인 것이 정렬 후보 순위 1 위로 오는 경우가 대부분이었으며, 2 위, 3 위로 나타날 경우는 1 위로 나타날 경우보다 현저히 떨어지고 있음을 알 수 있다. 전체적으로 비활용어의 정렬 성공률은 A, B 등급을 포함하여 66.3%이었으며, 활용어의 경우는 50.6%에 달하였다. 이러한 실험 결과값을 볼 때, 기존의 대역어 사전 개발 방법에 비해 상당히 적은 노력으로 1 차 초벌 대역 사전을 구축할 수 있음을 알 수 있다. 표 5.7 은 A 등급의 정렬 결과를 표 5.8 은 B 등급의 정렬 결과를 표 5.9 는 정렬 실패의 예를 보인 것이다. 추후의 연구에서는 한국어, 일본어 단어 단위 정렬 후보의 올바른 추출에 보다 노력을 기울여야 한다.

참고문헌

[1] Brown, P., Lai, J. and Mercer, R., "Aligning sentences in parallel corpora", *Proc. Of the Annual Meeting of the ACL*, 1991

[2] Church, K., "Char-align : A program for aligning parallel texts at the character level", *Proc. of the Annual Meeting of the ACL*, 1993

[3] Fung, P. and Church, K., "K-vec : A New Approach for Aligning Parallel Texts", *Proc. of 14th International Conference of Computational Linguistics*, 1995

[4] Gale, W. A. and Church, K. W., "Identifying Word Correspondence in Parallel Texts", *Porc. of Speech and Natural Language Workshop*, 1991

[5] Kay, M. and Roscheisen, M., "Text translation alignment", *Computational Linguistics*, Vol 19. No. 1, pp 121-142, 1993

[6] Kupiec, J., "An algorithm for finding noun phrase correspondences in bilingual corpora", *Porc. Of the Annual Meeting of the ACL*, 1993

[7] Matsumoto, Y., Ishimoto, H., Utsuro, T. and Nagao, M., " Structural matching of parallel texts", *Proc. of the Annual Meeting of the ACL*, 1993