Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 3, June 2013, pp. 191-195
http://dx.doi.org/10.5391/JKIIS.2013.23.3.191

링크확률과 개체명 인식을 이용한 영-한 교차언어 링크 탐색

English-Korean Cross-lingual Link Discovery Using Link Probability and Named Entity Recognition

강신재*
Shin-Jae Kang†

*대구대학교 정보통신대학 컴퓨터 · IT공학부
† School of Computer and Information Technology, Daegu University

요 약

본 논문에서는 방대한 웹 자원의 연결성을 더욱 증가시키기 위해 영어 위키피디아 문서로부터 한국어 위키피디아 문서로의 교차언어 링크를 자동으로 탐색하는 방법을 제안한다. 어구의 링크확률을 대략 추정하여 사용하던 기존의 방법에 비해, 본 연구에서는 위키피디아 문서 집합으로부터 추출한 제목 목록과 링크 확률과 같은 다양한 정보들과 개체명 인식 결과를 함께 사용하여 링크가 걸릴 앵커 후보를 선택한다. 앵커 후보를 한국어 대역어로 번역한 후, 대역어에 가장 적합한 한국어 웹문서를 찾아 교차언어 링크로 설정하게 된다. 실험한 결과 MAP 수치로 0.375를 얻었다.

키워드 : 교차언어 링크탐색, 링크 추천, 링크 확률, 위키피디아

Abstract

This paper proposes an automatic method for discovering cross-lingual links from English Wikipedia documents to Korean ones in order to increase connectivity among vast web resources. Compared to the existing methods roughly estimating link probability of phrases, candidate anchors are selected from English documents by using various information such as title lists and linking probability extracted from Wikipedia dumps and the results of named-entity recognition, and the anchors are translated into Korean words, and then the most suitable Korean documents with the words are selected as cross-lingual links. The experimental results showed 0.375 of MAP.

Key Words : Cross-lingual link discovery, Link identification, Link probability, Wikipedia

1. 서 론

웹에는 영어, 한국어 등 다국어로 작성된 방대한 양의 정보가 존재한다. 이러한 정보들로부터 링크 정보, 의미 정보 등 다양한 정보들을 자동으로 마이닝(mining)하여 지능형 검색이나 개인화 추천 등 보다 지능화된 웹 서비스의 구현에 활용하고자 하는 것이 최근 연구의 추세이다. 하지만 동일한 언어로 작성된 웹페이지 사이에는 링크가 다수 존재하지만, 서로 다른 언어로 작성된 웹페이지 사이에는 링크가 거의 존재하지 않기 때문에 다국어 웹 문서 사이에 존재하는 정보를 제대로 활용하지 못하는 측면이 있다. 이를 극복하기 위하여 최근 등장한 연구가 다른 언어로 작성된 웹 페이지 사이에 링크를 자동 설정해 주는 교차언어 링크 탐색(Cross-Lingual Link Discovery, 이하 CLLD)이다[1]. 영-한 CLLD의 예를 그림 1에 제시하였다. 영어 위키피디아(Wikipedia)에 있는 'Social Networking Service'라는 토픽(topic) 문서에서 링크가 걸린 만한 단어나 어구, 예를 들어 'e-mail'을 선택하고, 이에 해당하는 한국어 위키피디아 문서 '전자 우편'을 찾아서 링크로 걸어주는 것이다.

'링크드 오픈 데이터'(Linked Open Data, 이하 LOD)는 인터넷 상의 각 사이트에서 RDF형식으로 데이터를 제공하여 시스템 간 데이터를 공유/연결할 수 있도록 구현한 것으로, 시맨틱 웹을 통하여 다양한 지능적인 부가서비스를 개발하고자 할 때 필수적인 거대 지식베이스이다. 영-한 CLLD를 통하여 영어 정보 중심으로 구축되어 있는 LOD에 관련 한국어 정보를 어느 위치에 추가할 것인지를 판단할 수 있게 되며, 언어 장벽으로 인한 타 언어 정보 검색의 어려움을 해소할 수 있게 된다.

영-한 CLLD를 위해서는 일반적으로 3단계를 거치는데, 영어 문서에 포함되어 있는 단어나 구 가운데 링크가 걸릴 만큼 의미 있는 것(이하 앵커 또는 키워드)을 후보로 추출(anchor extraction)하고, 앵커 후보를 한국어로 번역

접수일자: 2013년 3월 20일
심사(수정)일자: 2013년 5월 3일
게재확정일자 : 2013년 5월 4일
† Corresponding author
본 논문은 2011년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(2011-0007025)

(anchor translation)한 다음, 번역된 앵커 대역어에 가장 적합한 한국어 웹 페이지를 선택(target link selection)하는 단계를 거친다.

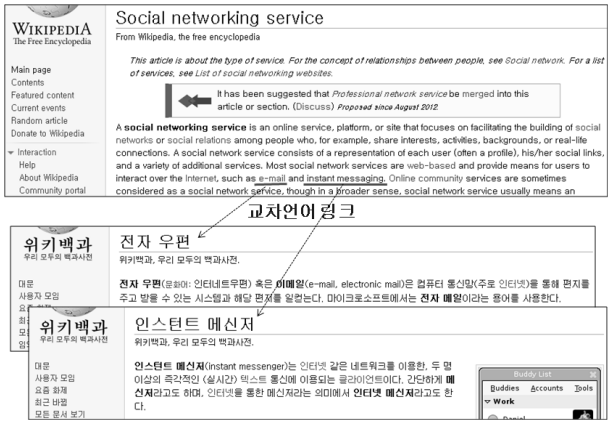

그림 1. 영-한 CLLD의 예
Fig. 1. Example of English-Korean CLLD

최근 웹 데이터 마이닝 연구 분야에서 많이 활용하는 리소스인 위키피디아를 살펴보면 영-한 페이지간 링크는 다소 포함되어 있기는 하지만, 실제로 연결이 가능한 영-한 페이지 사이의 모든 링크가 설정되어 있지는 않다.

본 논문에서는 CLLD 연구를 위한 평가 도구[2]가 공개되어 있는 위키피디아를 도메인으로 하고, 웹 문서 가운데서도 가장 많은 부분을 차지하고 있는 영어 웹 문서로부터 한국어 웹 문서로의 링크를 찾는 영-한 CLLD를 링크 확률과 개체명 인식 기법을 이용하여 해결하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 관련연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 영-한 CLLD 방법에 대해 설명한다. 4장에서는 실험 및 평가 결과를 제시하고, 마지막 5장에서 결론 및 향후 계획에 대하여 기술한다.

2. 관련 연구

Mihalcea[3]는 동일한 언어로 작성된 문서 사이의 링크를 탐색하기 위해, 단어나 어구가 링크될 확률을 의미하는 key-phraseness를 구하여 키워드 추출에 사용하였고, 위키피디아 문서 내 하이퍼링크의 주변 문맥을 추출하여 기계학습을 한 후 이를 키워드의 중의성 해소에 활용하는 방법을 제안하였다. 이 방법은 영어 위키피디아에만 적용되었다.

Tang[4]은 영-중, 영-일, 영-한 언어별로 다른 방법으로 CLLD를 구현하였다. 영어 문서로부터 중국어로의 링크를 찾기 위해서 구글 번역기의 기계 번역과 음역(transliteration) 기능을 이용하여 앵커 후보를 찾은 후, 페이지 이름 매핑과 링크 확률, 교차언어 정보검색 시스템 등을 적용하여 교차언어 링크를 추천하였다. 일본어로의 링크를 찾기 위해서는 개체명 인식과 위키피디아 라이브 검색 기능을 이용하였으며, 한국어로의 링크를 찾기 위해서는 페이지 이름 매핑 알고리즘만을 사용하였다. 언어쌍별 구현 방법이 다른 이유는 자연어 처리를 위한 언어별 리소스 구축이 쉽지 않고, 공개된 오픈소스도 한정되어 있기 때문이다. Tang이 사용한 기법 가운데 음역은 간단하지만 그다지 정확하지 않고, 페이지 이름 매핑도 간단하지만 페이지 제목이 일치하는 것만 찾을 수 있는 단점이 있다.

Kim[5]은 명사구, 개체명, 문서 제목, N-gram 단어 등으로부터 앵커 후보를 선택한 후, 링크 확률에 따라 랭킹하였다. 랭킹된 앵커들은 위키피디아 내에 존재하는 교차언어 제목쌍 목록 검색, 구글 번역, 다국어 사전 검색의 순으로 적용하여 번역하였고, 번역된 앵커와 정확히 일치하는 문서를 최종 링크로 설정하였다. 여기서 일치하는 문서가 없는 경우에는 링크가 걸릴 대상 문서로 들어오는 링크들의 앵커 텍스트를 색인한 후, 번역된 앵커를 질의어로 정보 검색하고 우선순위가 가장 높은 검색 결과의 문서로 최종 링크를 설정하였다. 언어 독립적인 방법을 제안하여 언어별 확장을 쉽게 하고자 하였으나, 언어에 의존적인 리소스를 활용하지 못한 단점이 있다.

Kang[6]은 영어 위키피디아의 링크 문맥정보를 이용하여 앵커를 추출하고, 영한 대역어 사전으로 앵커 대역어 후보들을 얻은 후, 영-한 토픽 문서 간 유사도를 계산하여 앵커와 앵커 대역어 간 단어의미 중의성을 해소하고자 하였다.

Adar[7]는 영어, 스페인어, 프랑스어, 독일어로 작성된 위키피디아를 대상으로 관련된 토픽 문서의 인포박스(Infobox) 매핑을 통하여 특정 언어의 페이지에 빠져있는 정보를 다른 언어의 페이지에서 찾거나, 상호 불일치하는 정보를 탐지하는 방법을 제안하였다. 기계학습을 이용하여 다국어간 페이지 정렬, 인포박스 정렬, 인포박스 병합의 순으로 처리하였다. CLLD가 문서 전체의 내용을 대상으로 링크를 설정하는 반면, 이 연구는 인포박스만을 대상으로 한다는 측면에서 차이가 있다.

3. 영-한 CLLD 과정

3.1 CLLD 필요 리소스 구축

위키피디아 문서의 제목에 사용된 단어들은 다른 문서 내에서 앵커로 사용될 가능성이 높으므로, 위키피디아 문서 집합을 분석하여 각 어구별 링크 확률을 계산하는 것이 필요하다. 먼저, 웹문서에서 태그와 기호, 숫자를 제거하고 리다이렉트(redirect)된 페이지 목록과 페이지로 들어오는 링크(incoming links), 페이지로부터 나가는 링크(outgoing links)에 연결된 앵커 텍스트를 추출하는 전처리 과정을 거쳤다. 이를 통해 정규화된 제목 목록을 구한 후 단어별 링크 확률 등의 정보를 구축하였다. 단어 w의 링크 확률 $LinkProb(w)$는 수식 (1)을 이용하여 구하였다. $TotalFreq(w)$는 문서 집합에서 단어 w가 출현한 빈도이고, $AncFreq(w)$는 문서 집합에서 단어 w가 앵커 텍스트로 사용된 빈도이다.

$$LinkProb(w) = \frac{AncFreq(w)}{TotalFreq(w)} \qquad (1)$$

Mihalcea[3]는 단어 w가 앵커로 사용된 문서의 수를 단어 w가 출현한 문서의 수로 나누어 링크 확률을 계산하였다. 이는 출현 빈도가 충분하지 않은 경우에 괜찮은 추정 방법이지만, 개별 단어의 출현 빈도를 직접 사용하지 않았기 때문에 왜곡된 링크 확률이 추정될 수 있는 문제도 가지고 있다. 본 연구에서는 위키피디아 전체 문서 집합을 대상으로 충분한 출현 빈도를 추출하였기 때문에 수식 (1)을 이용하였다. 2012년 7월에 덤프된 위키피디아로부터 총

3,733,511개의 앵커 후보 어구에 대해 링크 확률을 구축하였다.

다음은 영어와 한국어 위키피디아 문서 집합으로부터 추출한 문서 제목과 앵커로 사용될 수 있는 단어들의 링크 확률, 영-한 링크 목록 등의 정보를 추출함을 보여 준다.

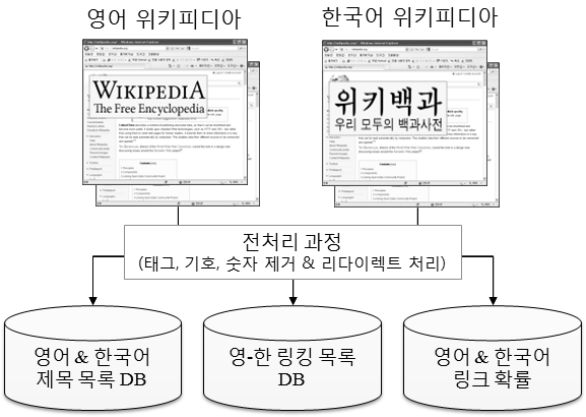

그림 2. 위키피디아 덤프로부터 CLLD에 필요한 리소스 구축

Fig. 2. Resource construction for CLLD from Wikipedia dump

다음 표 1에서 한국어 제목 목록의 경우는 최신 한국어 위키피디아 덤프로부터 더 많이 수집할 수도 있겠으나, 본 연구의 실험에서 사용한 NTCIR-9[8] 테스트 집합에서 교차언어링크를 설정할 한국어 위키피디아 문서 집합을 20만여 개로 한정시켜 놓았으므로, 이것만 추출하여 구축하고 추가 수집하지는 않았다.

표 1. 구축된 리소스

Table 1. Constructed resources

리소스 종류	갯수
영어 제목 목록	12,487,248
한국어 제목 목록	201,486
영-한 링크 목록 (영어 제목 - 한국어 제목)	134,920
영-한 대역어 사전 목록 (영어 단어 - 한국어 대역어)	186,773

3.2 앵커 후보 추출 & 랭킹

영어 위키피디아 문서에서 앵커를 선택하는 데 있어, 모든 단어를 대상으로 하기보다, 위키피디아 덤프에서 추출한 영어 제목 목록에 포함되어 있는 단어나 구를 대상으로 먼저 검색하는 것이 바람직하다. 이는 제한된 어휘(controlled vocabularies)를 이용하는 형태인데, 링크가 걸릴 가능성이 높은 앵커 후보를 우선적으로 선택할 수 있게 하는 효과를 가져 온다. 최장일치법을 사용하여 매칭하게 되며, 기 계산된 영어 링크 확률 정보를 이용하여 앵커 후보들 간에 랭킹을 한다. 또한 개체명에 해당하는 어구는 링크 확률이 없더라도 앵커 후보로 추출한다. 개체명 인식을 위해서는 NLTK(Natural Language Toolkit)에서 제공하는 파이썬 모듈을 이용하였는데, 이 모듈은 MaxEnt로 알려진 기계학습 알고리즘으로 구현되었다[9,10]. ORG(기관명), PER(인명), LOC(지명), GPE(지정학명) 등과 같은 개체명 또한 링크가 걸릴 가능성이 높기 때문에 이를 인식하여 앵커 후보로 추가하였다. 영-한 CLLD를 처리하는 전반적인 과정은 다음 그림과 같다.

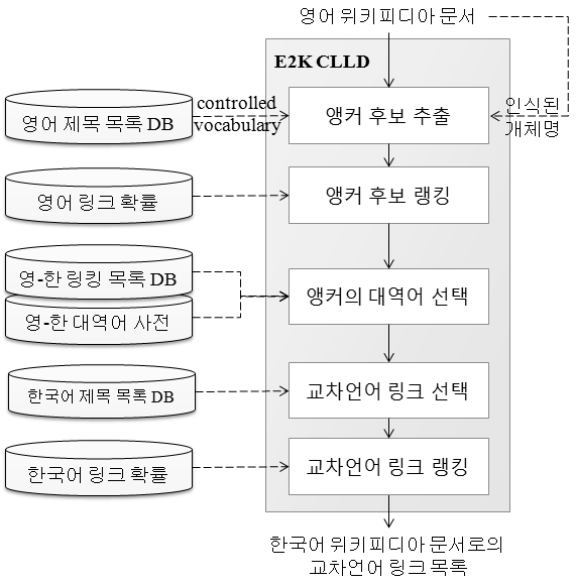

그림 3. 영-한 CLLD 처리 과정

Fig. 3. Processing flow of English-Korean CLLD

3.3 앵커 대역어 선택과 교차언어 링크 설정

영-한 링크 목록은 위키피디아에 존재하는 현존 교차언어 링크이므로 가장 확실한 힌트가 되지만, 그 양이 많지는 않다. 여기에 없는 앵커 후보들은 영-한 대역어 사전을 이용하여 2순위로 대역 후보를 찾게 된다. 영어 위키피디아 문서 내에서 앵커가 출현한 위치(offset)가 동일한 대역어가 여러 개인 경우에는 어구의 길이가 가장 긴 대역어만 선택한다. 긴 어구가 짧은 어구보다 구체적인 의미를 내포하고 있고 이에 해당하는 위키피디아 문서가 존재할 가능성이 크기 때문이다. 선택된 대역어를 이용하여 한국어 위키피디아 문서 집합을 검색하고 그 결과를 한국어링크확률에 따라 랭킹한 후, 상위권 문서를 교차언어 링크로 설정한다. 영-한 CLLD를 위한 구체적인 알고리즘은 표 2와 같다.

4. 실 험

NTCIR-9에서 CLLD를 위한 평가용 데이터 집합을 구축하였는데, 학습용 데이터 집합에는 'Australia', 'Femme fatale', 'Martial arts' 등 3개의 영어 토픽에 대해 한국어 위키피디아 문서로의 링크 정보가 포함되어 있으며, 테스트용 데이터 집합에는 25개의 토픽 문서가 영어 위키피디아에서 랜덤하게 선택되어 포함되었다. 한국어 위키피디아로의 링크를 설정하기 위해 제공된 문서는 총 201,596개이다. 평가를 위해서는 CLLD를 수행한 결과물을 다음과 같은 조건에 맞게 생성해야 한다. 테스트용으로 주어진 각 토픽 문서에 대해 최대 250개의 앵커를 설정할 수 있으며, 각 앵커의 교차언어 링크는 최대 5개까지 허용하였다. 따라서 CLLD가 끝난 하나의 토픽 문서에는 최대 1,250개의 교차언어 링크가 포함될 수 있다.

표 2. 제안하는 영-한 CLLD 알고리즘
Table 2. Proposed E2K CLLD algorithm

```
Algorithm E2K_CLLD (Document d) {
    // 1 단계: 앵커후보 추출
    fp = open(d);
    while (fp != EOF) {
      while (int gram=1; gram <= 10; gram++) {
        fp가 가리키는 단어부터 이후 gram개의 단어
            를 앵커 후보 w로 설정;
        if (w가 영어제목목록DB에 있으면)
            문서 d에서의 위치정보(offset)와 함께 w
            를 앵커로 등록;1)
      }
      fp가 다음 단어를 가리키도록 증가;
    }

    문서 d를 개체명 인식;
    인식된 개체명을 offset과 함께 앵커로 추가 등
       록;2)
    등록된 앵커들을 영어링크확률에 따라 랭킹;

    // 2 단계: 앵커 번역
    영-한 링킹목록 DB를 검색하여 앵커의 한국어 대
       역어 선택(1순위);
    대역어 선택이 되지 않은 앵커를 대상으로 영-한
       대역어사전을 검색하여 모든 대역어 선택(2순
       위);

    // 3 단계: 교차언어 링크 선택
    if (동일한 offset을 가지는 앵커들이 있으면)
       어구의 길이가 가장 긴 앵커만 남기고 나머지
       는 삭제;
    if (앵커의 대역어가 한국어제목목록DB에 존재하
       면)
       해당 앵커에 검색된 한국어 위키피디아 문서를
       링크로 설정;
    if (앵커의 한국어 교차언어 링크가 다수이면)
       한국어링크확률에 따라 랭킹;
}
```

본 연구에서 NTCIR-9의 평가용 데이터 집합을 이용하여 학습하고 실험한 결과는 다음 표와 같다. 'LP'는 앵커 후보 추출을 위해 링크 확률 정보만을 이용한 실험이며, 'NER'은 개체명 인식만을 이용한 것이고, 'LP_NER'은 링크 확률과 개체명 인식을 모두 이용한 실험 결과이다. 'LP_NER'이 MAP(Mean average precision)와 R-Prec (Precision at R-th position)에서 가장 좋은 성능을 보였다.

1) 문서에서 동일한 어구가 여러 번 나타나는 경우에는 가장 먼저 나타난 것만 앵커후보로 선택한다. 일반적으로 링크를 설정할 단어가 웹문서에서 반복하여 출현하는 경우, 가장 먼저 출현한 단어에만 링크를 설정하기 때문이다.

2) 1)과 동일하게 처리

표 3. 영한 CLLD 실험 결과
Table 3. E2K CLLD Experimental results (F2F automatic evaluation)

실험		MAP	r-prec
기존 연구	Tang[4]	0.122	0.208
	Kim[5]	0.337	0.440
	Kang[6]	0.328	0.437
NER		0.176	0.22
LP		0.37	0.442
LP_NER		0.375	0.442

앵커 후보를 결정하는데 있어 링크 확률이 가장 중요한 역할을 하고, 개체명 인식 결과는 약간의 도움이 됨을 알 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 영어 위키피디아 웹문서에서 한국어 위키피디아 웹문서로의 교차언어 링크를 자동으로 설정하는 방법을 제안하였다. 영어와 한국어 위키피디아 문서 집합으로부터 추출할 수 있는 페이지 제목 목록과 이들의 이형태(variants)인 리다이렉트된 페이지 목록, 단어별 링크 확률 등을 전처리과정을 통해 구축하였고, 또한 개체명 인식 결과를 이용하여 링크가 걸릴 앵커 후보를 확장하여 선택하는 방법을 취하였다.

CLLD 결과물은 단일어로 작성된 웹문서 집합으로부터 추출할 수 있는 정보의 한계를 넘어, 다국어 웹문서 집합 간에 존재하는 보다 다양한 정보의 추출이 가능케 하며, 언어의 장벽을 넘어서 다국어 정보 검색을 손쉽게 할 수 있는 기반이 될 것으로 기대된다. 향후에는 제안한 방법을 한-영 CLLD에 적용해 구현하고자 한다.

References

[1] CrossLingual Link Discovery Task, http://ntcir.nii.ac.jp/CrossLink/

[2] CrossLink,Evaluation, http://crosslink.googlecode.com/files/CrosslinkEvaluation-Training-20110715.zip

[3] R. Mihalcea, and A. Csomai, "Wikify! Linking Documents to Encyclopedic Knowledge", *In Proceedings of the CIKM'07*, pp.233-242, November, 2007.

[4] L. X. Tang, D. Cavanagh, A. Trotman, S. Geva, Y. Xu, and L. Sitbon, "Automated Cross-lingual Link Discovery in Wikipedia", *In Proceedings of the 9th NTCIR Workshop Meeting*, pp.512-519, December 2011.

[5] J. Kim and I. Gurevych, "UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery", *In Proceedings of the 9th NTCIR Workshop Meeting*, pp.487-494, December 2011.

[6] I. S. Kang, and R. Marigomen, “English-to-Korean Cross-linking of Wikipedia Articles at KSLP”, *In Proceedings of the 9th NTCIR Workshop Meeting*, pp.481-483, December 2011.
[7] E. Adar, M. Skinner, and D. S. Weld, “Information Arbitrage Across Multi-lingual Wikipedia”, *In Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, pp.94-103, February 2009.
[8] NTCIR-9 Home, http://research.nii.ac.jp/ntcir/ntcir-9/
[9] Natural Language Toolkit, http://nltk.org/
[10] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python, O’reilly, pp.281-284, 2009.

저 자 소 개

강신재(Shin-Jae Kang)
1995년 : 경북대학교 컴퓨터공학과 공학사
1997년 : 포항공과대학교 컴퓨터공학과 공학석사
2002년 : 포항공과대학교 컴퓨터공학과 공학박사
1997년～1998년 : SK Telecom 정보기술연구원 연구원
2007년 : 오스트리아 University of Innsbruck, DERI 연구소 방문교수
2002년～현재 : 대구대학교 컴퓨터·IT공학부 교수

관심분야 : 온톨로지, 시맨틱 웹, 자연어처리
Phone : +82-53-850-6584
E-mail : sjkang@daegu.ac.kr