WordTree Foundation에서, 우리는 이 탐구에서 특별한 흥미를 가지는 Mormon의 책을 포함한, 다양한 책들에서 유사성들과 영향들을 발견하는 방법들을 탐구했다. 과거에는, 우리는 여러 책들 사이의 유사점을 채점하는 흥미로운 방법(제한적이기는 하지만)이 될 가공되지 않은 비조건부 확률들을 보여줘왔다. 우리의 ‘A Comparison of The Book of Mormon and The Late War’의 온라인 발행은 본래의 연구의 결과였다.

나는 오늘 우리가 매우 빠른 Language Model generator인 KenLM을 사용한, Julia 포르그래밍 언어에서 작성했던 ‘previoius n-gram library’로 변형할 수 있다는 것을 생각하기 시작했다. 대신에, 결국 나는 LM과 데이터 포맷을 배우는 데 대부분의 날을 소비하였다. 나는 일반인들에게 내가 가진 이 지식과 어떤 이해들을 널리 알리고 싶다고 생각했다.

무엇보다도, LM은 WordTree Foundation에서 우리가 사용해왔던 N-gram 확률보다 더욱 괜찮다. 특히, LM은 텍스트에서 단어 발생의 확률들을 더욱 더 정확하게 평가하는 조건부 확률의 개념을 가진다.

내가 Book of Mormon과 같은 text를 가지고 그것을 KenLM’s “lmplz” 프로세서를 통해 전달할 때, 그것은 ARPA파일이라고 불리는 파일을 만들어 낸다. ARPA 파일의 구성 방식은 아래와 같다.

< 그림 >

이 음수들이 의미하는 것들이 처음에는 나에게 확실하지 않았다. 즉, “ ‘the’라는 단어를 볼 우연이 0.018023이다.”라고 말하는 대신에 ARPA 구성 방식은 확률을 얻기 위하여 “-1.7441652”를 사용한다. 이 확률의 묘사 형태는 확률들은 매우 작은 값이기 때문에 보통 더 작다. (예를 들어, “0.000000478630092”라고 적는 것 보다 우리는 “-6.32”라고 적을 수 있다.) 게다가 log 값으로써, 그들은 자연적으로 더 현저한 규모의 정렬을 나타낸다.

ARPA 파일에서, unigrams(1-grams)은 비조건부 확률이다. 반면에, DRW00002e3c1840또는 더 큰 값을 가지는 N-grams은 조건부 확률이다. 즉, 우리가 1-grams 부분에서 “-1.7441652 the”을 봤을 때, 그것은 간단하게 “the”라는 글자를 언어로 보여주는 로그 확률이다. 즉, 수학적 표기법으로 ‘P’ 또는 “ ‘the’의 확률”을 뜻한다. 하지만, 2-grams 부분에서 우리가 “the의 -0.5145896”을 볼 때, 이것은 “the”가 “of”뒤에 나올 확률을 뜻하는 것이다. 즉, 수학적 표기법으로 ‘P(the | of) 또는 “ ‘of’가 주어진 상황에서의 ‘the’의 확률”을 뜻한다.

LM에는 세가지 특별한 단어들이 존재한다. : <s> , </s> , <unk>

<s>는 문장의 시작을 의미하고, </s>는 문장의 끝을 의미한다. 이것은 문장의 시작과 끝의 개념을 모르는 우리의 본래 학문과는 다르다. 우리는 단지 문장의 경계들을 무시해왔다. 그리고 <unk>는 “unknown”을 뜻하는 특별한 단어이며, LM에서 Model안에 들어 있지 않는 단어들의 확률을 나타내는 단어로 사용된다.

조건부 확률을 사용함으로써, 구에 관한 의문들은 그 구의 빈도와 희소성을 더욱 정확하게 반영하는 확률을 결과로 낳을 것이다. 궁극적으로, 이것은 책들을 비교할 때와 텍스트들 사이에서 명백하지 않은 관계들 또는 영향들을 알아낼 때 더 정확한 알고리즘의 채점을 이끌 것이다.