김일성종합대학학보
정보과학
주체107(2018)년
제64권 제4호

TM을 리용한 기계번역의 한가지 방법

김동수, 리명일

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《프로그람을 개발하는데서 기본은 우리 식의 프로그람을 개발하는것입니다. 우리는 우리 식의 프로그람을 개발하는 방향으로 나가야 합니다.》(《**김정일**선집》 증보판 제21권 42페지)

우리는 TM(Translation Memory)을 리용하여 기계번역을 진행하는 한가지 방법을 고찰하였다.

EBMT[1－4]에서는 *EB*를 구축하고 입력문과 *EB*의 원천문과의 류사도를 계산하여 그 값에 따라 번역을 진행한다. 부족점은 *EB*의 구축에 많은 노력이 들고 리용률이 높지 못한것이다.

론문에서는 자료기지작성이 쉽고 리용률이 높은 TM을 작성한 다음 입력언어와 TM의 원천언어사이의 모호정합에 의한 류사도를 계산하고 류사도가 큰 TM의 목적어를 출력하는 방법으로 기계번역을 진행하였다.

1. CAT에 의한 기계번역

CAT(Computer Assisted Translation)[4]는 번역원의 번역과정을 지원하는 프로그람을 리용하여 언어번역을 진행하는 기계번역이다. CAT에는 여러가지 기능이 있는데 기본은 입력어에 대한 목적어의 출력기능, 원천어본문과 목적어본문을 각각 토막화하고 토막들에서 단어대응을 찾아 번역기억에 추가하는 기능이다.

EBMT에서는 *EB*를 구축한 다음 입력문과 *EB*의 원천문과의 류사도를 계산하고 류사도가 큰 원천문에 대응한 목적문을 출력하는 방식으로 기계번역을 진행한다. 여기서 *EB*는 원천문의 매 단어에 정확한 문장론적정보와 의미론적정보로, 목적문의 매 단어에 대역문생성에 필요한 여러가지 정보로 구성된다.

$E = KEY \ \& \ ES : KS$

KEY: 원천문장의 어휘개수

ES: 원천문(영어)

KS: 목적문(번역문)(조선어)

여기서 *ES*와 *KS*는 다음과 같이 표시된다.

ES: $EW_1[(Fu_1)]\ EW_2[(Fu_2)]\ EW_3[(Fu_3)]\ \cdots\ EW_n[(Fu_n)]$

EW_i: 원천문에서 i번째 단어(i=1, 2, …, n)

Fu_i: 원천문에서 i번째 단어의 품사정보

KS: $KW_1(n_1)[\backslash f_1]KW_2(n_2)[\backslash f_2]\ \cdots\ KW_m(n_m)[\backslash f_m]$

KW_j: 번역문에서 j번째 단어(j = 1, 2, …, m)

n_j: 대응되는 원천단어의 위치번호

f_j: 단어의 토정보

이것은 기계번역을 위한 실례기지를 작성하는데 많은 노력이 들며 자료기지의 정확성도 높지 못하다.

그러므로 기계번역을 위한 능률적인 실례기지를 만드는것이 중요한 문제로 나선다. 바로 이 문제를 해결하는것이 CAT의 번역기억 TM이다. TM은 병렬코퍼스를 그대로 리용하는데 다만 매 단어에 련결하는 원천단어의 위치번호를 붙여 목적문을 작성한다.

2. TM의 구축과 리용

TM을 다음과 같이 정의한다.

먼저 번역된 병렬코퍼스에서 원천어문장과 목적어문장을 토막으로 분리한다. 여기서 토막은 구, 문장, 단락으로 설정한다.

분리된 토막에 대하여 TM의 기록으로 정보를 추가하여 작성한다.

$TM = \{Re_i | i=1, 2, \cdots, N: N$은 기록의 총개수$\}$

$Key_i =$ (단어수 류형정보)

$SS = (sw_1, sw_2, \cdots, sw_n)$

$TS = (tw_1(t_1), tw_2(t_2), \cdots, tw_n(t_n))$

sw_i: 원천문의 단어, tw_i: 목적문의 단어

t_i : 목적문의 단어에 대응한 원천문의 단어순서번호

원천문의 대응한 순서번호는 영조기계번역기 《룡남산》의 영어－조선어단어대응설정도구에 의하여 자동적으로 진행된다.

입력문을 S라고 할 때 TM_i와의 류사성탐색의 고속성을 보장하기 위하여 구축된 TM의 구조에 따라 1차탐색(Key에 의한 탐색)과 2차탐색(문자렬편집거리를 리용한 모호정합도탐색)을 진행한다.

1차탐색은 TM에 있는 Key(원천어의 단어수와 문장류형)에 따라 진행한다. 2차탐색은 단어수가 1개 차이나면서 문장류형이 같은 원천어들에 대하여 진행한다.

2차탐색은 다음의 문자렬편집거리를 리용한 모호정합도계산에 의한 류사성탐색이다.

$$FMS(S_1, S_2) = 1 - \frac{dis(S_1, S_2)}{\max(|S_1|, |S_2|)}$$

여기서 $FMS(S_1, S_2)$는 문장 S_1과 S_2의 모호정합도, $|S_i|$는 문장 S_i에 있는 단어들의 총수, $dis(S_1, S_2)$는 문자렬편집거리이다.

문자렬편집거리는 문장 S_1과 S_2 사이의 단어대응관계를 고려하여 편집할 때 삭제, 삽입, 취환되는 단어의 개수이다.

모호정합도가 제일 큰 TM의 원천문장에 대응하는 기록을 선택하고 모호정합도계산에 기여한 단어들 즉 삽입, 제거, 치환하는 단어들을 선택하여 대응하는 목적어의 단어들에 대하여 삽입, 제거, 취환을 진행하고 번역문을 얻는다.

문자렬편집거리에 의한 모호정합도가 100%인 문장은 대단히 짧은 문장이며 그렇다고 하여 문장의 길이와 모호정합도사이에 호상관련이 있는것은 아니다.

제안방법과 선행방법들[1－4]을 비교하기 위하여 3 000개의 실례문을 가진 TM을 구축하고 어휘수가 10～12인 문장 300개, 13～15인 문장 300개, 16～20인 문장 300개 즉 900개의 문장을 준비하였다.

번역의 정확성은 다음과 같이 계산된다.

$$정확도(\%) = \frac{정확한\ \ 번역문장수}{전체\ \ 문장수} \times 100$$

실험결과는 표와 같다.

표. 실 험 결 과

시험본문	체 계		
	류사도에 의한 EBMT	실마리어에 의한 EBMT	제안방법
본문 1(10～12)	92	90	96
본문 2(13～15)	90	82	95
본문 3(16～20)	86	61	90

맺 는 말

제안방법은 선행방법들에 비하여 체계의 번역정확성이 높으며 번역을 위한 TM은 자동적으로 구축되므로 기계번역의 신속정확성과 유연성을 보장한다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 53, 1, 52, 주체96(2007).

[2] 김동수; 정보과학과 기술, 6, 25, 주체92(2003).

[3] E. Aramaki, S. Kurohashi; International Workshop on Spoken Language Translation (IWSLT), 91, 2004.

[4] C. Alabau et al.; Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 25, 2014.

주체107(2018)년 8월 5일 원고접수

A Method of Machine Translation Using TM

Kim Tong Su, Ri Myong Il

This paper proposed a method of machine translation using TM.

We created TM form parallel corpus and similarity search between input sentence and source sentence of TM carried out by the first search according to the number of words and second search based on Fuzzy matching score using string edit distance. Then the system produced the target sentence of TM with the highest similarity as translation sentence.

Key words: EBMT, TM