韓國 컴퓨터情報學會 2006하계학술발표논문집&학회지
第14卷 第1號, 2006. 6.

번역지원 시스템을 위한 유사 예문 검색

김 동 주*, 김 한 우**

Searching Similar Example Sentences for the Computer-Aided Translation System

JDong-Joo Kim*, Han-Woo Kim**

요 약

본 논문에서는 번역 지원 시스템을 위한 유사문장 검색 알고리즘을 제안한다. 이 알고리즘은 Needleman- Wunsch 알고리즘에 기반을 두고 있으며, 단어의 비교를 위해 단어의 표면어 정보, 표제어 정보, 품사 정보 계층으로 된 다층 정보의 융합을 통해 유사도를 계산하고 정렬을 수행하게 된다. 제안하는 알고리즘은 전기통신 분야의 문장 데이터에 대해 매우 우수한 검색 정확률을 보였다.

Abstract

This paper proposes an similar sentence searching algorithm for the computer-aided translation. The proposed algorithm, which is based on the Needleman-Wunsch algorithm, measures the similarity between the input sentence and the example sentences through combining surface, lemma, part-of-speech information of words with the multi-layered information . It also carries out the alignment between them. The accuracy of the proposed algorithm was very high in the experiment for the example sentences of the area of electricity and communication.

▸ Keyword : computer-aided translation, Needleman-Wunsch algorithm, translation memory, alignments

• 제1저자 : 김동주
*,** 한양대학교 컴퓨터공학과

Ⅰ. 서 론

번역 지원(Computer-aided translation) 시스템이란 번역 과정을 지원하여 인간 번역가가 번역을 용이하게 수행할 수 있도록 설계된 소프트웨어 시스템을 말하며, 번역 보조(Computer-assisted translation) 시스템이라고 부르기도 한다. 이러한 시스템은 완전한 자동번역을 수행하지 않는다는 점에서 기계번역과 근본적인 차이점을 갖는다. 일반적으로 번역 지원 시스템은 맞춤법 검사기를 내장하고 있으며, 전문용어 관리기, 용례 검색기, 대역어 사전, 등의 인간이 직접 번역을 수행하는데 도움이 되는 많은 도구들을 내장하고 있을 뿐만 아니라, 비록 제한적일지라도 심지어는 자동번역 기를 내장하기도 한다. 이 시스템에서 무엇보다 핵심은 번역 메모리(Translation memory)[1]라는 기술에 기반을 둔 유사예문 검색(혹은 매칭)이다.

유사예문 검색은 기번역된 번역 예문 쌍의 집합으로부터 입력문장에 대하여 구조적, 의미적으로 가장 유사한 해당 언어의 예제 문장들을 검색하고, 검색된 예문에 대한 대역 문장들을 사용자에게 제시해 줌으로써 번역을 지원하는 시스템이다. 이는 예제기반 기계번역(Example-based machine translation)[2]에서도 필수적인 요소이다. 번역 지원 시스템을 위한 유사 예문 검색에서 중요한 요소는 동일 언어로 작성된 문장 간의 유사성 정도를 계산하는 것이다. 이와 더불어 유사도 점수에 기여하는 부분이 두 문장에서 어느 부분인지도 제시할 수 있어야 한다. 이를 정렬(Alignment)이라고 부른다. 또한 원어 문장(Source language sentence)과 이에 상응하는 대역 문장(Target language sentence)의 예문 쌍에서 대응(Correspondence)관계, 혹은 대역관계에 놓인 부분 문장 조각들의 파악 또한 사전에 이루어져야 한다.

(1) a. He eats vegetables
그는 채소를 먹는다

b. She bring me potatoes
그녀는 내게 감자를 가져왔다

본 논문에서는 (1-a), (1-b)와 같이 대응관계가 파악된 영한 병렬 말뭉치[5]가 준비되어 있다고 가정한다. 영한 번역을 위한 유사예문 검색시스템을 구축하기 위해 영어 입력 문장과 영어 예문의 유사성 정도를 계산하면서 유사성 정도에 기여하는 부분들을 찾아내 정렬을 수행하는 알고리즘을 제시한다. 제시하는 알고리즘은 Needleman-Wunsch 알고리즘[3]을 문장 비교에 적합하도록 개량한 알고리즘이다.

5) 대역관계에 있는 예문 쌍들의 집합

II. 본론

입력 문장과 가장 유사한 예문들을 병렬 말뭉치 예문들로부터 검색하기 위해서는 동일 언어 문장들 간의 유사도(Similarity), 또는 거리(Distance)에 대한 척도를 정의해야 한다. 또한 일치하는 부분에 대한 대역 예문을 선택적으로 가려내어 유사성 정도에 기여하는 부분들을 파악하기 위해 번역하고자 하는 문장 언어 예문으로부터 일치하는 위치 정보도 알아내야 한다. 이를 위해 본 논문에서는 Needleman-Wunsch(NW) 알고리즘에 기반을 둔 전역적 문자열 비교 알고리즘을 사용한다.

2.1 Needleman-Wunsch 알고리즘

NW 알고리즘은 그림 1과 같이 생물정보학 분야에서 두 단백질에서의 아미노산열의 유사성을 판별하기 위한 알고리즘으로 편집거리(Edit distance)[4] 척도의 가중치 집합을 일반화한 것이다. 즉, 이 알고리즘은 아미노산열 $x_1 \cdots x_m$을 $y_1 \cdots y_m$으로 변환하기 위해 필요한 연산(일치, 대치, 삽입/삭제) 각각에 가중치 (w_m, w_s, w_d)를 주어 matrix A를 계산하면서 동시에 매치되는 문자를 알아내기 위한 경로 정보, 즉 $A(i,j)$의 값이 $A(i-1,j-1)$, $A(i-1,j)$, $A(i,j-1)$ 셋 중 어느 값으로부터 계산된 것인지에 대한 위치정보 $Ptr(i,j)$을 유지하게 된다. 정렬은 이 위치정보 $Ptr(m,n)$로부터 $Ptr(0,0)$까지 역추적에 의해 이루어진다.

$$A(0,0) = 0$$

$$A(i,0) = i \times w_d,\ A(0,j) = j \times w_d$$

$$A(i,j) = \max \begin{cases} A(i-1,j-1) + \alpha & \text{[case1]} \\ A(i-1,j) + w_d & \text{[case2]} \\ A(i,j-1) + w_d & \text{[case3]} \end{cases}$$

$$\text{where} \quad \alpha = \begin{cases} w_m & \text{if } x_{i-1} = y_{j-1} \\ w_s & \text{if } x_{i-1} \neq y_{j-1} \end{cases}$$

$$Ptr(i,j) = \begin{cases} Diag & \text{[case1]} \\ Up & \text{[case2]} \\ Down & \text{[case3]} \end{cases}$$

[그림 1] Needleman-Wunsch 알고리즘

2.2 다층 유사도 측정을 통한 유사 예문 검색

NW 알고리즘을 문장 비교 문제에 적용하기 위해서는 단어 단위의 비교를 필요로 한다. 따라서 (2)의 두 영어 문장에 대한 유사성 점수와 정렬을 구하면 그림 2와 같은 결과를 얻을 수 있다. 가중치 w_m, w_s, w_d는 각각 1, -1, -2로 설정하였다. 그림 2는 두 문장 (2-a)와 (2-b)의 유사성 점수는 0이며 'read'를 'buys'로, 'the'를 'a'로 대치하는 것이 최적의 정렬임을 의미한다.

(2) a. He reads the book
b. He buys a book

		He	buys	a	book
	0	-2	-4	-6	-8
He	-2	1	-1	-3	-5
read	-4	-1	0	-2	-4
the	-6	-3	-2	-1	-3
book	-8	-5	-4	-3	0

He reads the book
| | | |
He buys a book

【그림 2】 문장 (2-1)와 (2-b)에 대한 계산 예

그런데 문장 (3-a)와 (3-b)의 경우 두 문장이 구조적으로 (2)의 경우보다 더 상이함에도 불구하고 유사성 점수는 같다. 즉, (2-a) 문장은 (3-b) 문장보다 구조적으로 (2-b) 문장이 더 유사함에도 불구하고 유사도 점수가 동일하여 번역에 더 유용한 문장을 명확히 판단하지 못하는 경우가 발생한다. 더욱 극단적인 경우에는 구조적으로 덜 유사한 문장을 유사성이 더 높은 문장으로 잘못 판단할 수도 있다. 이 문제를 해결하기 위해 단어의 표면 정보만 반영하는 것이 아니라 단어의 품사 정보를 추가적으로 반영한다. 즉, 단어의 표면형과는 무관하게 비교하는 단어의 품사가 동일하다면 표면정보 가중치 외에 품사정보 가중치를 추가적으로 준다.

(3) a. He reads the book
b. He became the spokesman

추가적으로 고려해야 할 것은 형태론적 변형에 관한 것이다. 파생접사류와는 달리 시제, 수의 일치, 단복수의 구분을 위한 굴절접사류는 어간(Stem)의 의미를 변화시키지 않아 한국어로의 번역시에 번역되지 않거나 번역하지 않아도 의미상 문제가 없는 경우가 많다는 것이다. 따라서 형태론적 변형에 따른 급격한 유사도 변화를 완화하기 위해 표제어(Lemma)를 비교하여 표면 정보가 같지 않더라고 표제어가 동일하면 가중치를 준다. 또한 한정사(Determiner) 중 일반적으로 번역이 되지 않는 관사류는 유사도 계산에서 제외한다.

POS Layer (PL)	NNP	VB	DT	NN
Stem Layer (TL)	she	use	the	pencil
Surface Layer (SL)	She	uses	the	pencil

【그림 3】 다층 정보

【표 1】 세분화된 가중치 집합

	w_s(SL)	w_t(TL)	w_p(PL)
w_m(match)	w_s^m	w_t^m	w_p^m
w_s(mismatch)	w_s^s	w_t^s	w_p^s

이와 같이 본 논문에서는 문장간 유사성 검사와 정렬의 정확성을 높이기 위해 표면정보 외에 단어의 어간 정보, 품사 정보를 함께 반영하는 다층 유사성 척도를 사용한다. 즉, 그림 3과 같은 다층 정보는 표면층(SL), 어간층(TL), 품사층

(PL)으로 이루어져 있다. 표면층에서 질의문과 예문의 한 단어는 각각 s_i^x, s_j^y, 어간층에서 어간은 t_i^x, t_j^y, 품사층에서 품사는 p_i^x, p_j^y이고 $1<i<m$, $1<j<n$이라고 했을 때, NW 알고리즘의 α에서 문자 일치여부에 대한 가중치 w_m, w_s를 표 1과 같이 세분한다. 즉, 단어의 표층형에 관한 가중치 외에 단어의 어간이나 품사의 일치여부에 관한 가중치 w_t와 w_p를 추가하게 된다. 따라서 $A(i,j)$를 계산할 때, 대각 방향의 계산 $A(i-1,j-1)+\alpha$에서 α는 식 1과 같다. 삽입, 혹은 삭제 단어는 질의문이나 예문에서 해당 단어를 경계로 문맥적 단절의 가능성이 높아 두 문장의 유사성 정도에 기여하지 못하고 오로지 비유사성 정도만을 나타낼 뿐이다. 따라서 삽입, 삭제에 대한 가중치 w_d는 일치(w_m: w_s^m, w_t^m, w_p^m)와 대치(w_s: w_s^s, w_t^s, w_p^s)에 비해 커서는 않된다.

$$\alpha = w_s + w_t + w_p$$

$$w_s = \begin{cases} w_s^m & \text{if } s_{i-1}^x = s_{j-1}^y \\ w_s^s & \text{if } s_{i-1}^x \neq s_{j-1}^y \end{cases}$$

$$w_t = \begin{cases} w_t^m & \text{if } t_{i-1}^x = t_{j-1}^y \\ w_t^s & \text{if } t_{i-1}^x \neq t_{j-1}^y \end{cases}$$

$$w_p = \begin{cases} w_p^m & \text{if } p_{i-1}^x = p_{j-1}^y \\ w_p^s & \text{if } p_{i-1}^x \neq p_{j-1}^y \end{cases}$$

【식 1】 대각 방향(일치, 대치) 가중치

정렬은 유사도 계산과 동시에 저장되는 역추적 정보($Ptr(i,j)$)를 통하여 이루어지는데, 이는 번역 단계에서 대역어 선택에 있어 중요한 역할을 수행한다. 그런데 일치와 대치에 대해 역추적 정보는 모두 대각 방향으로 둘을 구분하지 못한다. NW 알고리즘에서 정렬 관계는 표면어의 일치여부에 따라 단 두 가지 α만을 갖기 때문에 일치와 대치 관계의 결정은 두 가지 값에 따르면 된다. 반면에 제안하는 척도에서 α의 값은 세 가지 정보에 대한 가중치의 합이므로 2^3가지 수가 존재한다. 그런데 표면어, 어간, 품사 정보의 계층 관계로 인하여 4가지 경우는 발생하지 않는다. 따라서 계층관계에 부합하는 4가지 경우만 발생하는데, 정렬 관계를 결정하기 위해 이 4가지 경우를 일치와 대치로 구분하여야 한다. 본 논문에서는 4가지를 다음과 같이 일치와 대치로 구분한다.

일치(M): 정렬 관계에 있는 단어의 어간과 품사가 동일한 경우

$$A(i,j) = A(i-1,j-1) + w_s^m + w_t^m + w_p^m$$
$$A(i,j) = A(i-1,j-1) + w_s^s + w_t^m + w_p^m$$

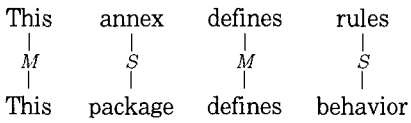
대치(S): 표층어와 어간이 일치하지 않는 경우

$$A(i,j) = A(i-1,j-1) + w_s^s + w_t^s + w_p^m$$
$$A(i,j) = A(i-1,j-1) + w_s^s + w_t^s + w_p^s$$

또한 길이 m인 질의문과 길이 n인 예문에서 하나의 정렬 관계 $r_l(1 \le l \le \max(m,n))$은 일치(M), 대치(S), 삽입(I), 삭제(D)로 정의한다. 이렇게 정의된 유사성 척도를 이용하여 질의 문장 (4-a)에 대해 예제 문장 (4-b)와의 유사성 점수 계산하고 정렬 관계를 역추적하면 그림 4와 같다. 이 예에서 가중치 집합은 $w_s^m = w_t^m = w_p^m = 1$, $w_s^s = w_t^s = w_p^s = -1$, $w_d = -3$으로 설정하였다.

(4) a. This annex defines rules
b. This package defines the behavior

		DT this This	NN package package	VB define defines	NN behavi behavi
	0	-3	-6	-9	-12
DT this	This	-3	3	0	-3	-6
NN annex	annex	-6	0	2	-1	-4
VB define	defines	-9	-3	-1	5	2
NN rule	rules	-12	-6	-4	2	4



【그림 4】 유사 점수와 정렬 관계 계산의 예

III. 실험 및 평가

실험에서 사용된 병렬 말뭉치는 'ITU-T 권고' 710 문장쌍[6]이다. Brill의 태거[5]로 태깅을 수행하였고, 사용된 품사 태그 집합은 Penn Treebank 태그 집합[6]을 기반으로 한국어의 번역에 영향을 비교적 주지 않는 NNS는 NN으로, VBP와 VBZ은 VB로 통합하여 총 33개 태그를 사용하였다. 표제어 추출(Lemmatization)을 위한 도구는 WordSmith를 사용하였다. 가중치는 $w_s^m = w_t^m = w_p^m = 1$, $w_s^s = w_t^s = w_p^s = -1$, $w_d = -3$으로 설정하였다. 알고리즘 평가를 위해 ITU-T 권고 710 문장쌍에서 모든 영어 문장을 한 번씩 질의문으로 사용하였으며 질의문으로 선택된 문장은 예문에서 제외하여 반복적으로 실험하였다. 각 질의문에 대하여 유사성 점수 S가 0이상인 것에 대한 검색 성공률과 유사성 순으로 상위 3개(T3)에 대해 검색 정확률을 평가하였다. 표 2와 같이 약 73.2%의 정확도를 보였다. 15개 이하의 단어수를 갖는 문장에 대한 검색 정확률은 매우 정확하였으나 21개 이상에 대해서는 매우 부정확했다. 통계량으로서 데이터의 양이 충분하지 않아 단정적으로 이야기하기 어렵지만, 긴 문장에 대한 부정확성은 NW 알고리즘의 특성인 전역적 유사성 비교에 따른 자료 부족 문제를 주요 원인으로 꼽을 수 있을 것이다.

【표 2】 유사 문장 검색 정확률

단어수	5이하	6-10	11-15	16-20	21이상	전체
문장수	100	107	298	197	8	710
$S>0$ 성공률(%)	100 (100)	92.5 (99)	83.2 (248)	68.0 (134)	37.5 (3)	82.3 (584)
T3 정확률(%)	98.0 (98)	89.8 (95)	74.8 (223)	55.3 (103)	12.5 (1)	73.2 (520)

6) 국제전기통신연합(ITU)에서 전기 통신 업무의 기술, 운용 및 요금 문제를 연구하여 그 결론을 권고로 공표하는 문서로 한국정보통신기술협회에서 번역하여 표준 번호 문서 단위로 대응되어있는 한영 병렬코퍼스를 수작업을 통하여 문장단위 대응 관계를 파악하였다.

IV. 결론 및 향후 연구 방향

본 논문에서는 전기통신 분야의 제한된 영역의 문장을 사용하여 번역지원 시스템을 위한 유사 예문 검색 알고리즘을 제안하였다. 매우 작은 량의 문장을 사용하였음에도 불구하고 비교적 긍정적인 결과를 얻을 수 있었다. 향후 과제는 제한된 영역의 문장이 아닌 일반 영역에서의 문장에 대해서도 평가가 이루어져야 할 것이다. 또한 해결되어야 할 문제점으로는 긴 문장에 대한 고품질의 번역을 위해서는 NW 알고리즘에서의 전역적 유사성 비교 외에 국부적(Local) 유사성 비교가 동반되어야 할 것으로 생각된다. 이와 더불어 최적의 가중치 집합을 결정하는 문제 또한 많은 연구가 필요할 것이다. 또한 실용화에 있어 무엇보다 우선적으로 해결되어야 검색 속도 문제는 [7]의 예에서와 같이 검색 대상 예문을 제한하는 방법론이나 적합한 새로운 파일시스템이 고안되어야 할 것이다.

참고문헌

[1] M. Kay, "The Proper Place of Men and Machines in Language Translation," Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, Calif., Reprinted in Machine Translation vol. 12, pp. 3-33 (1997), 1980.

[2] M. Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," In Artificial and human intelligence, A. Elithorn and R. Banerji (Eds.), Amsterdam: North-Holland, pp. 173-180, 1994.

[3] S. Needleman and D. Wunsch, "A General Method Applicable to the search for similarities in the amino acid sequence of two proteins," Journal of Molecular Biology Vol. 48, pp 443-453, 1970.

[4] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," Soviet Physics - Doklady, Vol. 10 No. 8, pp. 707-710, February 1996, Translated from Doklady Akademii Nauk SSSR, Vol. 163 No. 4 pp.845-848, August 1965.

[5] E. Brill, "Some Advances in Transformation- Based Part of Speech Tagging," Proceedings of the Conference of the American Association for Artificial Interlligence, 1994.

[6] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics, Vol. 19, No. 2, pp. 313-330, 1993.

[7] L. Cranias et al., "Clustering: A Technique for Search Space Reduction in Example-Based Machine Translation," Proceedings of International Conference on System, Man, and Cybernetics, Oct. 2-5, pp. 1-6. 1994.

저 자 소 개

김동주

1996년 한양대학교 전자계산학 공학사.
1998년 한양대학교 전자계산학 공학석사.
2001년 한양대학교 전자계산학 박사학위 과정 수료.

관심분야는 한국어 형태소 및 구문 분석, 정보검색, 맞춤법 검사, 기계번역

김한우

1975년 한양대학교 전자공학 공학사.
1978년 한양대학교 전자공학 공학석사.
1980년 한양대학교 전자공학 공학박사.
1981년~현재 한양대학교 전자컴퓨터공학부 교수.

관심분야는 정보처리, 자연언어처리, 기계번역.