제한된 도메인을 위한 코퍼스 기반의 하이브리드 번역 시스템

(A Corpus-based Hybrid Translation System for Limited Domain)

강 운 구 † 김 성 현 †† 이 병 문 ††† 이 영 호 †††
(Un-Gu Kang) (Sung-Hyun Kim) (Byung-mun Lee) (Young-Ho Lee)

요 약 본 논문은 RBMT, SMT, PBMT를 활용한 직렬 연결 방식의 하이브리드 번역 시스템을 제안한다. 번역 시스템은 입력된 문장에 대하여 구문 분석을 진행한 후, 이 정보를 바탕으로 구문 변환과 개체명 인식을 한다. 이 결과값을 의사 문장으로 변형, 문장 분리 규칙이 적용 가능할 경우, 분리된 문장에 대하여 다중 디코딩을 수행하고, 후처리기에서 접합 규칙에 따라 번역문을 생성하였다. 실험을 통하여 어순 배치의 경우 distortion 모델에 의존하지 않고 구문 변환(rule-based syntactic transfer) 규칙을 사용하는 것이 더욱 효과적인 것으로 나타났다.

키워드 : 하이브리드 번역 시스템, 절 단위 디코딩

Abstract This paper proposes a hybrid machine translation system which integrates SMT, RBMT, and PBMT in serial manner. SMT in our project has been implemented as a quasi-syntax-based system where monotone search is done, given a preprocessed string of foreign language. Preprocessing includes rule-based reordering, NE recognition, clausal splitting, and attaching pattern translation information at the end of the input text. For lengthy & complex sentences, clausal splitting turned out to generate better translation than normal input.

Key words : Hybrid machine translation, clausal splitting

1. 서 론

규칙기반 기계번역(RBMT:Rule-Based Machine Translation)에서는 원문의 언어학적 특징을 분석하여 문형을 분류하고 구문/어휘 변환규칙(transfer rule) 적용 프로세스를 거쳐 번역문을 생성하는 방법이 사용된다. 규칙 기반 시스템의 장점이 원문을 구문 분석(POS tagging and parsing)한다는 점에 있지만, 언어의 의미적 중의성(ambiguity)에 취약하고, 유지보수가 어렵다는 단점이 있다. 예를 들어, 번역 후보 단어가 다수일 때에 대비하여 적용범위가 넓은 단어 선택 조건을 어휘집(lexicon)에 미리 기입해주어야 하고, 새로운 도메인 적응도 쉽지 않다. 또한, 생성된 번역문이 자연스럽지 못한 것이 문제점으로 지적되어 왔다. 이러한 규칙기반 번역 시스템의 한계를 인식하고, 이 한계를 극복하기 위하여 도입된 것이 통계기반 번역 시스템(SMT: Statistical Machine Translation)이다.

번역규칙과 어휘집을 수동으로 구축하는 이전의 방식과 달리, SMT에서는 통계적 모델링 기법을 통해 병렬 코퍼스로부터 번역 모델을 훈련하고, 이 모델을 근거로 번역을 수행한다.

초기의 순수 데이터 기반 SMT는 규칙기반 번역 시스템의 문제점들을 극복하기 위한 대안으로서 제시되었다. 번역 지식을 언어전문가의 도움 없이 빠른 시간 안에 구축할 수 있고, 생성된 번역문이 비교적 자연스럽다

† 비 회 원 : 가천의과학대학교 정보공학부 교수
ugkang@gachon.ac.kr
†† 정 회 원 : (주)엘앤아이소프트 개발팀 선임연구원
hyunkim@lnisoft.com
††† 정 회 원 : 가천의과학대학교 정보공학부 교수
bmlee@gachon.ac.kr
leeyh@gachon.ac.kr
(Corresponding author임)
논문접수 : 2010년 1월 11일
심사완료 : 2010년 9월 10일

정보과학회논문지: 소프트웨어 및 응용 제37권 제11호(2010.11)

는 것이 강점이었다. 그러나 원문을 언어학적으로 깊게 분석(parsing, sense tagging)하여 얻은 정보를 번역 과정에서 활용하지 못하고, 번역과정에서 발생하는 원문-번역문 대응 단어의 어순 변경 이벤트에 대한 feature function인 distortion(reorder) 모델은 영어와 한국어와 같이 구문론(syntax)적으로 상이한 언어간 번역에서 발생하는 전역 어순 변경(global distortion)을 적절하게 포착하지 못하는 것이 문제점으로 제기되어 왔으며, 이 문제는 문장의 길이에 비례하여 발생한다.

이러한 문제점에 대한 인식을 바탕으로, 본 논문에서는 SMT와 RBMT, 그리고 패턴기반 기계번역(PBMT: Pattern-Based Machine Translation) 방법론을 결합하여 제한된 도메인을 위한 코퍼스 기반의 번역시스템 구축 방법론을 제안한다.

본 논문은 총 5장으로 구성되어 있다. 2장에서는 관련 연구로서 구문 정보를 활용하는 SMT에 대하여 소개하고, 3장에서는 하이브리드 번역 시스템의 구축 방법을 제안한다. 그리고 4장에서는 실험을 통해 제안한 시스템에 대한 성능을 평가하고, 5장에서는 결론을 맺고 향후 연구에 대하여 검토한다.

2. 관련 연구

1990년 IBM에서 발표한 "Statistical Approach to Machine Translation" 논문[1] 이후 SMT에 대한 논의가 본격화 되었다. 번역규칙과 어휘집을 수동으로 구축하는 이전의 방식과 달리, SMT에서는 통계적 모델링 기법을 통해 병렬 코퍼스로부터 번역 모델을 훈련하고, 이 모델을 근거로 번역을 수행한다. SMT에서의 번역 과정은 아래와 같은 최적화 프로세스로 표현된다

$$\hat{k}_1^I = \arg\max_{k_1^I} \{\Pr(k_1^I | e_1^J)\}$$

식 (1) 런타임 최적화 프로세스

영어 문자열 $e_i^I = e_1, \ldots, e_i, \ldots, e_I$가 입력으로 주어지면, 번역문일 확률이 가장 높은 한국어 토큰열 $k_i^J = k_1, \ldots, k_j, \ldots, k_J$를 생성한다. argmax는 상기 확률 모델(objective function: 목적 함수) 스코어를 최대화하는 한국어 토큰 집합(an optimal set of sub-solutions)을 찾아내라는 의미로서, 디코딩(run-time optimization)을 나타낸다. Shannon의 잡음 채널 모델로 알려진 초기 SMT의 번역 프로세스는 번역 모델(translation model)과 언어모델(language model)로 분해된다. 번역모델은 소스 단어가 목적 언어의 특정 단어로 번역될 확률을, 언어 모델은 생성된 목적 언어 단어들의 어순이 얼마나 문법적인가를 판단할 수 있는 확률을 각각 제공한다.

$$\hat{\theta} = \arg\max_{\theta} \left\{ \prod_{s=1}^{S} p_\theta(e_s | k_s) \right\}$$

$$\hat{\gamma} = \arg\max_{\gamma} \left\{ \prod_{s=1}^{S} p_\gamma(k_s) \right\}$$

식 (2) 최적 번역/언어 모델 스코어를 갖는 번역문 탐색

일반적인 SMT 개발 과정은 그림 1과 같다. 병렬코퍼스는 원문과 번역문이 1:1 대응하는 텍스트 데이터이며 번역지식(번역모델, 언어모델)을 학습하기 위한 자료로 활용된다. 코퍼스는 일반적으로 번역가에 의해 수동 구축되거나 web crawler또는 aggregator를 통해 반-자동으로 수집한다. 병렬코퍼스는 번역모델을 훈련하기 위해 필요하며, 단일 코퍼스는 목적언어의 단어배열순서확률(well-formedness, 경험적 문법) 스코어를 부여하는 언어모델을 훈련하는데 필요하다. 학습이란 번역 이벤트에서 확률 모델들을 유도하고, 개별 모델들의 최적 가중치를 결정하는 최적화 과정이며, 모델들은 번역모델, 어순 재배치(reorder) 모델, 언어 모델 등을 포함한다.

보다 최근의 방식은 상기 기본 모델뿐만 아니라 번역 과정에 참여하는 factor(mapping)들을 모두 모델링 하여 가중치에 따라 선형 결합하는 로그선형모델기반(log-linear) 아키텍처를 기본적으로 채택하고 있다. 로그선형 모델은 최대엔트로피[2,3]모델로도 알려져 있다.

$$\Pr(k_1^I | e_1^J) = p_{\lambda_1^M}(k_1^I | e_1^J) = \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(k_1^I | e_1^J)]}{\sum_{\tilde{k}} \exp[\sum_{m=1}^{M} \lambda_m h_m(\tilde{k}_1^I | e_1^J)]}$$

식 (3) 로그 선형 모델의 번역확률

h는 개별 factor (mapping)을 모델링 하는 휴리스틱 함수(heuristic feature function)를 나타이며, 이 함수집합 안에 번역모델과 언어모델이 포함되어 있다. 번역 프로세스에 영향을 미치는 것으로 판단되는 factor(=feature)의 확률 프로세스를 모델링하여 상기 프레임워크에 추가하게 된다. λ는 개별 자질함수의 가중치를 나타내는 모델 패러미터이고, M은 자질함수의 최대 개수이다.

$$\hat{\lambda}_1^M = \arg\max_{\lambda_1^M} \{\sum_{s=1}^{S} \log p_{\lambda_1^M}(k_s | e_s)\}$$

식 (4) 최적 모델 가중치 학습

다수의 모델(feature function)들이 번역문에 스코어

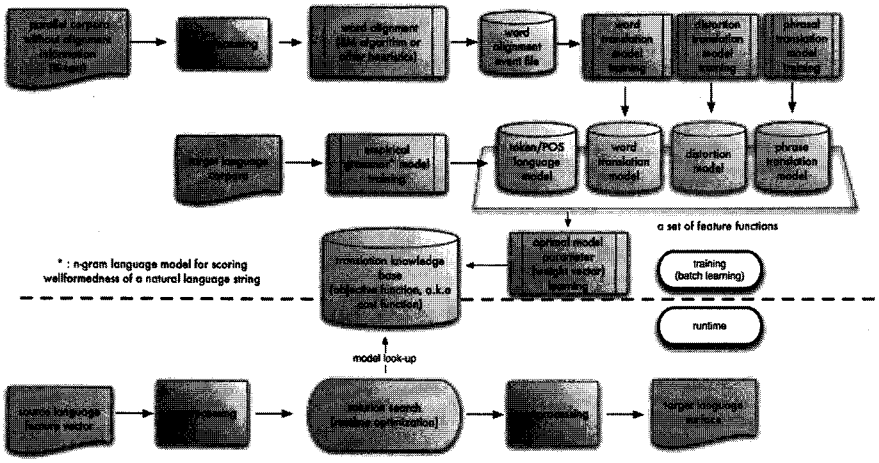


그림 1 통계 기반 번역 시스템의 일반적인 구축 과정

를 부여하므로 개별 모델들의 최적(optimal) 가중치 벡터를 학습(optimal model parameter tuning)하는 정규화 과정(regularization)이 필요한데, 현재에는 MERT (Minimum Error Rate Training)[3]가 널리 사용되고 있고, 관련 학습 프로그램이 공개되어 있다.

$$\hat{k}_1^{I} = \arg\max_{k_1^{I}} \{\sum_{m=1}^{M} \lambda_m h_m(k_1^{I}, e_1^{I})\}$$

식 (5) 로그 선형 모델 기반 번역 디코더 아키텍처

식 (1)과 마찬가지로 로그 선형 모델 디코딩도 입력문 e_1, e_2,...e_I를 생성하는데 가장 큰 영향을 미친 인자 k_1, k_2,...k_I로 이루어진 토큰 열을 그림 1의 다양한 최적화 알고리즘(run-time optimization algorithms)을 통해서 찾아내며, 출력 토큰 열 $\hat{k}_{1...I}$는 모델의 확률을 최대화하는 근사치로서 번역 시스템의 출력문이 된다.

초기의 순수 데이터 기반 SMT는 번역 지식을 언어 전문가의 도움 없이 빠른 시간 안에 구축할 수 있고, 생성된 번역문이 비교적 자연스럽다는 강점을 내세웠으나 원문을 언어학적으로 깊게 분석(parsing, sense tagging)하여 얻은 정보를 번역 과정에서 활용하지 못하는 점, 한국어와 같이 형태론적으로 복잡한 단어의 표층 생성(surface string generation) 과정의 단순함, 번역 이벤트에서 발생하는 원문-번역문 대응 단어의 어순 변경 이벤트에 대한 feature function인 distortion(reorder) 모델은 영어와 한국어와 같이 단어 구문론적으로 상이한 언어간 번역에서 발생하는 전역 어순 변경(global distortion)을 적절하게 포착하지 못하는 점 등이 문제점으로 제기되어 왔다.

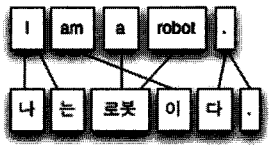


그림 2 로컬 어순 변경 이벤트의 예

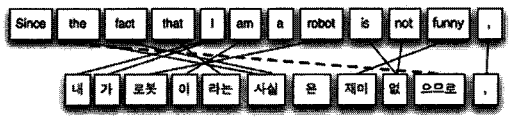


그림 3 글로벌 어순 변경(굵은 점선) 이벤트의 예

언어학적 분석 정보를 활용하는 이유는 여러 가지가 있는데, 본 연구에서는 긴 중문/복문을 절 단위로 분리하여 분리된 절 만큼 다중 디코딩을 함으로써 탐색 에러(search error) 가능성을 줄이고, 입력된 영어 중/복문에 인간이 직접 구축한 어순 변경 규칙을 적용함으로써 전역 distortion 확률 모델을 디코딩 시 사용하지 않거나 제한하는 방식을 활용한다. 본 논문에서 제안하는 방법론과 대상 언어 쌍이 동일하지 않지만, "clausal restructuring"을 통해 SMT의 약점을 보완하고자 하는 목적성 측면에서 유사한 연구 사례는 다음과 같다.

Lee et al.[4]은 한국어 원문의 품사 열(POS tag sequence) 정보를 바탕으로 장문을 적절한 위치에서 분리하여 단문으로 만드는 TBL(Transformation-based Learning) 기법을 제안하였다. 중/복문이 단문으로 변형되면 가설공간을 줄여 탐색 에러를 줄일 수 있기 때문이다.

구문 분석 정보를 획득, 원문의 어순을 목적 언어와 유사하게 재배치함으로써, global distortion을 최소화하는

방법도 제안되었다. Collins et al.[5]은 German-English 번역에서 원문을 구문 분석한 후, 몇 가지 reordering 규칙을 적용하는 방식으로 완벽하지는 않지만 distortion 문제를 어느 정도 해결할 수 있음을 보여주었다.

Xia and McCord[6]는 파스 트리 상에서 동일한 부모를 갖는 노드 배열이 그 순서를 바꿈으로써 global distortion을 최소화하는 규칙을 학습하는 방법을 제안하였다.

3. 하이브리드 번역 시스템 개발

3.1 필요 모듈 개발

병렬 코퍼스로부터 원문, 번역문 단어 대응 정보를 찾아내고, 번역 우도(likelihood)를 추정하는데 사용되는 단어 정렬 과정은 연구 초기에 오픈 소스인 giza++[1]에 의존하였으나, 중반에 자체 단어 정렬기인 Hermes를 개발하였다. 현재Hermes의 경우 IBM1~3 정렬 알고리즘과 HMM 정렬 알고리즘[12]이 구현되어 있는 상태이며 추가적인 연구를 통해 통사정보와 사전(prior) 정렬 지식을 활용할 수 있도록 성능향상을 진행할 예정이다.

최적 번역문을 생성하는 번역 코어 프로세스는 오픈 소스인 Moses나 Phramer등을 사용하지 않고, Qualia 번역 디코더를 설계하여 직접 구현하였다. Qualia에는 구 기반 beam search[7,8], finite state transducer[9]에 의한 탐색 알고리즘이 적용되어 있으며, 이러한 최적화 알고리즘은 계속적으로 쉽게 추가할 수 있도록 설계하였다.

번역 시스템을 구축하기 위해서는 단어 정렬기와 디코더 이외에도 많은 모듈들이 필요하다. 문장 분리기, 토큰 분리기, 개체명 인식기(named entity recognizer), 전/후 처리기, 품사 태거, 형태소 분석기, 구문 분석기, n-gram 언어 모델[10] 훈련기, 패턴 템플릿 추출기 및 패턴 템플릿 번역 모듈 등이 대표적이다. 영어 구문 분석기[2]를 제외한 나머지 모듈들은 모두 자체 제작하였다.

하이브리드 영어 품사 태거는 TNT Tagger[3]의 아키텍처에 근거하여 구현하였고, LDC(Linguistic Data Consortium)에서 유료로 제공하는 Penn Treebank를 활용하여 태깅(transition, emission) 모델을 훈련시켰다.

한국어 형태소 분석기는 세종계획 코퍼스를 이용, trigram 모델을 훈련하였고, 규칙과 확률모델을 결합한 하이브리드 아키텍처를 갖고 있다. 형태소 분석기를 사용하는 것은 단어 정렬(bitext alignment[11-13])시 한국어의 정렬 단위가 어절이 아닌 형태소 단위로 만들기 위해서이다.

1) http://fjoch.com/GIZA++.html
2) http://www.cs.brown.edu/~ec/
3) http://www.coli.uni-saarland.de/~thorsten/tnt/

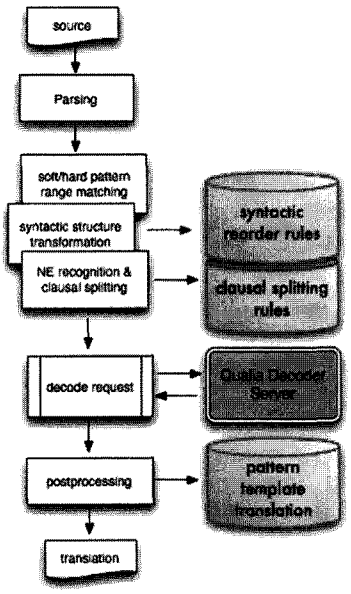


그림 4 번역 시스템의 프로세스

개체명 인식기는 한 개 이상의 토큰으로 구성된 고유명사 구간을 인식하여 단일의 의사단어(pseudo-word)로 치환하기 위해 사용하였다.

전처리기의 주요 역할은 영어 어순을 한국어와 가깝게 만들어 번역문 생성 과정이 monotone search 가 되도록 만드는데 있다. 따라서 distortion모델을 사용하지 않는 finite state transducer에 의한 디코딩 과정에서는 한국어 문장이 왼쪽에서 오른쪽으로 차례대로 생성된다.

후처리기에서는 생성된 의사 번역문을 원래의 형태로 복원하는 작업과 형태소 단위 토큰을 어절 단위로 붙이는 작업을 진행한다.

상기 모듈을 활용하여 번역 시스템을 구축하였다. 번역 시스템 내부의 디코더가 사용한 모델들은 다음과 같다.

- 5-gram Surface Language Model P(K)
- Translation Model P(E|K) for general domain
- Translation Model P(E|K) for equities market domain
- Reverse Translation Model P(K|E) for general domain
- Lexical Weight[14] Model Lex(K|E), Lex(E|K)
- Word penalty
- Phrase penalty
- NE Translation Model for equities market domain
- NP(POS only) Pattern Translation Model P(E|K)

단어, 구 단위 번역 모델은 giza++과 Koehn의 phrase extract/scoring 툴과 자체 제작한 Hermes를 활용하여

두 개의 독립된 번역 모델을 훈련시켰으나, 실제 테스트에서는 성능 상의 이유로 giza++기반 번역 모델을 사용하였다.

5-gram 한국어 언어 모델은 n-gram 모델 최대 우도 추정(maximum likelihood estimation) 및 Kneser-Ney 평탄화 알고리즘을 직접 구현하여 훈련시켰다. NE 번역 모델은 Financial Times에 포함된 개체명을 개체명 인식기로 추출한 후, crawler를 이용하여 웹에 존재하는 대응 번역어 후보들을 수집, Hermes로 NE 번역 모델을 훈련시켰다. NP 패턴 번역 모델은 3.9에서 설명한다.

3.2 하이브리드 번역 시스템 구축 방법론

하이브리드 시스템의 명확한 표준은 존재하지 않지만, 크게 두 가지 관점에서 정의 내릴 수 있다. 첫 번째는 하나 이상의 번역 시스템 구축 모델을 순차적 프로세스에 적용하는 것, 두 번째는 하나 이상의 번역엔진을 병렬로 통합하여 개별 번역엔진이 생성하는 번역문의 부분 최적 번역을 수집하여 단일 최적 번역문으로 결합하는 것이다.

본 논문에서는 RBMT, SMT, PBMT 번역 시스템 구축 모델을 직렬로 연결하는 방식을 선택하였다.

RBMT는 외국어 원문의 단어들의 품사를 결정하는 POS tagging, 그리고 구(phrase)의 계층 구조를 분석하는 구문 분석(syntactic parsing), 입력문의 토큰 수를 감소시키는 개체명 인식(NER), 그리고 이러한 정보를 통해 구문 구조의 변환(syntactic transfer: phrasal reordering) 과정을 포함한다.

SMT는 어휘 변환(lexical transfer)을 담당하고, 이 과정에서 패턴 번역이 적용되는 구간은 의미 중의성이 없는 것으로 간주 PBMT에 의한 번역문 생성 프로세스를 활용하였다.

이러한 직렬 구조를 사용한 이유는 시스템 구현의 난이도가 비교적 쉽다는 점을 들 수 있다. RBMT가 원문의 언어적 특성을 분석하는데 강점이 있기 때문에 SMT의 어순 변경 모델(reordering(=distortion) model)보다 안정적이다. 실제로 어순 변경 모델을 디코딩 과정에서 활용할 경우, 입력문이 길어질수록 번역문의 퀄리티가 떨어지는 것으로 관찰되었다.

번역문 생성(lexical generation) 측면에서는 SMT가 RBMT에 비해 강점을 갖는다. 다만 입력문이 길어질수록 디코딩 속도가 늦어지고 번역모델에 포함되지 않은 미등록어가 입력문에 들어있을 경우 좋은 번역을 기대하기 어렵다는 것이 단점이다.

이와 같은 문제를 보완하기 위해 패턴 번역 템플릿을 디코더에 추가하여 패턴 템플릿 구간에 대해서는 번역의 중의성이 발생하지 않도록 하였다. 또한 명사구 단위의 미등록어가 발생할 경우 번역 모델이 아닌 일반 사전에서 대역어가 생성되도록 시스템을 구축하였다.

3.3 규칙기반 어순 재배열

어순 재배열은 RBMT에서 수행하는 구문 구조 변환(syntactic transfer)을 말한다. 외국어와 번역문은 일반적으로 문법적 범주에 따라 단어/구 단위 순서 위치가 변경되기 마련인데, 구문 변환 단계에서 이 번역 이벤트를 처리한다.

어순 재배열 프로세스는 전처리기를 통해 구현하였다. 어순 재배열 프로세스는 다음과 같이 진행된다.

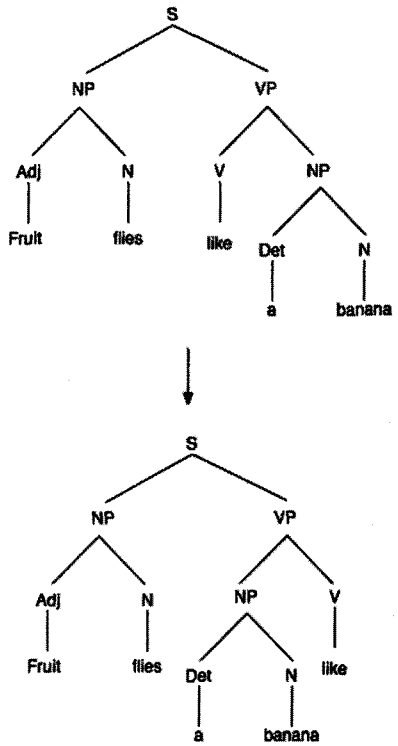


그림 5 어순 재배치

영어는 기본적으로 SVO, 한국어는 SOV 언어이기 때문에 재배치 규칙 사전을 구축하였다. 입력문의 통사 구조가 분석되면 parse 트리 정보가 전처리 서버에 전달되고, 트리의 개별 노드(node) 객체에 포함된 자식 노드 배열 정보를 확인, 어순 재배치 규칙에 적용되는지의 여부를 판단한다. 이 과정이 재귀적으로 진행되면 한국어 어순과 거의 일치하는 영어 문자열이 출력된다.

전처리기는 어순을 재배열하는 것뿐만 아니라 개체명 구간을 의사 단어(pseudo word)로 변환하고, 패턴 적용 가능 구간에 대한 추가 정보를 어순 변경된 문자열의 후미에 붙이는 작업도 동시에 수행한다.

전처리기의 결과는 디코더로 전송되어 최적화 과정을 거친다.

3.4 디코딩 프로세스

디코딩은 원문이 입력되었을 때, 가장 확률이 높은 번역문의 문자열을 출력하는 프로세스이다. 이 프로세스의 하이 레벨 알고리즘은 다음과 같이 요약된다.

디코딩 하이 레벨 프로세스
1. 외국어 토큰 열 입력
2. 부분 번역 이벤트 객체(노드) 생성
3. 환경설정에서 정의된 런타임 최적화 알고리즘 적용
4. 번역문 N-best 출력

번역 모델은 단어, 구 단위의 번역 및 번역 우도 정보를 포함한다. 이 때 '구'는 언어학적인 범주가 아니라 구 추출 알고리즘에 의해 생성된 단어의 연속체를 의미한다. n개의 토큰으로 이루어진 입력문은 n(n+1)/2 개의 구를 포함하며, 입력문이 주어지면 번역 모델로부터 입력문에 포함된 단어/구의 대응 번역 옵션 후보를 아래와 같이 획득한다.

표 1 번역 옵션 테이블의 예제

Source	Target	prob.
boy	소년	0.7234
boy	남자 아이 가	0.2341
the	그	0.6454
the boy	그 소년 은	0.7454
went	갔다	0.5533
away	멀리	0.2423
went away	떠났습니다	0.7234

디코딩을 진행하기 위해서는 상기 정보를 이용하여 부분 번역을 의미하는 노드 객체를 만들고, 두 노드 객체가 연결되었는지의 여부를 나타내는 매트릭스를 생성한다.

디코딩 프로세스는 multi-stack beam search과 finite state transducer 두 가지로 구현하였다. Beam search와 같은 heuristic 탐색은 monotone search가 아니기 때문에 부분 번역 가설들은 개별 iteration 마다 외국어 문장의 서로 다른 구간에 대한 번역을 진행한다. 따라서 현재 번역 우도가 높은 소스 구간을 먼저 번역한 부분 번역 가설 객체와 번역 우도가 낮은 소스 구간을 먼저 번역한 부분 번역 가설 객체를 비교하면 후자가 pruning 되는 경우가 발생한다. 이러한 문제를 해결하기 위해 부분 번역 확장(hypothesis expansion)시 현재까지 번역한 구간에 대한 번역 비용(prefix cost)와 아직 번역하지 않은 구간들에 대한 미래 비용(future cost)을 함께 고려한다[15]. 미래 비용은 Moses 시스템과 같이 Floyd Warshall all-pair shortest 알고리즘[16]을 통해 계산한다. 그리고 탐색 과정에서 하나 이상의 우선순위 큐(priority queue)를 사용했는데, 이는 현재 번역한 소스 문장의 토큰 수가 동일한 부분 번역 가설들 사이의 비교가 이루어지도록 하기 위함이다.

유한 상태 기계(finite state transducer)에 의한 디코딩에서는 어순 변경 모델(reorder model)을 사용하지 않았다. 이 방식에서는 전처리된 외국어 문자열을 이용하여 탐색 그래프를 만들고, monotone 탐색을 진행한 뒤, n-best 번역문을 생성하도록 했다. n-best 번역 후보를 만든 이유는 n-best reranker 를 디코더에 연결하기 위해서이다. n-best reranker 는 향후 휴먼 번역가의 추천에 의한 번역 시스템 성능 고도화 방법론을 연구하면서 구현할 계획이다.

3.5 절 단위 디코딩

절 단위 디코딩은 장문이 입력될 경우 최적 번역문을 생성할 가능성이 낮아질 것이라는 가설에 근거한 시도이다. 문장이 길어질수록 탐색 공간이 폭발적으로 증가하기 때문에 그만큼 탐색 오류가 발생할 가능성도 커질 것으로 판단했다.

절 단위 디코딩을 위해서 전처리기는 파싱 정보에 따라 절 경계를 판단한다. 한정용법의 관계대명사절은 번역문에서 명사구를 수식하기 때문에 별도로 디코딩하여 번역문이 모두 생성되면 해당 명사구의 앞에 연결하고, 부사절은 외국어 문자열의 위치에 따라 번역문에서 주절에 해당하는 구간의 앞/뒤에 연결시켰다. 목적어로 쓰인 명사절의 경우에는 번역문에서 안긴 문장에 속하기 때문에, 한국어 안긴 문장이 안은 문장의 중간에 삽입되도록 하는 등의 규칙을 사용했다.

○ 입력문

ReCaptcha was spun-off from Carnegie Mellon University, where Mr von Ahn is an assistant professor of computer science.

○구문 구조 변형, 개체명 인식, 절 단위 분리 결과

ReCaptcha <Carnegie Mellon University_institution> from spun-off was // there <Mr von Ahn_person> computer science of an assistant professor is.

○다중 디코딩 결과

ReCaptcha는 카네기 멜론 대학 에서 분리 되 았 다 //
거기 에서 미스터 본 Ahn 은 컴퓨터 과학 의 조교수 이 다 .

○후처리 결과

ReCaptcha 는 카네기 멜론 대학에서 분리되었는데, 거기에서 미스터 본 Ahn은 컴퓨터 과학의 조교수이다.

위의 "절 분리 규칙이 적용된 전/후 처리"의 예에서 알 수 있듯이 복문은 분리/디코딩/재결합 과정을 통해 번역을 하게 된다. 하지만 복문이 입력되더라도 분리-결합 규칙을 모두 적용할 수 있는 것은 아니다. 관계 대명

사절이 포함되어 있을 때, 선행사에 대한 개체명 인식이 실패하여 그 의미적 속성을 알 수 없을 때에는 문장 분리를 하지 않고 단순히 어순 재배치만 진행된다. 선행사가 언제나 물리적 장소만을 나타내는 것이 아니기 때문이다.

3.6 숙어 처리

숙어(idiom)는 개별 단어의 문자적인 의미와는 다르게 또 다른 비유적인 의미를 갖게 되는 단어/구이다. 현재 시스템에 'He kicked the bucket.'이라는 문장을 입력하면 '그는 양동이를 찼다.'라는 번역문이 생성된다. 이는 훈련 데이터 내에 'kick the bucket'에 대한 구 단위 번역문이 존재하지 않았거나, 존재하였더라도 그 수가 적어 스코어가 낮아지게 되므로 탐색 과정에서 폐기되기 때문이다.

숙어를 처리하는 능력을 높이려면, 전처리 단계에서 문장에 숙어에 해당되는 구간이 존재할 경우, 이 구간에 대한 탐색 노드를 번역 모델에서 가져오지 않고 숙어 데이터베이스에서 가져오도록 만들어야 한다. 이렇게 하면 이 하위 문자열에 대해서는 중의성이 존재하지 않게 되는 것이다. 현재 구축된 번역 시스템에서는 디코더에 입력문을 보낼 때, 다음과 같이 "forcePattern 객체가 처리하는 입력문의 부가 정보"가 함께 들어가도록 설계하였다.

He kicked the bucket. <PATTERN> (1-3)=[죽었다] </PATTERN>

상기 정보는 입력문을 구문 분석한 후, 전처리기가 VP 노드에서 동사구 형태의 숙어를 데이터베이스에서 검색했을 때 추가되는 패턴정보이다. 상기 패턴은 숫자나 요일과 같이 계열관계(paradigmatic relation[4])에 놓인 토큰 집합의 요소를 포함하지 않는 hard pattern 으로서 해당 구간(1-3)에 대해서는 중의성이 해소된 것으로 간주하여 디코딩 과정에서 단일의 탐색 노드 객체만 생성된다. 이 방식을 활용하면 탐색과정에서 발생하는 부분 번역 가설 객체의 수가 현저히 줄어드는 효과가 있다.

3.7 패턴 템플릿

Qualia 번역 디코더에 포함된 forcePattern은 섹션3.6에서 설명한 hard 패턴뿐만 아니라 soft 패턴도 입력할 수 있다. soft 패턴은 기본적으로 미등록어 처리되는 숫자 집합으로부터, 요일과 같이 수가 제한적인 요소를 포함하는 구를 의미한다.

Forecasts range from 5.1 per cent to 5.5 per cent annualized growth.

상기 "soft 패턴의 예" 문장에 포함된 숫자는 그 계열 요소가 무한대로 존재하기 때문에 번역모델이 모두 처리할 수 없다. 이러한 구간은 패턴으로 처리하는 것이 효율적이다. 위의 예에서는 "from 5.1 per cent to 5.5 per cent"가 전치사구 패턴으로 등록되어 있어, 부동소수(float-point representation)부분은 의사 토큰(FLOAT)으로 치환되고 나머지 요소들은 패턴 번역 데이터베이스에서 가져오도록 한다.

from FLOAT_1 per cent to FLOAT_2 per cent
=> FLOAT_1 퍼센트에서 FLOAT_2 퍼센트

상기 정보는 "soft 패턴 번역의 예"로서, 이렇게 소프트 패턴 템플릿에 적용이 가능한 구간은 디코더가 번역한 구간과 결합되어 번역문을 생성하게 된다.

3.8 번역 모델의 수동 교정

SMT에서 번역모델은 비교사 학습(unsupervised learning)에 의해 생성된다. EM 알고리즘[17,18]은 병렬코퍼스에 부재한 정렬 관계를 번역가의 도움 없이 대략적으로 복원하는 과정이므로, 아래와 같이 정렬 오류가 필연적으로 발생한다.

(accommodations ,) - (을)

구 추출 알고리즘에 의해서 학습한 번역 모델에는 언어전문가의 직관으로 보았을 때 명백하게 잘못 정렬된 번역지식들이 존재한다. 위의 경우에서처럼, 내용어 'accommodations,'가 기능어 '을'과 정렬된 경우 오류로 간주하였다. 물론 상기의 경우 번역 우도(likelihood)는 매우 낮지만 데이터의 특성에 따라 정렬되지 말아야 함에도 불구하고 높은 확률로 정렬되는 entry가 번역 모델 내부에 많이 존재한다.

이러한 정렬 오류는 런타임 번역 오류로 연결될 가능성이 있으므로 오프라인에서 번역지식을 갖춘 전문가로 하여금 해당 번역 모델 entry를 삭제하도록 할 필요가 있다.

3.9 도메인 특화 전략

기계번역 시스템의 성능 향상과 실제 사용을 위해서는 도메인의 제한이 필수적이다. 특정 도메인에서 사용되는 단어나 구는 의미적 중의성이 오픈 도메인보다 심하지 않고, 반복되는 표현을 쉽게 DB화할 수 있기 때문이다.

본 연구의 첫 번째 테스트 도메인은 Financial Times의 하위 도메인인 Equities Market이다. 주식 시장 관련 기사는 표현 패턴이 반복되는 경향이 많고, 구문 구조도 비교적 명확하여 파싱 오류를 줄일 수 있다고 판단했기 때문이다. 또한 잠재적인 사용자의 수요도 고려하였다.

도메인 특화 전략은 다음과 같다. 첫째, aggregator를 이용하여 수집한 코퍼스로부터 Financial Times의 개체

4) number = {0, 1, 2 … },
week = {Sunday-일요일, …, Saturday-토요일}

명을 추출, 고유명사 대역어 사전을 미리 만들어 둔다. 개체명은 "Bank of New York Mellon"과 같이 하나 이상의 토큰으로 구성되어 있어 번역 모델에 상기 개체명에 대한 정보가 부재할 경우 단어 단위 또는 더 좁은 범위의 구 단위 번역이 이루어지므로 좋은 번역을 기대하기 어렵다. 단어 정렬 후 훈련된 번역 모델에서는 이미 구축된 개체명 번역 사전과 교집합을 이루는 엔트리를 제거한다. 이것은 모델 자체의 에러를 사전에 차단하는 효과를 갖는다.

두 번째, NP 패턴 번역 모델은 구문 분석 정보에 의거, 구 단위 번역 모델이 처리할 수 있는 구의 최대 길이인 4를 초과하는 문법적 명사구 범위에 대하여 적용하는 구 번역 모델이다. 이 번역 지식을 훈련시키기 위해서 번역가로 하여금 Financial Times텍스트로부터, 4개 토큰 이상으로 이루어진 명사구를 추출, 명사구 단위 한국어 번역문을 만든 뒤 pos 태깅과 형태소 분석을 수행하였다. 이 병렬 데이터에서 surface factor는 일반적인 구 단위 번역 모델 학습을 위한 병렬 코퍼스에 추가하고, pos/morph tag factor만으로 이루어진 병렬 코퍼스에 대하여 NP 단위 패턴 번역 모델을 훈련시켰다. 전처리 과정에서 4개 토큰을 초과하는 명사구는 NP 패턴 번역 모델에서 부분 번역 이벤트 모델을 가져오도록 하고, 후처리 단계에서 번역 모델이 아닌 기존의 기계 번역용 사전에서 형태소 단위 대역어를 가져와 정렬 정보에 따라 결합시키는 방식으로 부분 번역문을 만들었다.

세 번째, 단어 정렬 시 일반 도메인 사전, 수동 구축된 경제 도메인 사전을 병렬 코퍼스에 추가한다. 병렬 코퍼스의 사이즈가 정규분포를 이룰 정도로 크다고 하더라도 자주 출현하지 않는 단어들의 정렬 정확도는 낮다. 특히 형용사나 부사 클래스 단어들은 정렬 정확도가 낮은 것으로 확인되었다. 따라서 이렇게 정렬 가능성이 떨어지는 단어를 병렬 코퍼스에 추가하여 번역 모델에 포함시키도록 한다.

네 번째, 전처리 과정을 통해 중/복문으로 된 원문(영어) 문장의 어순을 한국어 어순과 유사하게 만들고, 분리 규칙을 적용할 수 있을 경우, 단문으로 분리하여, 단어 정렬을 시도하였다. 이 방법은 정렬의 균일 분포를 가정하는 IBM 1번 정렬 모델로 인한 정렬 오류를 줄이기 위해 시도되었다. 전처리 과정은 디코딩 과정에서도 동일하게 적용되므로, 디코딩은 절 분리 규칙이 적용될 수 있을 경우 절 단위로 이루어지며 후처리 프로세스에 의해 접합, 한국어 복문으로 복원된다.

다섯 번째, 기능어 어순 재배치 규칙 적용으로 영어 문장을 한국어 어순과 유사하게 변형하였다. 한국어와 영어의 가장 큰 차이점은 기능어의 어순 차이이다. 특히, 명사절, 부사절 유도 접속사나 관계사 등의 어순 차이는 distortion 모델링으로 쉽게 포착되지 않기 때문에 전역 어순 재배치를 통해 한국어 어순과 유사하게 만들고, 4단어 이상으로 이루어진 전치사구의 핵어인 전치사를 구의 뒤로 이동시키는 등의 규칙을 적용하여 전처리를 하였다.

여섯 번째, 번역모델을 일반 도메인 번역 모델과 특수 도메인 번역 모델로 분리한다. 이 때, 도메인 전용 feature function의 가중치를 일반 분야의 그것보다 높게 정해야 한다. 주식 시장에 대한 뉴스를 전하는 번역 시스템을 만들 경우, 주식 시장 뉴스 코퍼스로부터 학습한 번역 모델에 높은 가중치를 주면 런타임 디코딩 과정에서 특수 도메인에 적합한 번역문을 포함하는 부분 번역 가설이 탐색 과정에서 탈락될 가능성이 낮아진다. 예를 들어, 'Stock market crashed.'라는 문장을 일반도메인 번역모델과 주식시장 도메인의 번역모델의 가중치가 동일한 번역 시스템에 넣을 경우 '주식시장이 충돌했습니다.'라는 번역문이 생성되었다. 그러나 주식시장 기사 코퍼스로 훈련시킨 번역모델의 가중치를 상대적으로 높였을 때에는, 'crashed'의 번역이 '급락했다'로 출력되었다.

4. 실험 및 평가

본 논문에서 제안한 번역 시스템의 번역 지식을 학습시키기 위해서 일반 도메인 병렬 코퍼스 600,000 쌍을 웹에서 획득하였다. 웹 데이터를 수집하는 crawler를 이용하여 웹 페이지를 파싱한 후 웹 페이지 별로 저장하고, hunalign[5]을 이용하여 문장 단위 정렬을 하였다. 그리고 번역가 2명을 활용하여 6개월에 걸쳐 Financial Times/Markets/Equities/Asia Pacific, U.S 섹션의 기사를 번역하여 15,500쌍의 도메인 병렬 텍스트를 구축하였다. 이 15,500쌍의 병렬 텍스트에서 무작위로 500개의 원문과 번역문을 테스트 셋으로 남겨두고 나머지 15,000쌍의 FT 도메인 코퍼스와 600,000쌍을 합쳐 학습 데이터로 활용하였다. 일반 도메인 병렬 코퍼스를 활용한 이유는 도메인 전용 코퍼스만 사용하게 될 경우 OOV (Out Of Vocabulary) 문제가 발생할 가능성이 있기 때문이다.

단어 단위 정렬의 경우 Hermes를 활용하지 않고 일반적으로 사용되는 giza++를 사용했다.

MERT 트레이닝은 별도로 수행하지 않고, 가중치를 수동으로 결정하는 방식을 택하였다. MERT 결과로 주어지는 모델 가중치를 적용하였을 때 번역문의 정성적 정확도가 눈에 띄게 좋지 않았기 때문이다.

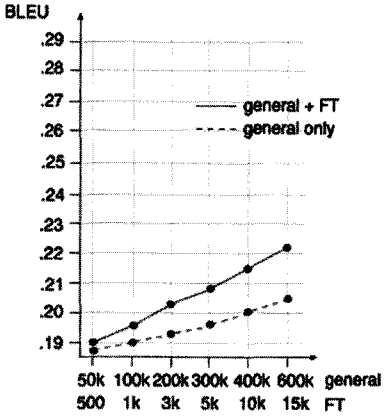
일반 도메인 구 단위 번역 모델에 가중치 0.2, 주식/

5) http://mokk.bme.hu/resources/hunalign

환율기사 코퍼스로 훈련한 번역모델은 0.3 언어모델에 0.3, NE 번역 모델에 0.1, NP 패턴 번역 모델에 0.1을 부여하였다. 번역 모델은 Source to Target 구 번역모델과 lexical weight 모델, Target to Source 구 번역모델과 lexical weight 모델, 그리고 phrase penalty로 나누어진다. 이 5개의 모델에 대한 가중치는 각각 0.7, 0.05, 0.1, 0.05, 0.1로 정하였다.

베이스 라인은 3.9섹션에서 소개한 특화 전략의 적용이나 전후 처리. 패턴 처리 등이 적용되지 않은 시스템으로 시작하였으며, 활용한 탐색 알고리즘은 beam search이고, distortion 최대값은 4로 정하였다. 평가 규준으로 BLEU[19]metric[6]을 활용하였다.

표 2 코퍼스 사이즈 변경에 따른 성능 변화 베이스라인 configuration



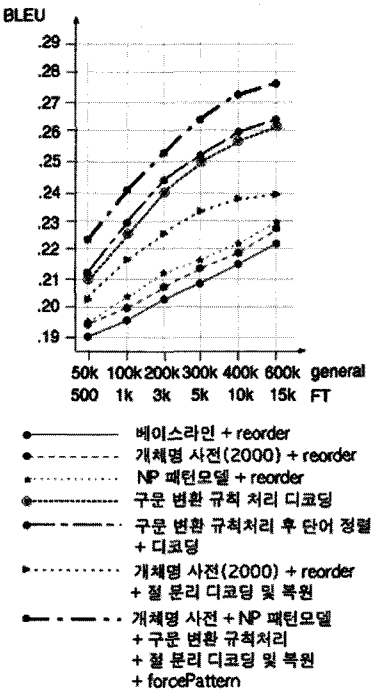
베이스라인 환경 설정으로 문장 단위의 코퍼스의 사이즈만을 증가시켰을 경우의 학습률(learning rate)은 상기와 같다. 테스트 데이터가 Financial Times 도메인에서 생성된 것이므로, 훈련 데이터에 해당 코퍼스가 포함된 시스템은 그렇지 않은 시스템과 비교하여 번역 정확도의 향상 비율이 높았다.

표 3은 여러 가지 성능 향상 기법들을 독립적으로 그리고 복합적으로 사용하면서 훈련 코퍼스의 사이즈를 증가시켰을 때의 번역 정확도 변화를 보여준다. 개체명 사전의 엔트리 사이즈는 2000으로 유지되었다. 별도의 표는 작성하지 않았으나 개체명 사전, forcePattern 기능 사용하면서 개체명 사전의 엔트리, 패턴 템플릿의 사이즈를 증가시키는 것도 번역 정확도 상승에 크게 기여하는 것으로 나타났다.

가장 영향력이 큰 정확도 개선 요인은 구문 변환 규칙을 통해 어순을 목적 언어와 유사하게 변형하는 프로

6) http://www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html

표 3 휴리스틱에 따른 번역 정확도 변화



세스로 관찰되었다. 구문 변환 규칙은 번역가의 도메인 지식을 규칙화한 것이기 때문에 로컬 및 전역 distortion을 적절하게 포착한 것으로 보인다.

실험을 통해 개체명 구간을 의사단어로 치환하는 등의 문장 단순화 및 사전(dictionary) 정보를 통해 정렬 정확도를 높이고, 구문 분석 결과를 토대로 문장 변형 규칙을 적용하여 언어간 구문 차이를 최소화함으로써 번역 정확도를 향상시키는 것이 가능하다는 것을 확인할 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 RBMT, SMT, PBMT를 활용한 직렬 연결 방식의 하이브리드 번역 시스템을 제안하였으며, 실험을 통해 번역 품질 향상이 가능하다는 것을 확인할 수 있었으며, 특히, 어순 배치의 경우 distortion 모델에 의존하지 않고 구문 변환(rule-based syntactic transfer) 규칙을 사용하는 것이 더욱 효과적인 것을 확인할 수 있었다.

하이브리드 번역 시스템은 각 번역 시스템 구축 모델의 장점들만을 통합하여 시너지 효과를 얻을 수 있을 것이라는 예상에서 출발했다. 실험 결과를 통해 확인할 수 있듯이, 최대 우도 추정(MLE)에 의한 distortion 이벤트 모델링만으로는 번역 프로세스에서 발생하는 복잡

한 어순 변경을 적절하게 포착하는 것이 어렵다. 왜냐하면 MLE는 인간이 이미 알고 있는 확실한 지식을 활용하지 않기 때문이다. 대신, RBMT가 사용하는 syntactic transfer 규칙사전을 활용하는 것이 전역 distortion 처리 측면에서 효율적이다. 그러나 syntactic transfer의 실패는 번역 정확도의 감소로 이어지기 때문에, 로컬 distortion 모델과 syntactic transfer를 적절하게 조합하는 방법에 대한 연구가 필요하다.

또한 특정 도메인에서 빈번하게 출현하는 패턴 템플릿 구간은 중의성이 존재하지 않는 구간으로 인정하고 단일의 번역 옵션만을 활용하여 탐색 공간을 대폭 줄일 수 있다는 것을 확인할 수 있었다.

본 논문에서 RBMT, PBMT, SMT에 의한 구축 모델을 직렬 연결 방식을 제안하였으나, 향후 연구는 독립된 번역 엔진들을 하나로 통합하여 개별 번역 시스템이 생성한 번역문의 최적 구간들을 수집, 하나의 번역문으로 합치는 과정을 구현하는 병렬 프로세스 기반 하이브리드 번역 시스템에 대한 연구가 필요하다.

참 고 문 헌

[1] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin: *A Statistical Approach to Machine Translation.* Computational Linguistics, vol.16, no.2, pp.79-85, June 1990.

[2] A.L. Berger, S.A. Della Pietra, V.J. Della Pietra: *A Maximum Entropy Approach to Natural Language Processing.* Computational Linguistics, vol.22, no.1, pp.39-72, March 1996.

[3] F.J. Och. Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of ACL,* pp.160-167, Morristown, NJ, USA, 2003.

[4] Gary Geunbae Lee, Jonghoon Lee, Donghyeon Lee, A transformation-based sentence splitting method for statistical machine translation. *Proceedings of the IJCNLP2008 workshop on technologies and corpora for Asia-pacific speech translation,* Hyderadad, Jan 2008.

[5] MichaelCollins, PhilippKoehn, and Ivona Kuc˘erova Clause Restructuring for Statistical Machine Translation. *In Proceedings of ACL,* pp.531-540, 2005.

[6] Fei Xia and Michael McCord, Improving a Statistical MT System with Automatically Learned Rewrite Patterns. *In Proceedings of the 20th international Conference on Computational Linguistics,* 2004.

[7] Koehn Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bo jar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. *In Proceedings of ACL Demo Session,* pp.177-180, Prague, Czech Republic. 2007.

[8] D. Jurafsky, J.H. Martin: *Speech and Language Processing.* Prentice Hall, Englewood Cliffs, NJ, pp.890-893, 2000.

[9] Philipp Koehn, *Statistical Machine Translation,* pp. 252-256, pp. 256, Cambridge University Press, 2010.

[10] R. Kneser, H. Ney, *Forming Word Classes by Statistical Clustering for Statistical Language Modeling.* In 1. Quantitative Linguistics Conf., pp. 221-226, Trier, Germany, Sept. 1991.

[11] I.D. Melamed: *Models of Translational Equivalence among Words.*, Computational Linguistics, vol.26, no.2, pp.221-249, 2000.

[12] S. Vogel, H. Ney, C. Tillmann: *HMM-based Word Alignment in Statistical Translation.* In COLING '96: The 16th Int. Conf. on Computational Linguistics, pp.836-841, Copenhagen, Denmark, Aug. 1996.

[13] K. Knight. *A Statistical Machine Translation Tutorial Workbook,* 35 pages, Aug. 1999. http://www.isi.edu/natural-language/mt/wkbk.rtf.

[14] Philipp Koehn, Franz Josef Och, and Daniel Marcu, *Statistical Phrase-Based Translation,* HLT/NAACL 2003.

[15] Ye-Yi Wang, Alex Waibel, *Decoding Algorithm in Statistical Machine Translation,* Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pp.366-372, Madrid, Spain, July 07-12, 1997.

[16] Thomas H. Cormen, Introduction to Algorithm, pp. 628-635, Ed. 2, The MIT Press, 2001.

[17] A.P. Dempster, N.M. Laird, D.B. Rubin: *Maximum Likelihood from Incomplete Data via the EM Algorithm.* J. Royal Statist. Soc. Ser. B, vol.39, no.1, pp.1-22, 1977.

[18] Sergios Theodoridis, *Pattern Recognition,* pp.44-49, Academic Press; Ed. 4, 2008.

[19] K.A. Papineni, S. Roukos, T. Ward, W.J. Zhu. Bleu: *A Method for Automatic Evaluation of Machine Translation.* Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, 10 pages, Sept. 2001.

강 운 구

2001년 2월 인하대학교 전자계산공학과 (공학박사). 2002년~2006년 뉴미디어연구소장. 2007년~현재 u-헬스케어연구소장. 1994년~현재 가천의과학대학교 정보공학부 교수. 관심분야는 소프트웨어공학, u-헬스케어, 의료정보

김 성 현

2003년 2월 연세대학교(원주) 문리대학 영어영문학 학사. 2005년 2월 연세대학교 일반대학원 문학 석사(논문명: 이상한 나라의 엘리스: 환상적 서사의 정치적 무의식 연구). 2001년~2006년 프리랜서 번역가. 2007년~현재 (주)엘엔아이소프트 자연언어처리연구팀 선임 연구원. 2007년 은닉마르코프 모델 기반 영어 POS 태거 개발. 2008년 최대엔트로피 기반 영어 고유명사 인식기 개발. 2008년~2010년 통계기반 번역 엔진 및 단어 정렬 모델 훈련기 개발. 관심분야는 패턴인식, 기계학습, 자동번역

이 병 문

1990년 2월 서강대학교 전자계산학과(공학석사). 2007년 2월 인천대학교 컴퓨터공학과(공학박사). 1990년~1997년 LG정보통신 중앙연구소선임연구원. 1998년~현재 가천의과학대학교 정보공학부 부교수. 2005년~현재 가천의과학대학교 u-헬스케어연구소 연구원. 관심분야는 u-헬스케어, 센서운영체제, 센서네트워크

이 영 호

1996년 2월 한국 외국어 대학교 응용전산학과(이학석사). 2005년 8월 아주대학교 의료정보학과(이학박사). 1999년~2002년 IBM Korea BI & CRM EM. 2002년~현재 가천의과학대학교 정보공학부 부교수. 2007년~현재 ISO/TC215전문위원. 2005년~현재 가천의과학대학교 u-헬스케어연구소 연구원. 2008년~현재 수송물류분야 단체표준 전문위원. 관심분야는 데이터마이닝, 의료정보, u-헬스케어