병렬 말뭉치를 이용한 도메인 특화 사전 자동 추출 연구

박은진○, 황금하, 김영길
한국 전자 통신 연구원 자연언어처리연구팀
{ejpark,hgh,kimyk}@etri.re.kr

A Study of Automatic Extraction of Domain Specified Dictionary

Eun-jin Park○, Kum-ha Hwang, Young-gil Kim
Natural Language Processing Team, Electronics and Telecommunications Research Institute

요 약

본 논문에서는 도메인별 병렬 말뭉치를 이용하여 해당 도메인에 특화된 한영 대역쌍을 Moses Toolkit을 이용하여 자동 추출하였다. 이렇게 추출된 대역쌍은 도메인 특화 자동 번역 시스템의 번역 사전으로 사용하기에는 많은 오류가 포함되어 있기 때문에, 본 논문에서는 이를 효율적으로 제거할 수 있는 식을 제안하였다. 본 논문에서 제안한 식으로 오류를 제거한 결과, 임계값 0.5를 기준으로 추출된 한영 대역쌍이 1,098개였고, 이는 실험에 사용한 기업 분야 병렬 말뭉치 42,200문장 중에서 29,292문장(69.4%)에 영향을 주었다. 자동으로 추출한 도메인 특화 번역 지식을 기존 자동 번역 시스템의 번역 지식에 적용한 결과 BLEU가 0.0054 향상되었다.

주제어: 병렬 말뭉치, 대역어 자동 추출, SMT, 특화 사전

1. 서 론

도메인 특화 패턴 기반 자동 번역 시스템은 일반 분야 자동 번역 시스템보다 번역 성능이 우수하다[1,2]. 그 이유는 해당 도메인에서 자주 사용하는 대역어와 전문 용어를 번역문에 사용하기 때문에 좀 더 자연스러운 대역문을 생성할 수 있다. 그러나 도메인 특화 자동 번역 시스템은 대역문 생성이 자연스러운데 반해서, 도메인 특화 번역 사전을 구축하는데 많은 시간과 비용이 드는 단점이 있다. 도메인 특화 사전을 구축하기 위해서는 해당 도메인의 전문가 혹은 그에 준하는 교육을 받은 사람이 해당 도메인에 맞는 대역어를 구축해야 되기 때문이다.

통계 기반의 자동 번역 시스템은 그 인코딩 단계에서 대량의 병렬 말뭉치를 자동으로 기계 학습하여 번역 지식을 생성하여, 시스템의 번역 지식으로 사용한다[3-5]. 이 시스템에서 자동으로 학습된 번역 지식은 학습 말뭉치의 통계 정보가 반영되어 있어서, 해당 도메인에 가장 적합한 대역어를 많이 포함하고 있다. 때문에 해당 도메인의 말뭉치에서 통계 기반으로 지식을 자동으로 학습하여, 도메인 특화 자동 번역 시스템의 번역 지식으로 사용할 수 있다면, 도메인 특화 자동 번역 시스템의 번역 지식을 구축하는데 드는 시간과 노력을 획기적으로 줄일 수 있을 것이다. 그러나 통계 기반 자동 번역 시스템의 인코딩 모델은 한국어-영어 언어 쌍과 같이, 문장 어순이 다른 언어일 경우에 아직 많은 한계를 보여 주고 있다. 이렇게 자동으로 학습한 지식에는 기계 학습 오류를 많이 포함하고 있다. 때문에 자동 학습된 사전을 지식의 정확도에 대한 요구가 높은 패턴 기반 자동 번역 시스템에 사용하기 위해서는, 자동 학습 사전의 정확도를 높이기 위한 필터링 과정이 필요하다.

본 논문에서는 통계 기반 자동 번역 시스템인 Moses Tookit으로 자동 학습된 구단위 번역지식 결과를 이용하여, 높은 정확도의 도메인 특화 번역 사전을 추출하는 방법을 제안한다. 자동으로 추출한 번역 사전이 사람의 검증을 거치지 않고, 도메인 특화 기계번역 시스템에 긍정적인 역할을 하는지 검증하기 위하여, 본 논문에서는 기존 번역 사전으로 번역한 결과와 자동 학습된 사전에서 추출한 번역 사전으로 번역한 결과를 자동 평가하여 성능이 향상되었는지를 검증하였다.

본 논문의 구성은 2장에서 관련 연구를 기술하고, 3장에서는 본 논문에서 제안하는 시스템 구성을 설명하고, 4장에서는 실험 결과를 분석한다. 마지막으로 5장에서는 결론 및 향후 과제에 대해 언급한다.

2. 관련 연구

병렬 말뭉치에서 자동으로 어휘 사전을 추출하는 연구가 있었다[6]. 이 연구에서는 언어 규칙에 기반한 도메인 특화 알고리즘을 통해서 한중 어휘 사전을 자동으로 추출하였다. 이 연구에서 88% 정확률과 66%의 리콜을 보였다. 본 논문에서는 별도의 알고리즘이나 언어 규칙을 사용하지 않고, 어휘의 출현 확률을 이용하였다.

통계 기반 자동 번역 방식의 번역 성능을 자동 평가 방법 중 가장 널리 쓰이는 방법으로 BLEU(Bilingual Evaluation Understudy)가 있다[7-10]. 이들 연구에서

는 BLEU 평가 결과가 사람이 하는 평가 결과와 비례한다고 말하고 있다. 본 논문에서는 이 BLEU 방법으로 자동으로 필터링한 한영 대역 사전의 성능을 측정하였다

3. 시스템 구성

본 논문의 시스템 구성은 그림1과 같다.

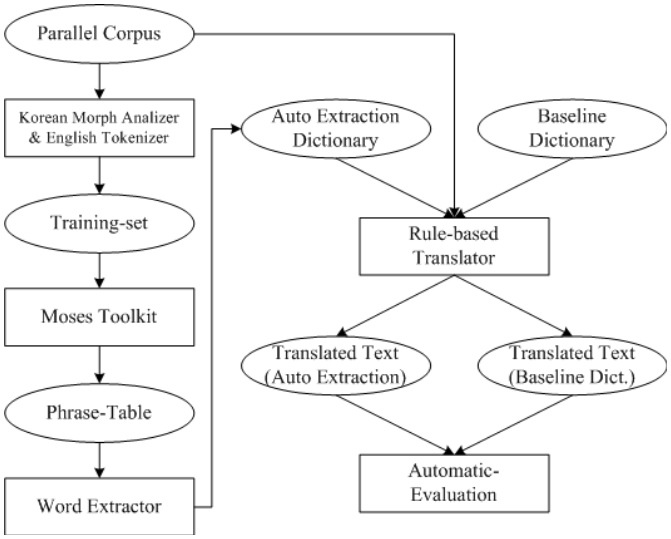

그림 1 시스템 구성.

그림 1에서, 병렬 말뭉치가 주어지면, 한국어 문장을 형태소 분석하고, 영어 대역문은 문장 부호와 영어 단어를 분리한다. 예를 들어 표 1과 같은 병렬 말뭉치가 주어지면 표 2와 같은 형태로 한국어 부분은 형태소 분석하고, 영어 대역문의 문장 부호와 영어 단어를 분리한다.

표 1 병렬 말뭉치 원문.

ANCHOR AWAY : 출항하기 위하여 UP & DOWN을 한 시각이며, LAST LINE을 벗긴 후 어떠한 사유로 인하여 ANCHORING을 한 경우 다시 출항하기 위하여 UP & DOWN ANCHOR를 한 시각. ANCHOR AWAY : the time when the anchor was up & down situation for sailing, and in case of dropping anchor after casting off last mooring line by any reason, the time of next up & down anchor for sailing is regarded as the time of anchor away.

표 2 SMT[1] 학습을 위한 Training-set 포멧.

ANCHOR[외국어] AWAY[외국어] :[컴마기호] 출항하[일반동사] 기_위하여[종속연결어미] UP[외국어] &[기타기호] DOWN[외국어] 을[목적격조사] 하[일반동사] ㄴ[관형사형전성어미] 시각[용언불가능보통명사] 이[긍정지정사] 며[대등연결어미] ,[컴마기호] LAST[외국어] LINE[외국어] 을[목적격조사] 벗기[일반동사] ㄴ_후[종속연결어미] 어떠하[지시형용사] ㄴ[관형사형전성어미] 사유[용언불가능보통명사] 로_인하여[부사격조사] ANCHORING[외국어] 을[목적격조사] 하[일반동사] ㄴ[관형사형전성어미] 경우[용언불가능보통명사] 다시[성상상태부사] 출항하[일반동사] 기_위하여[종속연결어미] UP[외국어] &[기타기호] DOWN[외국어] ANCHOR[외국어] 를[목적격조사] 하[일반동사] ㄴ[관형사형전성어미] 시각[용언불가능보통명사] .[문미기호] ANCHOR AWAY : the time when the anchor was up & down situation for sailing , and in case of dropping anchor after casting off last mooring line by any reason , the time of next up & down anchor for sailing is regarded as the time of anchor away .

표 2와 같은 형태로 만들어진 학습셋을 통계 기반 자동 번역 시스템인 Moses Toolkit으로 학습한다[3]. Moses 학습 결과로 생성된 Phrase-table에서 단어 정렬 정보를 추출한다[2]. 여기서 Moses는 표 3과 같은 단계로 학습한다.

표 3 Moses 학습 단계.

1. Prepare data
2. Run GIZA
3. Align words
4. Lexical translation
5. Extract phrase
6. Score phrases
7. Reordering model
8. Generation model
9. Configure file

본 논문에서는 1~8단계까지 학습한 결과인 Phrase-table을 이용한다. Phrase-table의 구조는 표 4와 같다.

표 4 Phrase-Table의 구조.

,[컴마기호] 폐기물[용언불가능보통명사] 관리[용언가능보통명사] 계획[용언가능보통명사] ||| , garbage management plans ||| (0) (1) (2) (3) ||| (0) (1) (2) (3) ||| 1 0.0423758 1 0.00871591 2.718

Phrase-table은 [원문] ||| [대역문] ||| [원문-대역문 정렬 정보] ||| [대역문-원문 정렬 정보] ||| [번역 확률]와 같은 구조로 되어 있다[11]. 본 논문에서는 원문-대역문 정렬 정보와 대역문-원문 정렬 정보를 이용한다[12]. 표 5는 표 3의 정렬 정보를 기준으로 추출한 대역어 정보이다.

표 5 단어 정렬 정보 추출.

한국어 -> 영어	
,[컴마기호]	,
폐기물[용언불가능보통명사]	garbage

1) 통계 기반 자동 번역 (Statistical Machine Translation)

2) GIZA++정렬 결과는 문장 단위 출현 빈도를 추출할 수 있지만, Moses Toolkit의 학습 결과에서는 구문 단위의 단어 출현 빈도를 추출할 수 있다. 본 논문에서는 Phrase-table의 구문 단위의 단어 출현 빈도를 사용하였다.

관리[용언가능보통명사]	management
계획[용언가능보통명사]	plans
영어 -> 한국어	
,	,[컴마기호]
Garbage	폐기물[용언불가능보통명사]
Management	관리[용언가능보통명사]
Plans	계획[용언가능보통명사]

Phrase-table에서 추출한 대역어 정렬 정보에는 표 6와 같은 정렬 정보가 많이 포함되어 있다. 표 6와 같은 정렬 정보는 통계기반 자동 번역 시스템에 필요한 정보이지만 도메인 특화 자동 번역 시스템에 사용하기에는 적합하지 않은 정보이다.

표 6 단어 정렬 오류

한국어	영어
3차로[용언불가능보통명사]	subset
가능[용언가능보통명사]	radar
가능성[용언불가능보통명사]	any
경증[용언불가능보통명사]	Solutions
경험[용언가능보통명사]	Now
과당[용언불가능보통명사]	for
과량[용언불가능보통명사]	15-years-old
관리[용언가능보통명사]	and
관성력[용언불가능보통명사]	6.1.2.10
관절염[용언불가능보통명사]	1468
괴혈병[용언불가능보통명사]	In

표 6와 같은 오류를 제거하기 위하여, 본 논문에서는 식 (1)을 제안한다.

$$x = \frac{f(kr_i, en_j)*2}{f(kr_i)+f(en_j)} \quad (1)$$

여기서, $f(kr_i)$는 i번째 한국어 단어의 출현 빈도이고, $f(en_j)$는 j번째 대역어 빈도이고, $f(kr_i, en_j)$은 i번째 한국어 단어 중에서 j번째 대역어가 나타나는 빈도이다.

본 논문에서 제안하는 식 (1)는 한영 대역쌍의 빈도가 한국어 혹은 영어 단어와 비슷한 빈도로 말뭉치에나타나면 높은 값을 가진다. 명사의 경우에는 거의 동일한 대역어를 가지기 때문에 잘못된 정렬 정보를 가지는 대역쌍을 제거할 수 있다. 다시 말해서, 영어 전치사의 경우, 영어 원문에 나타난 빈도가 매우 높으므로 식 (1)을 적용하면 낮은 값을 가진다. 그러나 전문 용어의 경우, 한국어 원문에 나타는 빈도와 영어 대역문에 나타난 빈도가 거의 같기 때문에, 식 (1)에서 1에 가까운 값을 가진다. 또한 동일한 의미를 가진 명사 단어라도, 병렬 말뭉치의 출현 빈도가 적용되기 때문에 높은 출현 빈도의 대역어가 높은 값을 가진다. 예를 들어서, 임의의 한국어 단어 "제안"이 한국어 원문에 100번 나타났고, 대응되는 영어 단어인 "proposal"이 100번, "offer"가 100번씩 각각 나타났고, 정렬 빈도 ("제안", "proposal")이 80번, ("제안", "offer")가 20번이라면, 식 (1)을 적용했을 때, ("제안", "proposal")은 0.8, ("제안", "offer")는 0.2로 계산된다. 즉, 본 논문에서 제안하는 식 (1)을 적용하면 한국어 단어 "제안"을 영어 단어 "proposal"로 선택하도록 되어 있다. 만약 병렬 말뭉치가 기업에서 사용된 문장들이라면, "제안"이라는 대역어로 "offer"보다는 "proposal"가 더 자연스러운 대역어라는 의미가 된다.

병렬 말뭉치의 한국어 원문을 자동으로 추출한 대역사전과 기존의 사전을 도메인 특화 자동 번역 시스템에 적용하여 번역한다. 두 개의 자동 번역문과 병렬 말뭉치의 영어 대역문을 자동 평가하여 BLEU 차이를 측정하여 자동으로 추출한 대역 사전을 평가한다.

4. 실험 및 결과

실험에 사용한 병렬 말뭉치는 기업 관련 문서 42,200 문장을 사용하였다. 본 논문에서는 Moses Toolkit으로 학습한 결과인 Phrase-table에서 명사 51,762개를 추출하였다. 추출한 명사 51,762개에 본 논문에서 제안하는 식 (1)을 적용한 결과 그림 2와 같은 분포로 나타났다.

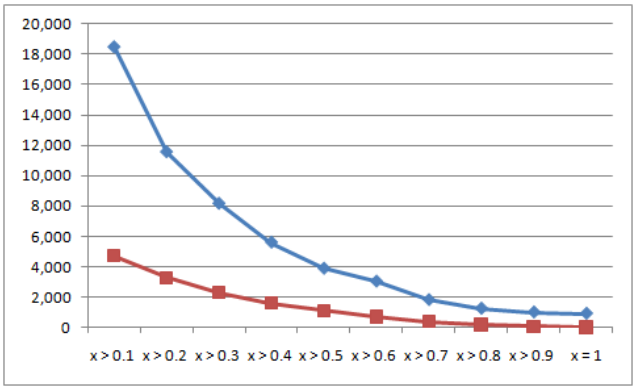

그림 2 Phrase-table 필터링 결과

그림 2에서 x축은 식 (1)의 값을 나타내고, y축은 Phrase-table에서 추출한 명사의 개수를 나타낸다. 실험에는 한국어-영어 대역어 빈도가 10 이상이고, 식 (1)의 x 값이 0.9~0.5가 되는 대역쌍을 추출한 결과 각각 59, 169, 390, 699, 1,098개가 되었다. 전체 병렬 말뭉치 42,200 문장 중에서 한영 대역 쌍이 나타난 병렬 문장을 추출하여 각각 2,315, 8,336, 14,603, 22,428, 29,292 문장을 추출하였다. 이렇게 추출한 문장을 기존의 번역 사전으로 자동 번역하고, 앞에서 자동으로 추출한 한영 대역쌍을 번역 사전에 반영하여 자동 번역하였다. 서로 다른 두 개의 사전을 각각 적용한 도메인 특화 자동 번역 시스템의 번역 결과와 병렬 말뭉치의 영어 대역문을 자동 평가한 결과, 그림 3과 같이 나타났다.

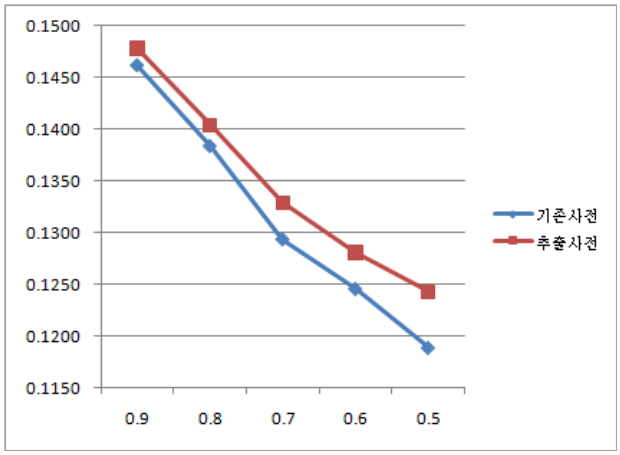

그림 3 자동 평가 결과

그림 3에서 x축은 식 (1)의 x 값을 나타내고, y축은 BLEU값을 나타낸다. 기존 도메인 특화 자동 번역 시스템의 번역 사전(기존사전)의 자동 평가 결과보다 본 논문에서 자동 추출한 사전이 반영된 자동 평가 결과가 성능이 더 좋게 나타났다. 식 (1)의 x 값이 0.5를 기준으로 기존의 번역 사전으로 번역한 번역문의 평가 결과가 0.1189 이였고, 자동으로 추출한 사전으로 자동 번역한 번역문의 평가 결과가 0.1243로 0.0054가 더 높게 나타났다. 0.5를 기준으로 한국어-영어 단어쌍이 1,098개이고, 이 단어쌍이 포함된 병렬 문장 수가 29,292개로 전체 입력 말뭉치의 69.4%를 포함하였다.

5. 결 론

본 논문에서는 통계 기반 자동 번역 시스템인 Moses Toolkit의 학습 모듈을 사용하여 자동 학습한 번역 사전에서 도메인 특화 자동 번역 시스템에서 사용할 수 있는 번역 사전을 자동 추출하였다. Moses Toolkit의 구단위 기계 학습 결과인 phrase-table은 도메인 특화 자동 번역 시스템의 번역 사전으로 사용하기에는 많은 오류가 있는데, 본 논문에서는 이러한 오류를 효율적으로 제거하는 방법을 제안하였다. 본 논문에서 제안한 방법으로 추출한 대역어 쌍을 도메인 특화 자동 번역 시스템의 번역 지식에 적용한 결과, BLEU가 0.0054 향상되는 결과를 얻었다. 이는 특히 자동 학습한 지식이 사람의 검증을 거치지 않고 자동 필터링으로만 얻어진 결과이기서 더욱 고무적인 결과로 보여 진다.

본 논문에서는 단일 명사 정렬 정보만을 추출하였다. 추후 연구로는 복합 명사의 대역어 추출 연구로 이어져야 될 것이다. 또한 추가적인 실험을 통하여 가장 최적의 임계값을 측정하는 연구로 이어져야 될 것이다.

참고 문헌

[1] Hong M.P., Kim Y.G., Kim C.H., Yang S.I., Seo Y.A., Ryu C. & Park S.K. *Customizing a Korean-English MT System for Patent Translation*. MT SummitX, pp.181—187. 2005.

[2] Ki-Young Lee, Sung-Kwon Choi, Oh-Woog Kwon, Yoon-Hyung Roh, Young-Gil Kim: *Domain Adaptation for English-Korean MT System: From Patent Domain to IT Web News Domain*. ICCPOL, pp.321-328. 2009.

[3] http://www.statmt.org/moses/

[4] P. Koehn, F.J. Och, and D. Marcu. *Statistical phrase based translation*. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT/NAACL). 2003.

[5] Philipp Koehn and Hieu Hoang. *Factored Translation Models*, Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, 2007.

[6] Necip Fazil Ayan, Bonnie Dorr, and Okan Kolak. *Domain Tuning of Bilingual Lexicons for MT*. In Proceedings of the Evaluation Workshop at the MT Summit IX, pp.3-11, 2003.

[7] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). *BLEU: a method for automatic evaluation of machine translation* in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp.311-318. 2002.

[8] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. *BLEU: a method for automatic evaluation of machine translation* in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp.311-318. 2002.

[9] Callison-Burch, C., Osborne, M. and Koehn, P. *Re-evaluating the Role of BLEU in Machine Translation Research* in 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006, pp.249-256. 2006.

[10] Doddington, G. *Automatic evaluation of machine translation quality using n-gram cooccurrence statistics* in Proceedings of the Human Language Technology Conference (HLT),

pp.128-132. 2002.
[11] Richard Zens and Hermann Ney. *Efficient Phrase-table Representation for Machine Translation with Applications to Online MT and Speech Translation*, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2007.
[12]http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases