대상 영역 코퍼스를 이용한 번역사전의 특정 영역화를 위한 워크벤치

노윤형** 이현아*○ 김길창*
한국과학기술원 전자전산학과*, 한국전자통신연구원**
yhnoh@etri.re.kr, halee@csone.kaist.ac.kr, gckim@cs.kaist.ac.kr

A Workbench for Domain Adaptation of an MT Lexicon with a Target Domain Corpus

Yoon-Hyung Noh** Hyun Ah Lee*○ Gil Chang Kim*
Dept. of Electrical Engineering & Computer Science, KAIST*, ETRI**

요 약

기계번역에서 좋은 품질의 번역 결과를 얻기 위해서는 대상으로 하고 있는 전문 영역에 맞게 시스템의 번역 지식을 조정해야 한다. 본 연구에서는 대상 영역 코퍼스를 이용하여 기계번역 시스템의 특정 영역화를 지원하는 워크벤치를 설계하고 구현한다. 워크벤치는 대상 영역의 코퍼스에서 대상 영역의 지식을 추출하는 영역 지식 추출기와, 추출된 지식을 사용자에게 제시하여 사용자가 사전을 편집할 수 있는 환경을 제공하는 영역 지식 검색기와 사전 편집기로 구성된다. 구현된 워크벤치를 이용하여 일반 영역 사전을 군사 정보 영역으로 특정 영역화를 해 본 결과, 효율성과 정확성에서의 향상이 있었다.

1 서론

일반적으로 자연 언어는 각 대상 영역에 따라 현저히 다른 어휘적, 구문적, 의미적 성향을 나타낸다. 따라서, 기계번역에서 좋은 품질의 번역 결과를 얻기 위해서는 일반 영역을 대상으로 하는 것보다, 대상 영역을 한정짓고 그에 맞게 번역시스템을 조정해 주는 것이 좋다. 번역 시스템을 대상 영역에 맞게 조정하는 작업은 번역의 품질과 범위를 결정하는 핵심 요소인 사전에 대상 영역의 특성을 반영하는 것으로 가능해진다. 예를 들어 그림 1의 왼쪽과 같은 일반 영역 사전으로 군사 정보 영역의 문장 "Additional armour is deployed by one tank company."를 번역하면 "추가 갑옷이 한 탱크 회사에 의해 배치되었다"와 같이 틀린 번역을 얻을 수 있다. 하지만, 그림 1의 오른쪽과 같이 대상 영역에 맞도록 번역어를 조정하고 복합어를 추가한 사전을 이용하면 "추가 전차가 한 탱크 부대에 의해 배치되었다"와 같이 올바른 번역을 얻을 수 있다.

이와 같이 특정 영역화된 사전을 얻기 위해서는 대상 영역의 고유한 특성을 파악하여 사전에 반영해야 한다. 대상 영역의 특성은 대상 영역 코퍼스를 분석하여 얻을

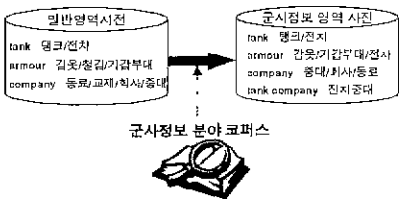

그림 1. 번역사전의 특정 영역화

수 있다. 이 때, 수작업을 통하여 대상 영역의 코퍼스를 분석하고 일관성 있게 사전에 반영하는 것은 많은 노력과 시간이 필요하고, 이를 완전 자동으로 수행하여 사전을 변경한다면 번역 사전의 신뢰도가 급격히 저하될 가능성이 높다. 만일 대상 영역 코퍼스를 자동으로 분석하고, 이를 참조하여 사전 개발자가 사전을 편집할 수 있는 환경이 제공된다면 효율성과 정확성을 모두 만족시킬 수 있다.

본 논문에서는 번역시스템, 특히 번역 사전을 대상 영역에 맞도록 조정하는 작업을 특정 영역화(domain adaptation)라고 정의하고, 이를 지원하는 워크벤치를 설계하고 구현한다. 번역 사전의 특정 영역화 워크벤치

는 영역 지식 추출기와 영역 지식 검색기, 사전 편집기로 구성된다. 영역 지식 추출기에서는 대상 영역의 코퍼스를 분석하여 대상 영역에서 나타나는 복합어나 숙어, 번역어에 대한 통계 정보 등을 구한다. 분석된 영역 지식은 영역 지식 검색기를 통해 사용자가 보고 이해하기 쉬운 형태로 제시된다. 제시된 내용을 참조하여 사전 작성자는 사전 편집기를 이용하여 번역 사전을 특정 영역화한다.

본 논문은 다음과 같이 구성된다. 2절에서는 특정 영역화를 위해 필요한 작업과 정보를 분석하고, 3절에서 특정 영역화 워크벤치의 영역 지식 추출기, 검색기, 사전 편집기에 대해서 설명한다. 4절에서 실험 및 평가를 하고 5절에서 결론을 맺는다.

2 번역 사전의 특정 영역화

그림 1에서 볼 수 있듯이, 번역 사전의 특정 영역화를 위해서는 대상 영역에서 자주 쓰이는 번역어나 복합어 등의 영역 지식을 추출하고 이를 사전에 반영해야 한다. 본 절에서는 특정 영역화가 되어야 할 사전 지식과 특정 영역화 워크벤치가 제공해야 할 정보를 일반적인 번역 과정인 어휘, 구문, 변환, 생성의 단계를 기준으로 살펴본다.

- 어휘 분석을 위한 정보
  번역 과정 중 어휘 분석에서는 어휘의 품사와 활용형을 분석한다. 특정 영역을 대상으로 하여 어휘를 분석하는 경우에는 일반 영역에서는 거의 나타나지 않지만 대상 영역에서는 빈번히 사용되는 단어나 복합어, 활용형들이 발생할 수 있다. 또한, 여러 품사나 여러 의미로 쓰이는, 즉 어휘적 애매성이 있는 단어가 대상 영역에서는 특정 품사나 의미로만 이용될 수도 있다. 이러한 정보를 사전에 반영하면 보다 정확한 어휘 분석 결과를 얻을 수 있다. 특정 영역화 워크벤치에서는 대상 영역 코퍼스에서 대상 영역 특유의 단어와 복합어를 추출하여 사전에 등록할 수 있게 지원하고, 각 단어의 활용별/품사별 통계 정보, 어휘 분석을 위한 문맥 정보 등을 제공하여 사전에 어휘적 애매성을 해소하기 위한 정보를 추가할 수 있게 지원한다.

- 구문 분석을 위한 정보
  구문 분석에서는 분석 규칙을 이용하여 입력 문장의 구조를 분석한다. 구문 분석 정보의 특정 영역화를 위해서는 대상 영역에 자주 나타나는 구문 패턴이나 숙어, 관용어를 분석하기 위한 규칙을 사전에 추가하고, 사전에 기술되어 있지만 대상 영역에는 적절하지 않은 숙어를 삭제하는 작업들이 필요하다. 이를 위해 특정 영역화 워크벤치에서는 코퍼스에서 자주 나타나는 구문 패턴이나 숙어의 빈도와 문맥 정보를 제공해야 한다.

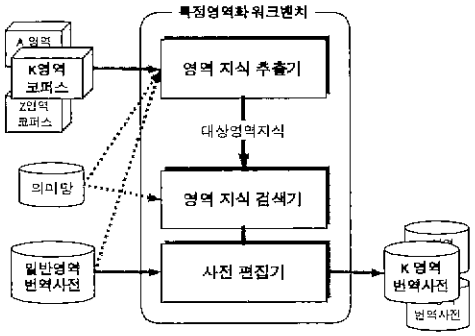

그림 2: 워크벤치 전체 구성

- 변환 및 생성을 위한 정보
  기계번역의 변환 단계에서는 구조 변환과 함께 번역어 선택이 이루어진다. 생성 단계에서는 선택된 번역어들을 이용하여 목적 언어 문장을 구성한다. 일반적으로 번역어 선택은 문맥에 나타나는 단어, 즉 공기 정보를 이용하는데, 만일 일반 영역을 기준으로 기술된 공기 정보를 사용하면 올바른 번역어 선택 결과를 얻기 힘들다. 특정 영역화를 위해서는 대상 영역에 맞는 공기 정보를 사전에 기술하여 정확한 번역어를 선택할 수 있게 한다. 또한 목적 언어 코퍼스에서의 번역어의 통계치를 이용하여 번역 사전 작성에 도움을 줄 수 있다.

본 논문에서는 위와 같은 정보들을 대상 영역의 코퍼스에서 분석하고 분석된 정보를 보기 쉽게 제시하는 번역 사전의 특정 영역화 워크벤치를 설계하고 구현한다. 다음 절에서는 특정 영역화를 위한 워크벤치를 소개한다.

3 특정 영역화를 위한 워크벤치

특정 영역화를 위한 워크벤치의 전체 구성은 그림 2와 같다. 워크벤치는 대상 영역 코퍼스와 의미망을 이용하여 사용자가 일반 영역 사전을 대상 영역에 맞게 특정 영역화를 할 수 있는 환경을 제공한다. 워크벤치는 영역 지식 추출기와 영역 지식 검색기, 사전 편집기로 구성된다. 영역 지식 추출기는 대상 영역 코퍼스를 분석하여 2절에서 나열한 지식들을 추출한다. 추출된 영역 지식은 영역 지식 검색기를 통하여 사용자에게 보기 쉬운 형태로 제시된다. 사용자는 제시된 정보를 참조하여 사전 편집기를 통하여 번역 사전 내용을 수정한다. 아래에서는 각각

에 대해서 자세히 설명한다. 본 논문에서는 일반 영역의 사전을 군사 정보 영역으로 특정 영역화하는 경우를 예제로 사용한다.

3.1 영역 지식 추출기

그림 3은 대상 영역 지식을 추출하는 과정을 나타낸다. 영역 지식 추출기에서 추출되는 정보는 크게, 어휘 분석을 위한 어휘 정보와 복합어 정보, 구문 분석을 위한 숙어 정보, 변환과 생성을 위한 문맥에 나타난 공기 단어의 클러스터, 번역어 통계 정보로 구성된다.

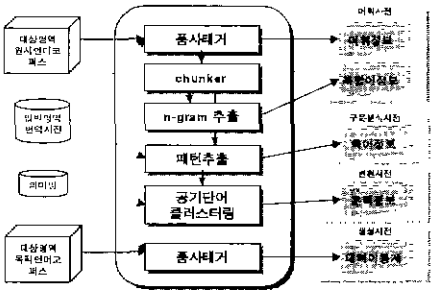

그림 3. 대상 영역 코퍼스로부터 지식 추출

3.1.1 어휘 정보의 추출

어휘 정보의 추출에서는 품사 태깅 작업을 통하여 각 단어의 활용형과 품사 등에 대한 통계치를 제공한다. 아래 표는 군사 정보 영역 코퍼스에서 추출한 'company'의 각 품사별 활용형에 대한 통계치를 보인다. 표와 같이 추출된 정보에 의해 워크벤치 사용자는 군사 영역 코퍼스에서 'company'는 주로 명사로 이용된다는 사실을 파악할 수 있다.

company	-	-s	-ed	-ing
noun	193	291	0	0
verb	0	0	0	0

어휘 정보 추출을 위해 사용한 품사 태깅 시스템이 대상 영역 이외의 코퍼스에서 추출한 통계치를 이용한다면 추출된 정보를 신뢰할 수 없다. 워크벤치의 영역 지식 검색기에서는 용어 색인(concordance) 기능을 제공하므로 사용자는 추출된 정보와 코퍼스 용어 색인, 두가지 정보를 동시에 참조하여 사전 변경 여부를 결정할 수 있다

3.1.2 복합어 추출

복합어는 정규 표현을 이용하여 명사구를 중심어순으로 정렬하는 Termight[6]의 방식을 이용하여 추출한다. 이때 명사구 인식(NP chunking)은 [7]의 방식을 따른다. 이에 덧붙여, 본 논문에서는 대상 분야에서 많이 나타나는 패턴을 처리하기 위한 규칙을 추가하여 성능 향상을 꾀한다. 예를 들어 'Reconstruction Bank of Korea', 'Under Secretary of Commerce'와 같이 고유 명사가 연속된 패턴들은 다음과 같은 규칙으로 처리된다.

name → (nnp|nnps|i)+ ((in|of) (nnp|nnps|i)+)*

복합어 추출 과정을 통해, 'company'에 대해서는 아래의 표와 같은 결과가 얻어진다. 워크벤치 사용자는 워크벤치에서 재시한 복합어 중에서 복합 명사로 등록할 것들을 선택하여 번역 사전에 반영할 수 있다. 만일 사용자가 'tank company'를 사전에 등록하고 번역어로 '전차중대'를 입력한다면, 'tank'와 'company' 각각에 대한 번역어 선택 과정을 통해 '탱크 회사'로 번역되는 문제도 해결된다.

빈도	분석된 복합어
20	oil company
2	Western oil company
4	security company
3	tank company

3.1.3 숙어 추출

숙어 추출의 과정은 크게 세 단계로 구성된다.

step 1 명사구 및 동사구를 인식(chunking)하고, 동사구는 중심어의 원형만을 남긴다.

step 2 [8]에서 제시한 방법으로 단어의 n-gram을 구한다. (n-gram은 인식된 명사구를 명사구의 중심어로 바꾼 것과 명사구를 구문 심볼 NP로 바꾼 것 두가지에 대해서 추출한다.)

step 3 추출된 n-gram, 즉 구문 패턴에 적당한 점수를 부과한다. 이 때 정량적인 중요도를 고려하기 위해 빈도수를 사용하고, 단어간의 결합성으로 MI(Mutual Information)을 이용한다.

위와 같은 과정으로 'has made increasing use of the area'에 대해서는 make use of area, make use of NP, make NP of NP 등을 숙어로 추출할 수 있고, 이러한 패턴들의 빈도수를 이용하면 대상 영역에서 자주 발생하는 숙어들을 추정할 수 있다. 아래는 'bear'에 대해서 얻어진 숙어 정보 중에서 명사구를 구문 심볼로 대체한 경우들의 예이다.

빈도	분석된 숙어
20	bear/v NP
9	bear/v on NP
4	bear/v out of NP
2	bear/v in NP
	. .

3.1.4 공기 어휘 클러스터링

숙어 추출에 의해 얻어진 구문 패턴은 하나 이상의 의미를 가질 수 있다. 공기 어휘 클러스터링에서는 구문 패턴이 가지는 공기 어휘를 클러스터링하여, 사용자가 각 패턴의 번역어를 판단하기 쉽게 지원하고, 동시에 각 번역어로 선택되기 위한 공기 어휘 정보를 제공한다.

본 논문에서는 공기 어휘를 클러스터링할 때 모든 단어 쌍에 대해 유사도를 한꺼번에 고려하지 않고, 클러스터의 중심 어휘(seed)를 먼저 선별하여 추출한 뒤에 그 외의 단어들이 중심 어휘에 대하여 가질 수 있는 유사도를 계산하여 클러스터를 결정하는 방법을 이용한다. 중심 어휘를 추출하는 방법은 [8]에서 제시한 선택적 샘플링(selective sampling)을 응용한다.

아래의 표는 conduct NP 구문 패턴의 명사구 위치에 나타나는 단어들에 대해서 중심어휘를 추출하고 이를 이용하여 클러스터링을 수행한 결과를 보인다. 이 분류를 바탕으로 'conduct NP'는 대상 영역에서 주로 '수행하다'의 의미로만 쓰인다는 것을 쉽게 분석할 수 있다.

패턴 : conduct NP

중심어휘	공기어휘
exercise	exercise, activity, programme, work, survey, investigation, analysis, research, trial
strike	strike, attack, mission, operation
firing	*firing*
raid	raid
evaluation	evaluation, review, study, test

3.1.5 번역어 통계 정보

아래는 일반 영역 사전에 등록된 영어 단어 '*company*'의 각 번역어가 군사 영역 코퍼스에서 발생된 빈도를 보인다.

의미	번역어(빈도)
company[1]	동료(89), 친구(124)
company[2]	손님(38), 방문자(17)
company[3]	교제(6), 사교(2)
company[4]	회사(426)
company[5]	중대(487)

이 정보를 이용하면 유사한 의미를 가지고 있는 두 단어 '교제'와 '사교' 중에서 '교제'라는 단어가 대상 영역에서 많이 이용한다는 정보와 함께, 'company'의 여러 번역어 중에서 '회사'와 '중대'라는 표현이 대상 영역에서 많이 발생하다는 등의 정보를 알 수 있다. 이 때 정보를 추출한 코퍼스가 양언어가 정렬된 코퍼스가 아니므로 'company'라는 단어가 '회사'와 '중대'로 자주 번역되었다고 판단할 수는 없다. 하지만, 번역어를 사전에 기술할 때 어떤 단어를 기본 번역어로 기술한 것인가 등의 대상 영역의 목적언어 특성을 파악하기 위한 자료로 이용할 수 있다.

위와 같은 방법으로 추출된 영역 지식은 영역 지식 검색기를 통해 워크벤치 사용자에게 제시되고 사용자는 사전 편집기를 통해서 특정 영역화 작업을 수행하게 된다. 영역 지식 검색기와 사전 편집기는 하나의 사용자 인터페이스로 통합되어 구성된다. 다음에서 각각에 대해서 설명한다.

3.2 영역 지식 검색기

그림 4는 사용자 인터페이스에서 영역 지식 검색을 위한 부분을 보인다

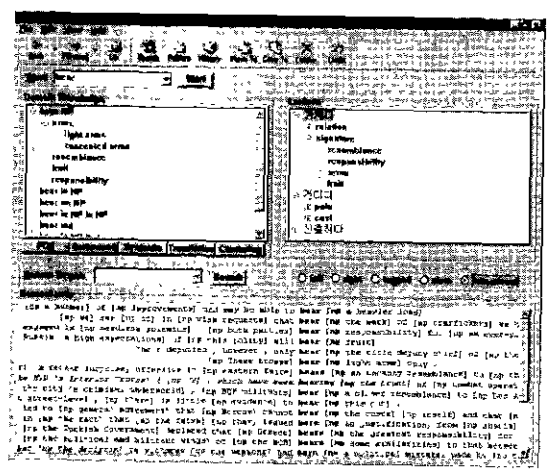

그림 4: 워크벤치 인터페이스

인터페이스는 영역 지식 보기, 사전 내용 보기, 코퍼스 용어 색인(concordance)의 세개의 창으로 구성된다.

그림 왼쪽 위의 Domain Knowledge 창에서는 영역 지식 추출기에서 분석된 영역 지식을 제시한다. 제시되는 정보는 품사(POS), 복합어(Compound) 구문 패턴(Syntactic), 공기어휘 클러스터(Clustering), 번역어 통계 정보(Translation)로 나뉜다. 그림 4에서는 'bear'에 대한 구문 패턴 정보를 보인다. bear NP를 선택하면 실제 대상 영역 코퍼스에서 명사구 NP의 위치에 자주 나타나는 어구를 계층적으로 보여주는 것을 볼 수 있다.

그림 4의 오른쪽 위의 Lexicon 창에서는 현재 사전의 내용을 보인다. 그림 4에서는 왼쪽 창에서 선택된 bear NP 에 대한 현재 사전 내용을 보인다. bear NP에 대해서 명사구에 'relation'과 'signature'와 유사한 단어들이 발생하면 '가지다'라는 번역어를 가지는 것으로 현재 사전이 작성되어 있음을 나타낸다.

그림 4의 아래쪽 Concordance 창은 대상 영역 코퍼스에서 특정 단어가 발생하는 위치를 찾아 해당 단어를 기준으로 정렬하여 단어의 문맥을 보여준다. 이 때 문맥의 좌/우 정렬 여부, 품사별 정렬 여부, 활용형별 정렬 여부, 원형 복귀 여부를 선택사항으로 지정할 수 있다. 또한 패턴 검색을 통해 어휘, 품사, 구문심볼, 단어의 원형, 구문의 중심어 등을 이용하여 검색할 수도 있다 그림 4에서는 bear NP에 대한 코퍼스 문맥 정보를 보인다. 이 정보는 왼쪽 위의 창에 제시된 bear NP를 선택하거나, 패턴 검색에서 '$bear %NP'를 직접 입력하는 두가지 방식 모두를 통해서 얻을 수 있다($는 원형을 표시하는 태그이고 %는 구문심볼임을 표시하는 태그이다).

그림에서 볼 수 있듯이 지식 검색기에서는 어휘, 구문, 번환과 생성에 대한 영역 지식을 계층구조로 표시하여 사용자가 영역 지식을 효율적으로 탐색할 수 있도록 한다. 또한 계층 구조상의 노드를 클릭할 때마다 해당되는 현재 사전 내용과 문맥 정보를 자동으로 검색하여 보여준다.

영역 지식 검색기를 통해서 영역 지식을 살펴본 워크벤치 사용자는 사전 편집기를 이용하여 사전을 직접 수정하게 된다. 사전 편집기는 그림 4의 인터페이스에서 메뉴 선택을 통해 실행된다. 다음에서는 사전 편집기에 대해서 설명한다.

3.3 사전 편집기

사전 편집기에서는 기본 기능으로 사전 엔트리의 검색, 수정, 삽입, 삭제 기능이 제공된다. 이에 더하여 효율적인 작업과 안전성을 위해 유사 예제 제시와 사전 작업의 안전성을 보장하기 위한 기능을 지원한다.

사전편집 과정에서 가장 많은 시간을 요구하는 작업은 새로운 항목의 추가이다. 이를 위해 워크벤치에서는 기본 사전에 있는 내용 중에서 새로 입력하려는 항목과 가장 유사한 예제를 제시하여 사용자가 쉽게 사전 항목을 추가할 수 있게 돕는다 유사도 판정기준으로는 어휘적, 구문적, 의미적 유사성을 고려한다. 예를 들면 armoured personnel carrier라는 복합어를 등록하기 위해서는 'personnel carrier', 'carrier', 그리고 의미망에서 각 의미노드에서의 상위어에 대한 사전의 내용을 제시한다. in search of NP의 경우 품사 패턴을 이용하여 'in favor NP'나 'in behalf of NP' 등을 제시한다. 그림 5는 유사 예제 제시의 예를 보인다. tank company에 대해서 'company', 'institute', 'unit' 등에 대한 사전 내용을 볼 수 있게 지원하고, 사용자는 'company'가 '중대'로 번역되는 예제를 참조하여 사전을 작성할 수 있다.

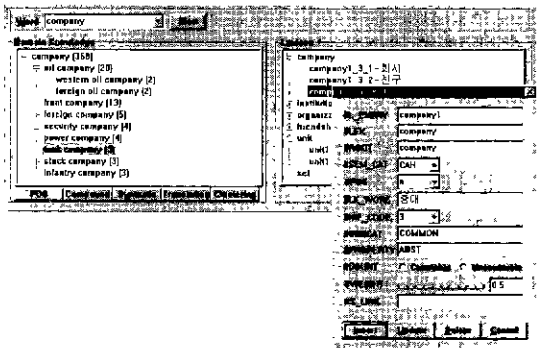

그림 5: 사전 편집기

인터페이스에서는 사전 변경이 이루어질 때 이로 인하여 전체 사전의 일관성(consistency)이 유지되는지를 점검해 주기 위한 기능을 지원한다. 의미분별을 위한 공기어휘 추가 및 변경의 경우 코퍼스에서 추출된 공기 어휘를 이용한 점검을 통해 추가된 지식으로 인해 기존에 형성된 일관성이 깨어지는 것을 방지할 수 있다.

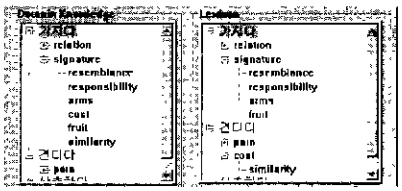

그림 6: 'bear NP'에 대한 일관성 유지 점검 예

그림 6은 bear NP에 대한 일관성 점검 예를 나타낸다. 영역 지식창에서는 대상 영역 코퍼스에서 'bear'의 목적어로 'similarity'가 발생하는 것을 보인다. 우측 창에서는 목적어에 'relation'이나 'signature'가 오면 'bear'를 '가지다'로, 'pain'이나 'cost'가 오면 '견디다'로 번역되도록 현재 사전이 구성되어 있음을 보인다. 이 때, 그 하위 구조는 의미망을 이용하면 'similarity'가 'signature'보다 'cost'와 더 유사하다고 의미망에서 판단됨을 나타내고 있다. 사용자는 이를 보고 'bear similarity'가 '가지다'로 번역되기 위해서는 'similarity'도 사전의 번역어 선택을 위한 공기어휘로 등록해야 한다는 사실을 알 수 있고, 사전 내용을 변경하여 올바른 번역을 얻을 수 있다.

4 실험 및 평가

사전 특정 영역화에 대한 기존의 연구는 코퍼스로부터 새로운 사전을 구축하는 방법[3]이나 양국어 정렬된 코퍼스

를 이용하는 방법[4], 용어 사전과 코퍼스에서 새로운 어휘 사전을 구축하는 방법[5] 등이 있다. 이러한 기존의 연구들에서는 사전 작성이나 코퍼스 분석 등의 개별적인 작업을 위한 도구만을 구축했을 뿐, 코퍼스 분석과 사전 점검, 사전 편집 등의 통합 환경을 제공한 경우는 찾아볼 수 없다. 또한 기존의 연구들에서는 정렬된 코퍼스와 같이 획득하기 어려운 지식원을 사용하지만, 본 논문에서는 기존의 번역 사전, 의미망, 독립적인 원시 언어 코퍼스와 목적 언어 코퍼스 등 비교적 획득하기 쉬운 지식원을 이용하면서도 특정 영역화를 위한 정보를 효율적으로 추출하는 방법을 제시했다.

본 논문에서는 평가를 위하여 군사 정보 분야의 전문 잡지인 "Jane's Intelligence Review"의 4만여 문장을 대상 영역 코퍼스로 사용하여, 롱맨영영한 사전을 기본으로 기술된 일반 번역 사전을 군사 정보 분야로 특정 영역화해 보았다. 평가는 6명의 사전 작성 전문가를 3명씩 2그룹으로 나누어 워크벤치의 통합환경을 사용하지 않는 경우와 사용하는 경우에 대한 작업 성능 비교로 이루어졌다.

먼저 사전 작업의 편의성을 살펴보기 위해 복합어 추가 작업을 대상으로 사용자 편의성과 작업 효율에 대해서 평가했다. 본 논문에서 제시한 워크벤치를 이용한 경우 코퍼스를 분석하여 대상 영역의 복합어 후보들을 찾아서 사용자가 보기 쉬운 형태로 보여주고, 유사 예제를 제시하여 사전 작성 작업을 돕는다. 30개의 복합어 추가에 대한 실험 결과에서 개별적인 도구를 이용하여 코퍼스를 분석하고 사전을 편집하는 경우에 비하여 워크벤치를 사용한 경우에 작업 시간이 약 82% 정도 감소하는 것으로 나타났다.

일관성 검사의 유용성을 알아보기 위해서는 10개의 'v NP' 구문에 대하여 코퍼스로부터 각각 15개의 명사구의 중심어를 추출하여 사전 문맥정보에 의한 의미분석을 수행했다. 잘못 분석된 단어들에 대해 문맥정보 추가작업을 수행할 때, 일관성 검사를 통하여 검사를 하지 않은 경우에 비해 27%의 의미분별 정확성 향상이 있었다.

살펴본 바와 같이 워크벤치를 통해 사용자를 위한 작업환경의 편의성, 사전 작업의 효율성, 사전 작업의 안전성에서 향상이 있었다.

5 결론

본 논문에서는 일반 영역 기계번역 사전을 특정 영역화하기 위한 워크벤치를 설계하고 구현하였다. 이를 위해 특정 영역화의 대상이 되는 정보들과 이에 필요한 작업들을 분류하고, 각 정보를 자동으로 추출하는 방법을 고안하였다. 워크벤치는 자동으로 추출된 정보를 보기 쉬운 형태로 제시하고 사용자는 추출된 정보에 기반하여 사전을 특정영역화 할 수 있다. 실험결과 워크벤치를 통해 기존의 단순반복적이거나 불필요한 작업을 줄일 수 있고 사전에 반영되는 정보의 정확성 및 일관성을 향상시킬 수 있었다.

참고문헌

[1] 노윤형, "대상 영역 코퍼스를 이용한 번역 사전의 특정 영역화를 위한 워크벤치," 한국과학기술원 전산학과 석사학위논문 2000

[2] 임철수, 이현아, 최명석, 장병규, 이공주, 김길창, "어휘화된 규칙에 기반한 영한 기계번역시스템," 한국정보과학회 추계학술대회 발표 논문집, 1997

[3] Larie Gerber and Jin Yang, "Systran MT Dictionary Development," *In Proceedings, Machine Translation Summit IV*, 1997.

[4] Sayori Shimohata and Toshiki Murata and Atushi Ikeno, "Machine Translation System PENSEE System Design and Implementation," *In Proceedings, Machine Translation Summit VII*, 1999.

[5] Jan W. Amtrup and Karine Megerdoomian, "Rapid Development of Translation Tools," *In Proceedings, Machine Translation Summit VII*, 1999

[6] Ido Dagan and Ken Church, "Termight: Coordianting Humans and Machine in Bilingual Terminology Acquisition," *Machine Translation (MT).12(1):89–107*, 1997.

[7] Steven Abney, "Rapid Incremental Parsing with Repair," *In Proceedings, the 6th New OECD conference Electric Text Research, pages 1–9* , 1990.

[8] Nagao, M and S Mori, "A New Method of n-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese," *In Proceedings, International Conference on Computational Linguistics (COLING), pages 611–615*, 1994.

[8] Atsushi Fujii, "Corpus-Based Word Sense Disambiguation," *Phd. thesis, Department of Computer Science, Tokyo Institute of Technology*, 1998.