Concept Unification of Terms in Different Language for IR

고려대학교
컴퓨터 정보통신 대학원
미디어 공학과

전희원
2006.09.02

Opening

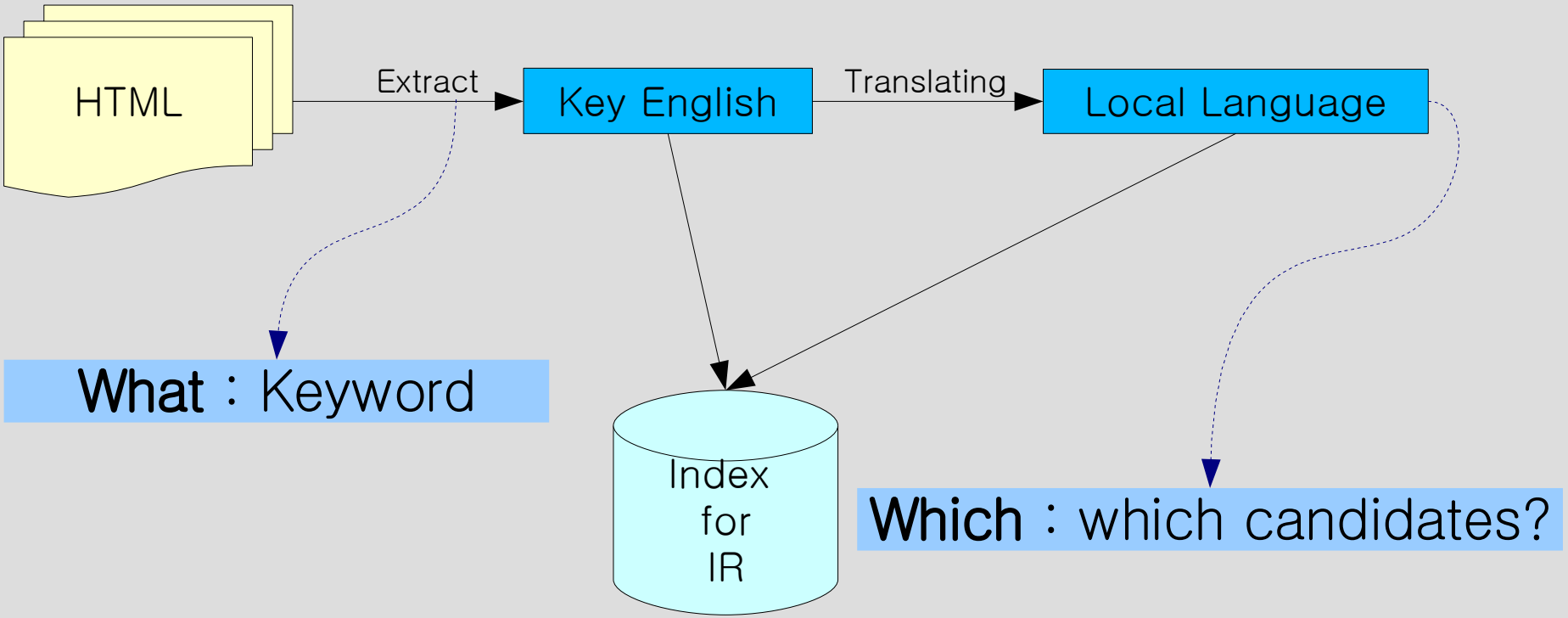
- 외국어를 한국어로 발음되는 단어로 씀으로써 검색 성능의 저하를 가져온다 .(eg. 디지털 , 디지틀 , 디지탈)
- 약어 및 복합어의 관계 (eg. 세계무역기구 (WTO), 서울대 (Seoul National University) 등)

To slove this problem

- ‘Digital’ 같은 단어의 로컬 언어로의 음성표기한 단어를 동일한 색인으로 구성하는 맵핑 방법론 <- ‘back transliteration’

- 하지만 약어 및 복합어의 관계 때문에 색인어의 의미적인 해석에 기반한 맵핑 방법론의 대두가 필요하다 .

Concept Unification



Assumption

- 한 페이지에서 영문 색인어와 그에 대응하는 한글 색인어가 나올 가능성이 많다 .
- 따라서 이 논문에서는 검색엔진으로 검색된 결과물을 이용해서 한글 색인어 후보들을 추출해 낸다 .

What Keyword?

- “ ”, ( ) 와 같이 중요한 의미에 표시되는 문장 부호와 함께 쓰인 단어를 대상으로 한다 .

Which candidates?(1)

- 최대 후보의 길이를 정해놓고 그 범위 안의 모든 n-gram 후보를 추출한다 .

Which candidates?(2)

- 카이제곱 검정 (Chi-square) 방법을 이용한 통계적 방법으로는 정확도가 떨어진다 .
- 그래서 이 논문에서는 '의미론적 (sementic)', ' 음성론적 (phonetic)' 인 방법론을 함께 사용해서 후보자를 추려내고자 했다 .
- eg. 클론의 습격 (Attack of The Clones)

Which candidates?(3) Statistical model

- 기존의 논문에서는 단어가 함께 출현하는 **빈도수**와 **두 단어의 문서 안에서의 거리**에 기반한 후보 랭킹 산정 방식을 채용했다 .

- 이 논문에서는 위의 사항과 추가해서 **후보 단어의 길이 정보**도 후보책정에 영향을 주는 요소로 넣었다 .

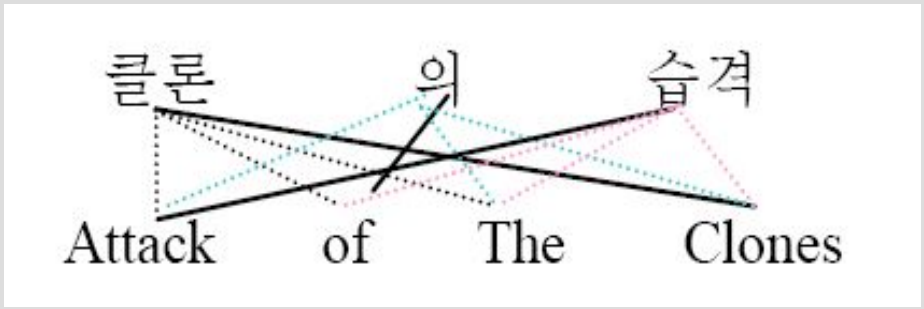
$$w_{FL}(q,c_i) = \alpha \times \frac{len(c_i)}{\max_{len}} + (1-\alpha) \times \frac{\sum_k \frac{1}{d_k(q,c_i)}}{\max_{Freq-len}}$$

Which candidates?(4) Phonetic and Semantic model

- SSP(Score of sementic and phoneme) 를 정의한다 .

$$SSP=\text{argmax}\sum_{i=1}^{n}\omega(kw_i, ew_{\pi(i)})$$

- SSP 에 쓰이는 w 값은 semantic weight 과 phoneti weight 의 합이다 .



Which candidates?(5) Phonetic Weight

- 음성 가중치 (Phonetic weight) 는 영어와 해당 로컬 언어사이의 음역 확률이다 .(HMM 관련 내용 참고)

$$\omega_{phoneme}(EW,CW)=\omega_{phoneme}(e_1,\ldots,e_m,c_1,\ldots,c_k)$$

$$=\omega_{phoneme}(g_1,\ldots,g_n,c_1,\ldots,c_k)=\frac{1}{n}\sum_j \log P(g_j \mid g_{j-1})P(c_j \mid g_j)$$

$$\omega_{phoneme}(E,C)=\frac{1}{n}\sum_j \log P(g_j g_{j+1} \mid g_{j-1} g_j)P(c_j c_{j+1} \mid g_j g_{j+1})$$

Which candidates?(6) Sementic Weight

- 2 개국어 사전을 이용한 단순 해석에 이은 후보단어와의 매핑을 조사한다 .

$$w_{semanteme}(E,C) = argmax \frac{No.\ of\ overlapping\ units}{total\ No.\ of\ \ units}$$

- eg“ 인하대 (Inha University)” -> 0.33

Which candidates?(7)
Get final translation

- Phonetic, Sementic weight 값의 normalization
- SSP 값이 어떤 임계값보다 작다면 확률모델 값이 가장 큰것을 최종 후보로 삼는다 .
- 그렇지 않으면 확률 모델값에 기반한 후보 확보후 SSP 값이 많이 뛰는 곳에 가상의 라인을 두고 그 이상의 후보만을 최종 후보로 책정한다 .

Conclusion

- 14.9 % 의 정확도 향상

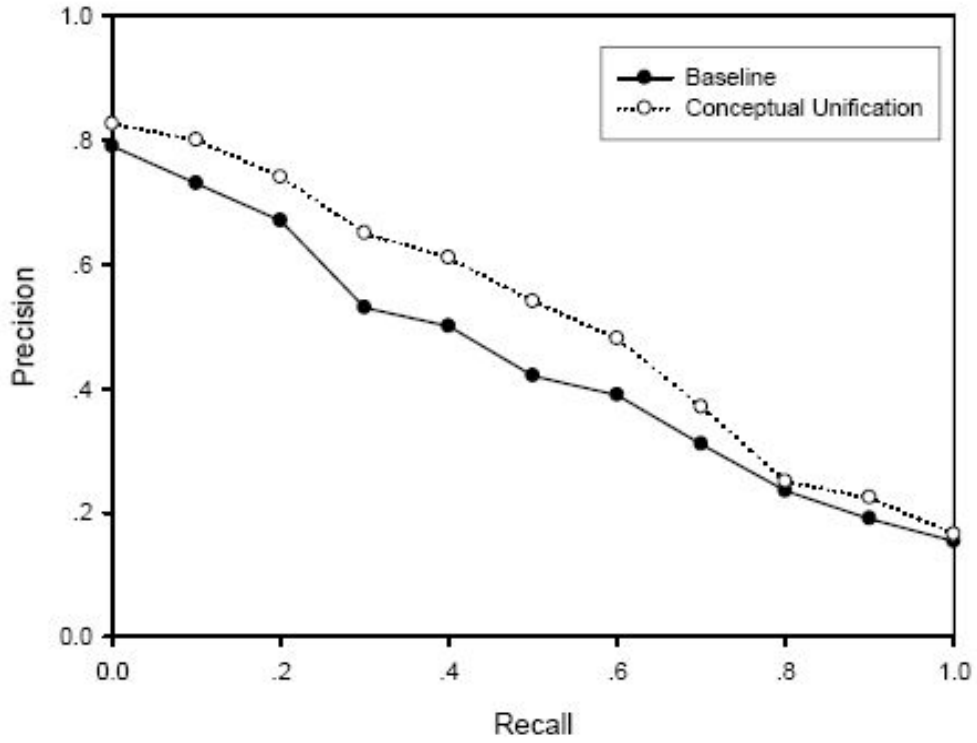


Figure 3. Korean Monolingual IR