품사간 정렬 경향을 반영한 통계 기반 영한 단어 정렬 후처리 방법

이재희○, 이승욱, 황영숙, 김상범, 임해창
고려대학교 정보통신대학 컴퓨터전파통신공학과
SK 텔레콤 미래기술원
{jlee, swlee, rim}@nlp.korea.ac.kr {yshwang, sangbum.kim}@sktelecom.com

A Postprocessing method for Statistical English-Korean Word Alignment Reflecting Alignment Tendency Between Parts-of-Speeches

Jae-Hee Lee○, Seung-Wook Lee, Young-Sook Hwang, Sang-Bum Kim, Hae-Chang Rim
Dept. of Computer and Radio Communications Engineering, Korea University, Seoul, Korea
Institute of Future Technology, SK Telecom

요 약

병렬 말뭉치 내에서 서로 대응되는 단어를 찾아내는 단어 정렬 작업은 기계 번역에서 가장 기본적으로 수행되는 작업이고 다양한 분야에서 유용하게 사용된다. 본 논문에서는 영한 단어 정렬에서 기존의 통계 기반 정렬 모델의 문제점을 파악하고 이를 해결하기 위해 영한의 품사간 정렬 경향을 단어 정렬에 반영하는 방법을 제안한다. 실험을 통해서 기존 통계 기반 영한 단어 정렬 결과와 비교하여 제안된 방법이 정확률, 재현율, F-measure 측면에서 모두 향상시키는 것을 보였다.

주제어: 통계 기반 단어 정렬, 품사간 정렬 경향

1. 서 론

단어 정렬이란 동일한 의미의 문장을 두 개 이상의 언어로 표현한 병렬 말뭉치 내에서 문장에서 사용되는 단어들 간의 대응 관계를 찾아내는 작업이다. 단어 정렬 결과는 규칙 기반 기계번역(Rule-based Machine Translation)에서의 번역 규칙 추출이나, 예제 기반 기계번역(Example based Machine Translation)에서 번역틀을 생성하는 데 활용가능 하다. 뿐만 아니라 대역어 사전 구축, 복합어 인식, 의미 중의성 해소, 구문 분석 등 자연어 처리의 여러 분야에서 유용하게 사용 가능하다.

본 논문에서는 영한 단어 정렬에 있어서의 난점을 파악하고 이를 개선할 수 있는 방법을 제안하고자 한다. 영어와 한국어의 상이한 어순의 차이로 인해 정렬 대상이 되는 영어, 한국어 단어들의 위치의 격심한 차이가 발생할 수 있다. 의미를 나타내는 단위가 영어에서는 단어(word)인데 반해 한국어에서는 형태소(morpheme)이기 때문에, 직접적으로 단어 정렬을 시도하기 보다는 형태소 분석과 같은 부가적인 처리가 요구된다는 점 또한 영한 정렬을 힘들게 한다.

한국어 형태소 중 기능형태소(조사, 어미 등)들은 그 정렬 대상이 되는 영어 단어들을 파악하기 모호하다는 문제점이 있다. 한국어의 어절은 최소 하나 이상의 형태소가 합쳐져 구성되는데, 주격을 나타내는 '는', '가'나 목적격을 나타내는 '을', '를'과 같은 형태소들이 이에 해당된다.

또한, 문장을 구성하는 단어들의 의미적, 통사적 중의성 문제가 존재한다. 예를 들어, 한국어에서 '나는'이라는 어절에서 이 뜻은 대명사(나는 학교에 갔다)의 의미와 동사(하늘을 나는 비행기)의 의미로 쓰일 수 있으며, 비슷하게 영어에서 'fly'라는 단어가 명사(파리)의 의미와 동사(날다)의 의미로 사용될 수 있다.

이러한 특성들은 한영 단어 정렬의 성능을 저하시키는 요인으로 작용한다. 기존에는 두 언어의 구조적 특성을 일치시키기 위한 전처리 방법에 치중하였다 하지만 여전히 오류가 존재하고, 이들을 최소화하기 위해서는 자동으로 오류의 패턴을 탐지하여 수정하는 후처리 기법이 요구된다.

본 논문에서는 기존의 단어 정렬 결과로부터 품사간 정렬 경향성을 파악한 후, 이를 후처리 기법을 통해 반영하여 단어 정렬 결과를 개선하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 단어 정렬에 관한 기존 연구를 간략하게 소개하고, 3장에서는 본 논문이 제시한 방법에 대해 설명한다. 4장에서는 실험 및 결과를 분석하고, 마지막으로 5장에서 본 논문의 결론과 향후 계획에 대해 알아본다.

2. 관련 연구

단어 정렬에 관련된 다양한 연구가 존재하며 이들은 크게 사전 기반 방법과 통계 기반 방법으로 구분된다. 사전 기반 방법[1]은 특정 언어의 대역사전, 유의어 시소러스를 기반으로 정렬을 수행한다. 대용량의 병렬말뭉

표 1. 한국어 동사와 정렬된 영어 품사들의 빈도

영어 품사	빈도
일반 형용사	34,699
단수 일반명사	104,748
복수 일반명사	45,994
단수 고유명사	58,507
복수 고유명사	5,104
기본형 동사	34,682
과거형 동사	26,948
현재진행형 동사	17,361
과거분사형 동사	25,230
1인칭 현재형 동사	9,691
3인칭 현재형 동사	8,838

치를 필요로 하지 않는다는 장점이 있지만, 대상 언어에 의존적이고, 사전에 존재하지 않는 단어를 정렬하기 힘들다는 단점이 있다. 반면 통계 기반 방법[2]은 대용량의 병렬말뭉치를 이용한 학습과정을 거쳐 단어 정렬을 수행한다. 이 방법론은 언어적 정보를 정렬에 반영하지 못한다는 한계점이 있지만, 정렬하고자 하는 언어의 종류에 관계없이 적용가능하다는 유연성을 가진다. 또한 대량의 병렬말뭉치만 주어지면 자동으로 정렬을 수행할 수 있으므로 최근 많은 연구가 수행되고 있다.

통계 기반 단어 정렬 방법 중 대표적인 연구 중 하나로 IBM 모델을 들 수 있다[3]. IBM 모델은 병렬말뭉치에서 단어들의 공기 정보, 위치 정보 등을 이용하여 단어 정렬을 수행한다. 이때 발생 가능한 자료 부족 문제를 완화하기 위해 [4]는 단어들의 원형을 복원한 후 정렬을 수행하였다.

IBM 모델의 한계점 중 하나로 언어의 구조적, 문법적 정보를 고려하지 못한다는 점을 들 수 있다[5,6]. 이를 개선하기 위해 품사 정보를 통계적으로 반영하기 위한 연구가 존재한다[7]. 영중 기계번역을 위한 단어 정렬을 위해 영어와 중국어의 품사 정보를 통계학적 모델로 학습하여 품사의 정렬 경향을 반영하고자 하였다. 하지만 영어-중국어 간의 언어적인 차이는 문장을 구성하는 어순이나 기능어의 활용에서 영어-한국어 간의 언어적인 차이보다 적기 때문에 이를 영한 단어 정렬에서 반영하기에 어려움이 있다. [8]에서는 병렬말뭉치의 각 단어마다 품사 정보를 부착하고, 한국어 형태소 분석을 수행한 결과를 말뭉치로 사용하는 방법을 사용하였다. 또한 영어 단어에 대응되지 않는 한국어 품사들(조사, 어미 등)을 말뭉치에서 삭제하는 방법을 사용하였다.

이러한 기존 방법들은 단어 정렬을 수행하기 전에 입력 말뭉치를 변형하는 일종의 전처리 기법들이다. 본 논문에서는 IBM 모델의 단어 정렬 결과를 바탕으로 품사 정렬 경향을 파악한 후, 후처리 방법을 통해 정렬 결과를 수정하는 방법을 제시하고 실험을 통해 최종적인 단어 정렬 결과의 성능이 향상됨을 보인다.

3. 통계 기반 단어 정렬 성능 향상을 위한 방법

통계 기반 단어 정렬의 성능 향상을 위해서 본 논문에서는 기존에 연구된 전처리 방법 외에 품사 정보를 활용하여 단어 정렬 결과를 후처리 하는 방법을 제안한다.

표 2. 한국어 품사 집합

NNIN1	고유명사
NNIN2	일반명사
NNDE1	비단위성 의존명사
NNDE2	단위성 의존명사
PN	대명사
NU	수사
AN	관형사
ADCO	성분부사
ADSE	접속조사
CJ	접속사
VBMA	동사
AJMA	형용사

표 3. 영어 품사 집합 (Penn Treebank tagset)

CC	coordinating conjunction
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subordinating conjuction
JJ	adjective
MD	modal
NN	noun
NNP	proper noun
PDT	predeterminer
POS	possessive ending
PRP	pronoun
RB	adverb
TO	to
UH	interjection
VB	verb
WP	wh-pronoun
WRB	wh-adverb

3.1 품사간 정렬 경향 분석

IBM 모델을 이용하여 단어 정렬을 수행한 결과를 바탕으로 영어와 한국어 품사 간의 정렬된 경향을 파악하였다. 표 1은 말뭉치 내에 존재하는 한국어 동사와 주로 정렬 된 영어 품사들의 정렬 빈도를 나타낸다. 언어적 직관에 따르면 영어의 고유명사가 한국어 동사와 정렬이 될 가능성은 희박하다고 볼 수 있지만, IBM 모델을 통한 단어 정렬 결과를 수치적으로 분석한 결과, 다른 동사들보다 단수 고유명사와 월등히 높은 빈도로 정렬되었음을 확인할 수 있었다. 이 원인 중 하나로, 한국어에 비해 지나치게 세분화된 영어 동사류 품사들로 인

```
JJ,JJR,JJS -> JJ
NN,NNS -> NN
NNP,NNPS -> NNP
PRP,PRP$ -> PRP
RB,RBR,RBS -> RB
VB,VBD,VBG,VBN,VBP,VBZ -> VB
```

표 4. 품사간 정렬 빈도

	NNIN1	NNIN2	NNDE1	NNDE2	PN	NU	AN	ADCO	ADSE	CJ	VBMA	AJMA	총합
CC	13	404	1	3	2	1	1	260	3385	2156	188	437	6851
CD	8153	49485	458	10303	603	76440	3404	1839	257	56	14359	1396	166753
DT	25	14046	2	39	788	135	4961	430	678	1	378	2030	23513
EX	5	60	0	1	68	2	14	27	16	0	33	57	283
FW	62	138	4	1	5	4	3	24	1	0	51	10	303
IN	152	18229	185	1372	215	1251	56	1579	44	2	5712	832	29629
JJ	13959	139112	1948	1935	1664	3166	4893	9606	746	141	37560	33420	248150
MD	11	7860	1454	971	13	1	12	88	5	0	708	1914	13037
NN	67268	834283	6161	42177	8642	36161	8827	26804	4149	1054	150742	30376	1216644
NNP	174058	295984	2492	9486	4969	14642	3850	10450	1563	541	63611	9773	591419
PDT	0	95	2	0	21	3	442	133	4	0	3	87	790
POS	29	404	67	1	6	0	1	2	0	0	89	55	654
PRP	1227	22889	536	68	28433	582	2150	864	105	0	840	745	58439
RB	1448	33732	1469	1199	1648	768	1509	19098	8471	10	10980	9480	89812
TO	0	816	1	45	0	0	3	0	0	0	8	0	873
UH	108	218	3	7	59	5	3	41	21	0	91	21	577
VB	4799	75104	3384	1586	1329	1210	1648	6286	580	34	122750	7101	225811
WP	52	479	22	7	1288	42	537	237	13	0	65	63	2805
WRB	17	1239	66	5	349	0	34	1828	47	0	149	129	3863
총합	271386	1494577	18255	69206	50102	134413	32348	79596	20085	3995	408317	97926	

표 5. 단어 정렬 성능

실험 구성	Precision	Recall	F-measure
Baseline(품사부착, 형태소 분석)	42.9%	68.5%	52.8%
+ Proposed	41.2% (-1.7%)	73.5% (+5%)	52.8% (+0%)
Baseline + 영어 원형	43.9%	69.6%	53.9%
+ Proposed	41.7% (-2.2%)	74.4% (+4.8%)	53.5% (-0.4%)
Baseline + 영어 원형 + 기능형태소 제거	61.2%	65.8%	63.4%
+ Proposed	**63.9% (+2.7%)**	**73.8% (+8%)**	**68.5% (+5.1%)**

해 동사-동사간 정렬 빈도가 분산되었기 때문에 각각의 빈도가 단수 고유명사보다 낮아졌기 때문이다.

따라서, 유사한 품사 군을 통일할 필요성이 있다. 아래와 같은 방법으로 여섯 가지 영어 품사군을 선정하여 대표적인 품사로 통일하였다. 사용된 영어 및 한국어 품사 집합은 각각 표 2와 표 3에서 확인할 수 있다.

유사한 품사를 통일한 후 다시 수행한 정렬 결과에서 품사들 간의 정렬의 빈도수는 다시 살펴보았다. 표 4에 나타난 바와 같이 품사간 정렬 빈도에서 영한 단어 정렬에서 유사한 품사들 간의 정렬이 상대적으로 많이 일어나는 것을 볼 수 있다. 예를 들어 말뭉치 내에 존재하는 전체 영어 고유명사의 정렬 빈도(591,419회)중 한국어 고유명사와 정렬이 된 경우(174,058회)는 29.4%로 일반명사를 제외하고 가장 높은 비율을 보인다. 그 외에도 전체 영어 동사의 정렬 빈도(225,811회)중 한국어 동사와 정렬이 된 경우(122,750회)는 54.5%의 비율을 보인다. 이와 같이 자동으로 수행한 단어 정렬 결과를 기반으로 분석한 통계이기 때문에 오류를 포함하고 있지만 전반적으로 유사한 품사들끼리 정렬되는 경향을 보이는 것을 확인할 수 있다. 계산된 품사간 정렬 경향은 단어 정렬 시 주요한 정보로 활용 가능하다.

3.2 품사간 정렬 경향을 이용한 IBM 모델 후처리

IBM 모델을 통하여 생성된 단어 정렬의 결과에는 단어-단어의 정렬 확률을 포함하고 있다. 제안하는 방법은 이 확률에 품사간 정렬 경향을 반영하여 아래 수식과 같이 단어 정렬 점수를 계산하였다.

$$Score(w_e; w_k) = P_{IBM}(align = true | w_e, w_k) \times P(POS(w_e) | POS(w_k)) \times P(POS(w_k) | POS(w_e))$$

P_{IBM}은 IBM 모델에서 계산한 두 단어간의 정렬 확률이고, POS는 해당 단어의 품사를 알려주는 함수이다. 품사간 정렬 확률을 계산하기 위해 아래와 같이 상대빈도를 이용하였다.

$$P(POS(w_e) | POS(w_k)) = \frac{aligncount(POS(w_e); POS(w_k))}{count(POS(w_e))}$$

$$P(POS(w_k) | POS(w_e)) = \frac{aligncount(POS(w_k); POS(w_e))}{count(POS(w_k))}$$

$count$는 특정 품사가 병렬말뭉치 내에서 출현한 빈도를, $aligncount$는 병렬말뭉치 내에서 특정 두 품사가 정렬된 빈도를 나타내는 함수이다. 제안하는 방법은 변경된 정렬 점수인 $Score$를 기반으로 IBM 모델의 정렬 결과를 수정한다. 그림 1에 나타난 예제에서 변경 전에 '매장'(명사)이 'visiting'(동사)에 대응될 확률이 'store'(명

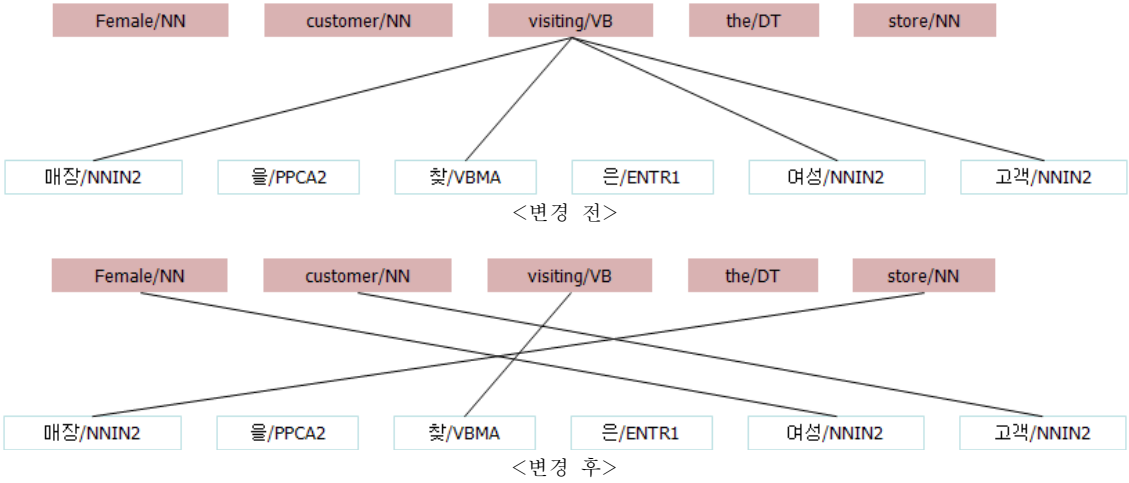

그림 1. 단어 정렬 결과의 변경 예제

사)에 대응될 확률보다 높게 추정되어 '매장'이 'visiting'에 대응되는 오류를 포함하고 있다. 제안하는 품사 경향을 고려하여 새롭게 정렬 점수를 계산한 결과, 동일한 명사 품사를 가진 '매장'과 'store'가 올바로 정렬 되게 수정한 것을 확인할 수 있다.

4. 실험

본 논문에서는 통계 기반 단어 정렬 성능 향상을 위해 품사간 정렬 경향을 IBM 모델의 결과에 반영하는 방법을 제안하였다. 이번 장에서 실험을 통해서 소개한 방법이 실제로 단어 정렬을 성능 향상시키는 것을 검증하였다.

실험에 사용된 말뭉치로 웹상에 공개된 영한 뉴스 기사쌍들을 웹로봇을 통해 수집하였다. 말뭉치는 약 21만 문장 쌍으로 이루어져 있다. 평가를 하기 위해서 해당 말뭉치에서 1,000개의 문장을 임의로 선별하여 수작업으로 단어 정렬 정답 집합을 구축하였다. 단어 정렬 정답 집합은 숙어에 해당되는 표현인 경우 기능 형태소를 포함하여 정렬을 하였으며, 그 외의 경우 정렬에 모호성이 없는 형식 형태소만을 정답으로 하였다. 평가 지표로 정확률(precision), 재현율(recall), F-measure를 사용하였다. 단어 정렬은 IBM 모델을 구현한 공개 툴킷 GIZA++를 사용하였다[9].

기존 연구에서 전처리 방법으로 제안된 말뭉치에 품사를 부착하고 한글 형태소 분석을 한 것을 Baseline으로 설정하였고, 그 외에도 영어 원형 정보 복원과 기능형태소 제거 등의 기존 연구를 추가하여 모두 비교 대상으로 삼았다.

표 5에서 각 전처리 방법들의 GIZA++결과에 대해 본 논문에서 제안한 방법으로 실험을 수행한 결과를 확인할 수 있다. Baseline으로 단어 정렬을 수행한 결과는 42.9%의 정확률, 68.5%의 재현율, 52.8%의 F-measure 값을 보였으며, 여기에 3장에서 소개한 품사간 정렬 경향을 반영하여 후처리 한 결과 41.2%의 정확률, 73.5%의 재현율, 52.8%의 F-measure 값을 보였다. F-measure의 성능이 기존 Baseline을 향상시키지 못했으며, 정확률은 하락하고 재현율이 상승하였다. 이는 GIZA++에서 단어-단어 대응 확률이 존재하더라도 정렬을 시키지 않는 단어들을 보정된 점수에 따라서 정렬이 추가되었기 때문이다. 그 중에서도 특히 한국어 정렬이 되지 않았던 기능 형태소들이 정렬이 됨에 따라서 정확률이 하락하였다. 하지만 재현율이 5% 상승함을 보이는 것은 잘못된 정렬을 수정하여 올바른 정렬로 변경되었음을 나타낸다.

두 번째 Baseline은 영어 원형을 복원하여 단어 정렬을 수행한 경우이다. 첫 번째 Baseline에 비해 정확률, 재현율, F-measure 수치가 미미하게 향상된 것을 확인할 수 있다. 여기에 제안한 후처리 방법을 적용한 경우 첫 번째 Baseline에서 적용한 결과와 비슷한 결과를 보인다. 마찬가지로 한국어 기능 형태소의 정렬이 큰 영향을 주기 때문으로 분석된다.

세 번째 Baseline은 추가적으로 한국어 기능 형태소를 제거한 경우이다. 이전 두 Baseline들에 비해 정확률을 향상시켰지만, 재현율을 하락시켰다. 이는 숙어적인 표현 외에도, 영어의 전치사 등 일부 단어들로 번역될 수 있는 형태소가 분명히 존재하기 때문이다. 이 문제를 해결하기 위해서 기능 형태소들에 대한 추가적인 분석이 필요하며, 세분화된 기능형태소의 제거가 필요하다. 여기에 제안한 방법을 적용한 결과 최종적으로 가장 높은 63.9%의 정확률, 73.8%의 재현율, 68.5%의 F-measure 값을 보인다. 이전 두 Baseline들에 포함되어 있던 기능형태소들로 인해 효과가 미미했던 제안하는 방법이 기능형태소의 제거에 따라 높은 효과를 보인 것으로 분석할 수 있다. 따라서 제안하는 방법을 효과적으로 사용하기 위해서는 기능형태소를 제거하는 전처리가 반드시 필요하다.

결론적으로 본 실험을 통하여 제안한 영한 단어 정렬에서 품사간 정렬 경향을 반영하여 단어 정렬의 결과를 변경하는 것이 성능 향상에 도움이 되는 것을 확인하였다.

5. 결론 및 향후 계획

본 논문에서 통계 기반 단어 정렬의 성능을 향상시키기 위해 품사간 정렬 경향을 정렬 결과에 후처리로 반영하는 방법에 대해서 소개하였다. 통계 기반 단어 정렬에 언어의 특성을 반영하기 위한 후처리 방법을 제안하였다.

실험에서 확인된 문제로, 기능형태소를 제거하였을 때 단어 정렬의 재현율이 하락하는 문제를 해결하기 위해서 기능 형태소들에 대한 추가적인 분석이 필요하며, 세분화된 기능 형태소의 제거가 필요하다.

품사 정보를 활용한 단어 정렬 결과의 변경의 방법으로 영한 단어 정렬과 한영 단어 정렬의 합집합을 이용하는 것을 고려해 볼 수 있다. 이 경우, 단어간 정렬 확률을 이용하지 않고, 대응될 수 없는 품사쌍들을 파악하여, 이를 합집합으로부터 제거하는 방향을 고려할 수 있다.

단어 정렬에서 품사 정보를 고려하는 것은 영한 단어 정렬이 아닌 다른 언어 간의 단어 정렬에서도 공통적으로 사용할 수 있다. 언어들 사이에서 사용되는 품사들의 종류는 다르겠지만, 서로 유사한 품사들간의 경향성을 파악한다면 이를 그대로 반영할 수 있을 것으로 예상된다.

감사의 글

본 연구는 SK텔레콤의 “영한/중한 기계번역을 위한 한국어 언어모델 및 번역지식 구축도구 개발” 과제의 지원에 의해 이루어졌음.

참고문헌

[1] 리금희. 외. 중-한 대조분석정보를 이용한 단어정렬. 제14회 한글 및 한국어 정보처리 학술발표 논문집 pp. 40-46. 2002.

[2] Peter F. Brown., et al., A Statistical Approach to Machine Translation. Computational Linguistics, vol. 16, pp. 79-85. 1990.

[3] Peter F. Brown., et al., The Mathematics of Machine Translation : Parameter Estimation. Computational Linguistics, 19(2). 1993.

[4]Sharon Goldwater and David McClosky. Improving Statistical MT through Morphological Analysis. ACL-2005.

[5] Kenji Yamada and Kevin Knight. A Syntax-based Statistical Translation Model. ACL-2001.

[6] Ye-Yi Wang. Grammar Inference and Statistical Machine Translation. Ph.D. thesis, Carnegie Mellon University. 1998.

[7] Chang, J. S. and M. H. C. Chen. Using Partial Aligned Parallel Text and Part-of-speech Information in Word Alignment. In Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA). pp. 16-23. 1994.

[8] Jonghoon Lee, Donghyeon Lee, and Gary Geunbae Lee. Improving Phrase-based Korean-English Statistical Machine Translation. INTERSPEECH 2006.

[9] Franz Josef Och. GIZA++: Training of statistical translation models. http://www.fjoch.com/GIZA++.html 2001.