김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제6호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 6 JUCHE106(2017).

도조기계번역에서 조선어대역선택을 위한 SNoW토대의 학습체계구축에 대한 연구

신혁철, 한승주, 량철호

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《당의 과학기술중시로선을 철저히 관철하여 첨단과학기술분야를 개척하며 나라의 과학기술을 높은 수준에 올려세워야 합니다.》(《**김정일**선집》 증보판 제23권 502페지)

대역선택문제는 단어가 가지는 의미적애매성을 해소하기 위한 문제이다.

현재 기계번역체계에서 의미적애매성해소(WSD: Word Sense Disambiguation)에는 최대엔트로피모형, *n*그람법을 비롯하여 여러가지 기계학습방법들이 리용되고있다.[2－4]

코퍼스의 량이 점차적으로 늘어나는데 맞게 대역선택의 정확도를 점차적으로 개선하는 체계를 구축하기 위하여 론문에서는 직결식학습방법인 SNoW[1]를 리용하는 조선어대역선택학습체계를 구축하였다.

1. SNoW를 리용한 조선어대역선택학습체계구축방법

1) SNoW토대의 조선어대역학습체계의 목표

① 학습자료에 대하여 100% 성능에 도달하도록 하는것이다.

② 학습자료에 포함되지 않은 미지단어들에 대해서도 일정한 성능을 보장하도록 하는것이다.

③ 학습자료량의 증가에 따라 일반화정확도가 계속 높아지며 갱신된 대역선택체계가 번역체계와 주기적으로 반결합될수 있도록 하는것이다.

이러한 목적은 직결식학습방법인 SNoW의 특성으로부터 비교적 쉽게 달성될수 있으므로 론문에서는 SNoW를 리용하여 조선어대역학습체계를 구축하였다.

2) 특징설계

SNoW는 잡음견딤성이 비교적 강하고 학습과정에 특징의 무게를 자동적으로 결정하기때문에 특징모임이 충분히 설계되여야 한다.

론문에서는 다음과 같은 특징들을 리용하였다.

① 어휘특징

문장안에 있는 단어들의 어휘문자렬들을 특징으로 리용한다. 도이췰란드어에서는 단어의 형태변화가 매우 복잡하기때문에 출현단어가 아니라 원형단어를 특징으로 리용함으로써 학습체계의 부하를 줄이고 일반화성능을 높일수 있게 하였다. 한편 어휘특징에 주목하는 단어를 포함시켜 그 단어의 대역별출현빈도가 학습과정에 간접적으로 반영되도록 하였다. 학습체계의 과학습을 방지하고 일반화정확성을 높이기 위해 주목하는 단어에 대해

서는 어휘만을 특징으로 리용하였다.

② 구문관계특징

이 특징은 주목하는 단어가 문장속에서 가지게 되는 구문적관계정보를 반영한다. 대역코퍼스구축이 기계번역체계를 리용하여 반자동적으로 진행되기때문에 문장의 구문구조에 대한 추가적인 정보가 존재하게 된다. 현재 구문해석과정에 얻어지는 구문관계들은 다음과 같다.

명－동, 동－명, 명－명(생격), 형－명, 대－명, 명－병접－명, 명－전－명, 동－전－명

우의 구문관계들은 대역선택에서 선택제한정보들을 음적으로 포함하고있기때문에 주목하는 단어와 구문관계를 이루는 단어들에 대해서는 어휘, 의미, 대역정보들을 특징으로 반영한다. 구문관계를 이루는 단어들의 대역들을 특징으로 리용하였기때문에 조선어에서의 련어적관계가 간접적으로 반영되게 된다.

③ 의미특징

문장안에 있는 명사단어들에 대해서 의미를 특징으로 리용한다. 이때 일반화성능을 높이기 위하여 명사의 의미 그자체만이 아니라 그것의 상위개념까지도 함께 특징으로 반영한다.

3) SNoW토대의 조선어대역선택학습체계의 구성

기계학습문제로서의 대역선택문제는 일반적으로 다중선택문제이므로 SNoW의 특성에 잘 맞는다. 다시말하여 대역코퍼스에서 매 단어에 할당되여있는 대역은 SNoW에서 그 단어의 정의 실례로, 단어가 가지게 되는 다른 대역들은 부의 실례로 학습하게 된다.

SNoW토대의 조선어대역선택학습체계에서 SNoW의 클라스모임은 단어가 가지는 대역들로 구성할수 있다. 그러나 조선어대역으로만 클라스들을 설정하는 경우 동음이의어들의 선택에서 애매한 문제가 발생될수 있다. 론문에서는 이와 같은 현상을 막기 위해 대역과 의미정보를 결합하여 SNoW의 클라스모임을 구성한다. 이렇게 하면 대역코퍼스에 들어있지 않은 미지어들에 대한 대역선택에서도 일정한 효과를 볼수 있다.

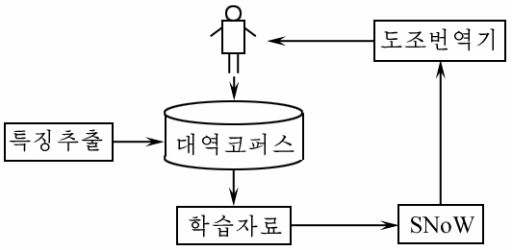

그림. SNoW토대의 조선어대역 학습체계의 구성도

도조기계번역에서 SNoW토대의 조선어대역학습체계구성도는 그림과 같다.

2. 실험 및 평가

실험에서 SNoW의 증가파라메터는 1.2, 감소파라메터는 0.5로 설정하였다. 학습자료에 대한 체계의 목표를 달성하기 위해 일정한 주기마다 SNoW를 반복학습시켰다. 학습회수 60일 때 회복도는 99.5%에 달하였다. 일반화성능을 평가하기 위하여 학습되지 않은 의학분야의 100개 문장에 대한 번역실험을 하였다. 번역실험결과 시험본문속에 들어있는 총 622개의 다의어들에 대한 선택정확도는 70.1%로서 베이스법을 리용했을 때의 47.1%에 비하여 약 23% 개선되였다.

맺 는 말

반자동적으로 구축되는 대역코퍼스를 리용하는 대역선택체계에서 SNoW를 리용하면 도조기계번역에서 제기되는 대역선택의 정확성을 보다 높일수 있다.

참 고 문 헌

[1] A. R. Golding et al.; Machine Learning, **34**, 107, 1999.

[2] C. H. Chueh et al.; Computational Linguistics and Chinese Language Processing, **11**, 1, 37, 2006.

[3] P. F. Brown et al.; Computational Linguistics, **18**, 4, 467, 1992.

[4] F. M. Tyers et al.; In Proceedings of the 16th Conference of the European Association for Machine Translation(EAMT), 213, 2012.

주체106(2017)년 2월 5일 원고접수

Building the SNoW-based Learning System for Korean Sense Selection in German-Korean Machine Translation

Sin Hyok Chol, Han Sung Ju and *Ryang Chol Ho*

We presented the WSD (Word Sense Disambiguation) learning system using SNoW, online learning algorithm. We set up the object of the WSD learning system according to the characteristics of SNoW and G-K MT, and designed the feature set including lexical, semantic, and syntactical features. We showed that our object was able to achieve in the online learning system using SNoW, based on the sense-corpus that is semi-automatically made.

Key words: machine translation, SNoW, WSD