일영 통계기계번역에서 의존문법 문장 구조와 품사 정보를 사용한 클러스터링 기법

(A Clustering Method using Dependency Structure and Part-Of-Speech(POS) for Japanese-English Statistical Machine Translation)

김 한 경 † 나 휘 동 †
(Hankyong Kim) (Hwi-Dong Na)
이 금 희 † 이 종 혁 ††
(Jin-Ji Lee) (Jong-Hyeok Lee)

요 약 클러스터링 기법은 다양한 분야에서 이용되어 왔으며, 통계 기반 기계번역에서도 익히 사용된 기법이다. 그러나 기존의 연구에서는 깊이 있는 문법적인 분석 없이 기계학습 기법을 사용하거나, 문장구조의 정보를 사용하더라도 정규식을 이용하여 판별하는 선에서 그치는 경우가 많았다. 본 논문에서는 각 문장의 의존관계 문법에 따른 구조와 조사 등의 품사 정보를 사용하여 문장구조를 파악하고 유형별로 분류하여 각각에 특화된 언어모델을 획득하는 방법과, 이를 구 기반 통계기계번역에 추가적인 정보로 사용하여 번역성능을 향상하는 데 이용하는 방법을 제안한다.

· 본 논문은 2009년도 두뇌한국21사업, 지식경제부 및 정보통신진흥연구원의 정보통신선도기반기술개발사업, 한국과학재단 기초연구사업(No.2009-0075211), 그리고 MSRA(2009)의 지원으로 수행되었습니다.
· 이 논문은 2009 한국컴퓨터종합학술대회에서 '일영 통계기계번역에서 의존문법 문장 구조와 품사 정보를 사용한 클러스터링 기법'의 제목으로 발표된 논문을 확장한 것임.

† 비 회 원 : 포항공과대학교 컴퓨터공학과
arch@postech.ac.kr
leona@postech.ac.kr
ljj@postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
jhlee@postech.ac.kr

논문접수 : 2009년 8월 14일
심사완료 : 2009년 10월 8일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제12호(2009.12)

키워드 : 통계기계번역, 클러스터링, 문장구조, 의존문법

Abstract Clustering is well known method and that can be used in statistical machine translation. In this paper we propose a corpus clustering method using syntactic structure and POS information of dependency grammar. And using this cluster language model as additional feature to phrased-based statistical machine translation system to improve translation quality.

Key words : SMT(Statistical Machine Translation), Clustering, Syntactic Structure, Dependency Grammar

1. 서 론

구 기반 통계기계번역(phrase-based SMT)에서 원시언어의 문장을 목표언어로 번역하는 과정은 번역모델(translation model)과 언어모델(language model)을 사용한다. 번역모델은 원시언어의 구(phrase)가 목표 언어의 어떤 구로 번역될 확률을 나타내며, 언어모델은 목표언어의 단어들이 연속하여 출현할 확률을 의미한다. 이 두 가지 모델을 사용하여 원시언어에서 목표언어로 의미의 전달과 자연스럽고 유창한 목표언어 문장의 생성을 하게 된다.

이때 사용하는 각 확률모델을 획득하는 과정에서 클러스터링 기법을 사용하면 세부 도메인에 특화된 모델을 얻을 수 있다. 이러한 모델을 추가로 사용하면 번역성능을 향상할 수 있고, 이에 관련된 연구는 이전부터 다양하게 진행되어 왔다. 기존의 방법들은 번역모델을 다룬 연구와 언어모델을 대상으로 한 연구, 그리고 두 모델 모두를 대상으로 한 연구가 있으나 클러스터링 방법을 기준으로 나눈다면 기계학습 기법을 사용한 연구와 언어학적 지식을 사용하여 문장의 유형에 따라 분류한 연구로 나눌 수 있다.

기계학습 기법을 사용한 연구로 Yamamoto등[1]은 문장에 대하여 계산한 각 유형별 언어모델의 엔트로피 값을 거리함수로 사용하여 말뭉치를 분류하고, 유형별 언어모델과 번역모델을 추가로 사용하였다. Ito등[2]은 단어 유사도를 사용하는 클러스터링 도구로 말뭉치를 분류하고 세부 도메인에서 획득한 번역모델을 추가로 사용하여 실험하였다.

한편 Hasan등[3]은 정규식을 사용하여 말뭉치를 문장 유형별로 분류하고, 해당 유형의 언어모델을 추가로 사용하였다. Yasuda등[4]은 일-영 특허번역에서 대상 말뭉치를 다루고 있는 분야에 따라 IPC(International Patent Classification)에 의거하여 분류하여 번역모델과 언어모델에 적용하였다. Li등[5]은 문장에 포함된 단문의 종류에 따라 전체 말뭉치를 분류하고 획득한 유형별

언어모델을 추가로 사용한 결과를 발표하였다.

기존연구 중 기계학습기법을 사용한 실험에서는 주로 단어의 출현빈도를 주요자질로 사용하여 분류하였고, 번역모델에 적용하였을 때 우수한 결과를 얻었다. 그리고 언어학적 지식을 사용하여 말뭉치를 분류한 실험 사례들에서는 문장구조의 유사성이 언어모델의 성능을 향상하는데 도움이 됨을 발견하였다.

이 논문에서는 NTCIR-7 특허번역작업[6] 말뭉치에서 일본어를 영어로 번역하는 방향으로 실험하였다. 이 말뭉치는 특허번역 문서로 각 문장의 길이가 긴 편이며, 단문보다는 중문이나 복문이 많이 나타나는 특성을 가지고 있다. 이것은 먼 거리 의존관계의 번역에 곤란을 겪는 구 기반 통계기계번역의 단점으로, 상대적으로 낮은 번역성능을 보이는 원인이 된다.

이를 보완하기 위해서 본 논문에서는 다음과 같은 일련의 가설을 세우고 실험하였다. 먼저 복문, 중문으로 이루어진 일본어 문장에서 각각의 절(clause)은 영어 문장에서도 독립된 절로 대역될 것이라 가정하였다. 다음으로 각 언어에서 문장에서 내포된 절의 수와 절 사이의 의존관계가 유사한 문장은 전체 문장의 구조 또한 같다고 볼 수 있다. 이렇게 문장구조가 유사한 말뭉치를 형성하면 언어모델의 성능을 향상할 수 있을 것이다. 따라서 일본어 문장에서 문장구조를 분석하여 유사한 구조를 가진 문장을 모아 추가의 언어모델을 획득하고, 이 새로운 언어모델을 통계기계번역도구에 추가정보로 사용할 경우 번역성능의 향상을 얻을 수 있으리라 가정하고 실험하였다.

실제 실험에서는 병렬말뭉치에서 일본어 문장을 의존문법 파서인 cabocha[7]를 사용하여 얻은 결과를 분석하여 분류하였고, 나누어진 세부 도메인에서 영어의 언어모델을 획득하여 통계기계번역도구인 Moses[8]에 추가의 정보로 사용하였다.

2. 본 론

2.1 의존관계 정보를 사용한 클러스터링 방법

본 논문에서 사용한 NTCIR-7 특허번역작업 말뭉치의 경우 한 문장에 내포하는 절이 여러 개인 경우가 많다. 이렇게 한 문장에 여러 개의 절이 속해 있을 때, 내포된 절의 수와 각 절의 관계를 자질로 사용하여 문장을 분류하면 유사한 형식의 문장들로 이루어진 말뭉치를 얻을 수 있다. 이 실험에서는 일본어에서 영어로 번역을 수행하므로 전체 말뭉치의 일본어 문장을 의존관계 문법에 따라 분석하고, 그 의존관계 트리를 분석하였다.

일본어 문장의 의존관계 트리는 우리말의 어절과 유사한 단위인 문절(文節)을 하나의 단위로 하여 트리를 형성한다. 이 의존관계 트리에서 문장의 본용언이 포함된 문절이 루트노드가 되며, 그 외의 문절은 루트 노드에 직접적인 의존관계를 가지는 노드와 그 아래의 자식 트리를 형성하는 노드가 된다. 이러한 일본어 문장의 의존관계 트리를 분석한 결과 일본어 문장에 내포된 절의 수는 루트에 직접적인 의존관계를 가지는 노드의 수와 관계가 있다고 판단하였다. 물론 의존관계 트리의 형태만으로 내포된 절의 수를 알 수 있다고 보기에는 곤란하기에, 이 실험에서는 루트에 직접적인 의존관계를 가지는 노드와 그 자식트리가 하나의 절을 형성할 조건으로 일정한 조사나 문장 부호를 요구한다.

정리하자면 언어학적 분석결과 일본어 문장의 의존관계 트리에서 루트에 직접 연결된 자식노드를 형성하는 문절에 병립조사(並立助詞), 계조사(係助詞), 접속조사(接續助詞), 부사화조사(副詞化), 인용격조사(格助詞-引用), 쉼표(讀点)의 품사를 가지는 형태소가 있을 경우 이 문절과 그 자식트리에 해당하는 문절들은 독립된 절을 형성한다고 간주하였다.

이를 기본 구상으로 하여 일본어 문장의 의존관계 분석결과에서 일정한 규칙에 따라 패턴을 추출하여 말뭉치를 분류하였으며 그 방법은 다음과 같다.

우선 일본어 문장의 의존관계 트리에서 루트와 직접적인 의존관계를 가지는 문절만을 패턴 추출의 대상으로 한다. 이러한 조건을 만족하는 각 문절의 의존문법 파싱 결과에서 접속조사, 계조사, 인용조사, 부사화조사, 병립조사, 쉼표를 품사로 가지는 형태소가 나타날 경우 해당 품사를 패턴으로 추출하고, 서로 다른 문절에서 추출된 패턴 사이에는 '/'를 구분자로 삽입하였다. 그리고 어떠한 대상 문절에서도 해당하는 품사를 가지는 형태소가 없을 경우의 패턴은 '-'와 같이 표현하였다.

이러한 방법으로 일본어 문장에서 패턴을 추출한 예는 아래와 같다.

표 1의 일본어 문장을 의존문법으로 분석한 결과는 그림 1에 나타난 것과 같다. 이를 확인하면 두 문절이 루트와 직접적인 의존관계를 가지며, 그 중 한 문절에서 계조사와 쉼표를 발견할 수 있다. 다른 문절에서는 이 실험에서 절 사이의 관계를 나타내는 것으로 간주하는 품사를 가진 형태소가 없으므로 이 문장에서 추출되는

표 1 일본어 문장과 대역되는 영어 문장

전체 문장	일본어	この変形例でも、図12のキャップ4を使用している。
	영어	In this modified example also , there is used the cap 4 shown in figs . 12a and 12b .
절 1	일본어	この変形例でも、
	영어	in this modified example also ,
절 2	일본어	図12のキャップ4を使用している。
	영어	there is used the cap 4 shown in figs . 12a and 12b .

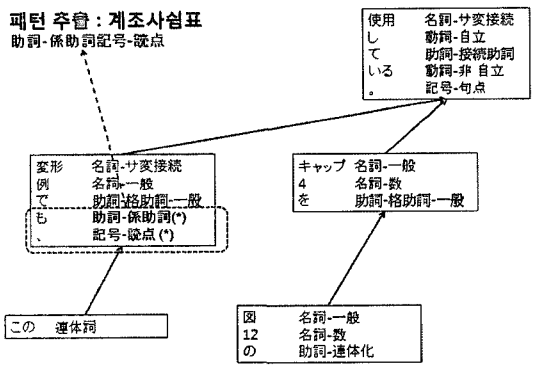

그림 1 의존관계 트리와 패턴추출 예 1

패턴은 '계조사쉼표'의 형태가 되며, 한 문절에서만 패턴이 추출되므로 구분자가 사용되지 않는다. 여러 문절에서 패턴이 추출되는 경우에는 표 2와 그림 2에서 나타난 바와 같이 각 문절에서 추출된 패턴의 사이에 구분자를 삽입하여 혼란의 여지를 줄이는 것과 함께 문장에 내포되는 절의 수를 표시할 수 있도록 하였다.

상기한 규칙을 기반으로 학습할 병렬말뭉치 전체 1,172,709문장 중 동일한 패턴의 출현빈도가 10,000번 이상인 패턴을 이 실험에서 클러스터로 선택하여 클러스터링을 수행하였다.

2.2 실험 환경

기본실험은 상기한 바와 같이 NTCIR-7 특허번역작업 말뭉치를 사용하였으며 훈련말뭉치에서 다음과 같은 전처리를 수행한 후 사용하였다. 우선 훈련말뭉치의 영어 및 일본어의 각 문장에 대하여 40단어 이상의 긴 문장을 제거하였다. 다음으로 일본어의 전각문자는 반각문자로, 영어의 대문자는 소문자로 치환하였으며 문장부호를 분리하였다. 이러한 전처리를 수행한 전체 훈련말뭉치에서 언어모델과 번역모델을 획득하여 통계기계번역도구 Moses에 사용하였으며, 이때 각 모델의 가중치는 학습말뭉치와는 별도로 튜닝말뭉치를 준비하여 획득하였다. 기본실험은 이러한 시스템에서 테스트말뭉치를 번역한 결과이며 이 번역결과를 BLEU[9]를 사용하여 채점하였다.

클러스터링을 사용한 실험에서는 기본실험에서 구축한 통계기계번역 시스템에 각각의 유형별 세부 도메인으로 나누어진 말뭉치에서 획득한 언어모델을 추가로 사용하여 번역하였다. 이때 전체 말뭉치에서 획득한 언

표 2 일본어 문장과 대역되는 영어 문장

전체문장	일본어	このとき、図2に示されるシェアドセンスアンプ回路1200bは活性化されない。
	영어	In this operation, shared sense amplifier circuit 1200b shown in fig. 2 is not activated.
절 1	일본어	このとき、
	영어	in this operation ,
절 2	일본어	図2に示されるシェアドセンスアンプ回路1200bは活性化されない。
	영어	shared sense amplifier circuit 1200b shown in fig . 2 is not activated .

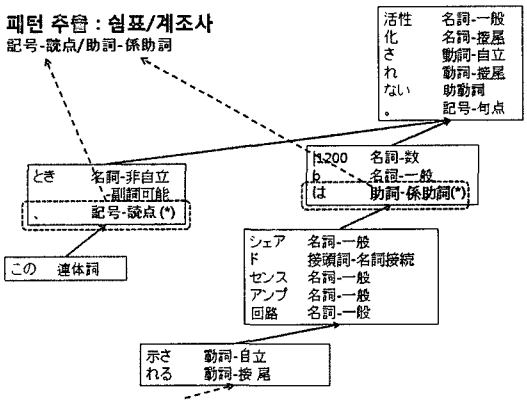

그림 2 의존관계 트리와 패턴추출 예 2

표 3 말뭉치 정보

말뭉치 종류	문장 수
학습	1,172,709
튜닝	609
테스트	1,381

표 4 말뭉치 클러스터링 결과

유형	패턴	말뭉치별 출현 빈도	
		학습	테스트
1	계조사쉼표	183,297	200
2	쉼표	142,521	136
3	계조사	113,855	65
4	- (추출 패턴 없음)	88,046	77
5	쉼표/계조사	39,454	53
6	쉼표/쉼표	38,563	58
7	쉼표/계조사쉼표	37,472	56
8	계조사쉼표/쉼표	33,502	42
9	접속조사쉼표	25,657	25
10	쉼표/접속조사	22,320	24
11	쉼표계조사	21,971	27
12	쉼표계조사쉼표	20,121	29
13	계조사쉼표/접속조사	19,191	26
14	접속조사	16,862	6
15	쉼표/접속조사쉼표	11,900	22
16	인용조사	11,203	10
17	계조사/접속조사	10,930	7
18	계조사/쉼표	10,257	9
합계(%)		847,122 (72.23%)	872 (63.14%)

어모델과 추가로 사용하는 클러스터별 언어모델의 가중치를 조절하여 Moses에서 사용한 두 언어모델 가중치의 합은 기본실험에서 전체 언어모델에 주어진 가중치와 동일하도록 유지 하였다. 그리고 기본실험에서 사용하였던 전체 언어모델과 추가로 획득한 클러스터별 언어모델의 비율은 각 유형별로 200문장씩 별도의 튜닝을 위한 말뭉치를 준비하여 모든 클러스터가 최적의 성능을 발휘할 수 있도록 서로 다른 비율을 가질 수 있게 튜닝하였다.

즉 기본실험에서 사용한 언어모델을 LM_g라 하고, 각 유형별로 나뉜 말뭉치에서 획득한 언어모델을 LM_c(c = {1, 2, …… 18})라 할 때 클러스터 c에 속하는 일본어 문장을 번역하기 위해 사용하는 언어 모델 'LM_c는

$$'LM_c = (1 - \lambda_c)\ LM_g + \lambda_c\ LM_c \quad (1)$$

와 같이 표현할 수 있고 이 'LM_c를 구하는 과정에서 λ_c는 각 클러스터 c마다 최적의 값을 가지도록 별도의 말뭉치를 사용하여 학습하였다.

이 λ_c의 학습 방법은 모든 클러스터에 대하여 λ_c를 0.0부터 1.0까지 0.1단위로 변화시키면서 해당 클러스터의 λ_c를 학습하기 위해 준비된 200문장씩의 말뭉치를 번역하여 가장 좋은 번역성능을 내는 값을 선택한다.

따라서 일본어 문서를 번역할 때, 번역할 문장이 어떤 클러스터 c에 속할 경우에는 상기의 과정을 통하여 준비된 언어모델 'LM_c를 사용하여 번역하였으며, 어떤 클러스터에도 소속하지 않을 경우에는 LM_g를 사용하여 번역하였다.

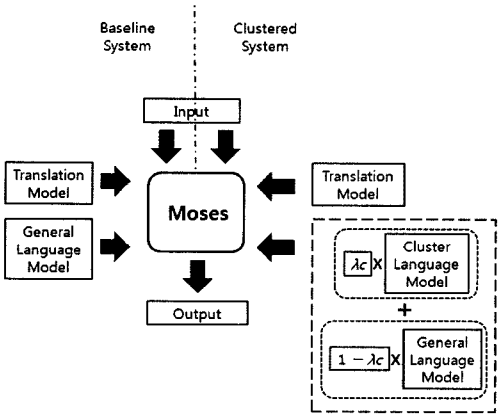

그림 3 번역 시스템 구성

2.3 실험 결과

일본어의 문장구조를 기준으로 수행한 이 실험에서 과연 클러스터링이 잘 이루어졌는지 확인하기 위해서 각 유형별로 나누어진 말뭉치의 혼란도(perplexity)를 SRILM[10]으로 측정하여 보았다. 이 혼란도의 개념은 언어모델과 해당언어의 문장이 주어질 때 문장이 언어모델에 적합한 단어의 배열순서를 따르고 있는가를 계산한 값으로써 낮을수록 더 유창한 문장이라 할 수 있다. 본 논문의 주장에 따르면, 병렬말뭉치에서 일본어 문장의 형식을 기준으로 말뭉치를 분류하였으나 영어 문장도 유사한 구조를 가지는 문장들이 모이게 된다. 즉 유형별로 나누어진 훈련말뭉치의 영어 언어모델과 해당 말뭉치의 복잡도를 계산하고, 이 결과를 다른 유형의 말뭉치에서 획득한 영어 언어모델로 계산한 복잡도와 비교하면, 일본어 문장의 의존관계와 품사를 기준으로 수행한 이 실험이 영어에서도 올바른 클러스터링을 수행하였는지 확인할 수 있다. 이를 계산한 결과 일부 예외를 제외하면 학습 말뭉치의 각 유형별 말뭉치에서 해당 클러스터의 언어모델을 사용하여 계산한 혼란도가 다른 클러스터의 언어모델을 적용한 경우들의 평균에 비하여 감소한 것을 확인하였다.

또한 유형별 언어모델을 추가로 사용였을 때의 번역 성능 향상치는 혼란도 감소치를 계산했을 때와 같이 각 클러스터에 해당하는 말뭉치를 200문장씩 별도로 준비하여 이것을 클러스터링을 사용한 시스템과 기본실험으로 준비된 시스템에서 각각 번역하여 그 결과의 차를 번역성능의 향상치로 정리한 것이다.

실제 테스트 말뭉치를 번역한 결과를 역시 BLEU를 사용하여 그 성능을 확인하면 클러스터링을 사용한 실험이 테스트 말뭉치 전체를 대상으로 하였을 경우 0.2%, 클러스터에 속하는 문장만을 번역하는 경우 0.4%의 향상이 있었다. 그리고, 번역된 결과의 각 문장에 대하여 문장별 BLEU를 사용하여 결과를 확인하였더니

표 5 유형별 언어모델 적용 결과

유형	혼란도 감소	번역성능 향상
1	13	1.4694095
2	26	2.46069144
3	20	0.19342083
4	17	2.0686309
5	-1	0.93707835
6	18	0.86346791
7	6	0.69908884
8	24	0
9	27	0.44860996
10	14	0.80450379
11	12	1.01025412
12	17	1.02690598
13	44	0.80370017
14	-1	0.58310531
15	21	1.03668469
16	48	1.603904797
17	85	0.535861153
18	19	0.720646638

표 6 테스트 말뭉치의 번역성능

	전체	클러스터 국한
기본실험 시스템	25.66(%)	25.33(%)
클러스터 시스템	25.86(%)	25.72(%)
최적 실험 시스템	26.16(%)	26.41(%)

기본실험과 결과가 동일한 유형 8번을 제외한 830개 문장 중 265개 문장에서 최대 0.62%에 이르는 성능향상이 있었고, 231개 문장에서는 최고 0.39%까지 기본실험에 비하여 떨어지는 결과를 보였다. 그리고 말뭉치에서 학습된 λ값이 아닌 실제 테스트 말뭉치에 최적화된 λ값을 사용했을 때의 번역성능, 즉 최적 실험 시스템에서는 최고 1.1%의 성능 향상이 있었다.

3. 결 론

본 논문에서 실험을 통하여 증명하고자 한 바는 두 가지로 정리할 수 있다. 일영 특허번역 말뭉치에서 일본어 문장의 의존관계와 품사를 통한 클러스터링이 영어에서도 유사한 문장구조의 말뭉치가 형성됨으로써 언어모델의 성능향상이 가능한 점과, 이 언어모델을 추가의 자질로 사용하여 구 기반 통계기계번역 시스템의 성능향상을 이끌어 낼 수 있다는 점이다.

단, 실제 번역성능의 향상 측면에서 클러스터 전체 및 유형별 적용률의 향상, 최적의 λ값을 학습하는 방법 등 여러 개선점을 발견하였다는 점에서 추가적인 언어학적 분석과 기계학습 기법에 대한 연구가 필요하다고 생각된다.

참 고 문 헌

[1] Hirofumi Yamamoto and Eiichiro Sumita : "Bilingual cluster based models for statistical machine translation," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.514-523, June 28-30, 2007.

[2] Takeshi Ito, Tomoyosi Akiba and Katunobu Itou : "Effect of the Topic Dependent Translation Models for Patent Translation - Experiment at NTCIR-7," *Proceedings of NTCIR-7 Workshop Meeting*, pp.425-429, December 16-19, 2008.

[3] Saˇsa Hasan and Hermann Ney : "Clustered Language Models based on Regular Expressions for SMT" 10th EAMT conference "Practical applications of machine translation," pp.119-125, 30-31 May 2005.

[4] Keiji Yasuda, Andrew Finch and Hideo Okuma : "System Description of NiCT-ATR SMT for NTCIR-7," *Proceedings of NTCIR-7 Workshop Meeting*, pp.415-419, December 16-19, 2008.

[5] Jin-ji Li, Hwi-dong Na, Hankyong Kim, Chang-Hu Jin and Jong-Hyeok Lee : "The POSTECH Statistical Machine Translation Systems for NTCIR-7 Patent Translation Task," *Proceedings of NTCIR-7 Workshop Meeting*, pp.445-449, December 16-19, 2008.

[6] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro : "Overview of the Patent Translation Task at the ntcir-7 Workshop," *Proceedings of NTCIR-7 Workshop Meeting*, pp. 389-400, December 16-19, 2008.

[7] Taku Kudo and Yuji Matsumoto : "Fast Methods for Kernel-based Text Analysis," *Proceedings of ACL-2003*, pp.24-31, 7-12 July 2003 Available at http://www.chasen.org/~taku/software/cabocha/

[8] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin and Evan Herbst : "Moses: Open Source Toolkit for Statistical Machine Translation," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2007.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. : "BLEU: A method for automatic evaluation of Machine Translation," *Technical Report RC22176*, IBM. 2001.

[10] Andreas Stolcke. : "SRILM - an extensible language modeling toolkit," In *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP)*. pp.693-696, 2002.