김일성종합대학학보
정보과학

주체109(2020)년
제66권 제2호

감독 및 무감독방식의 적응본문선택에 대한 대비분석

김철, 리혁철

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《나라의 과학기술을 세계적수준에 올려세우자면 발전된 과학기술을 받아들이는것과 함께 새로운 과학기술분야를 개척하고 그 성과를 인민경제에 적극 받아들여야 합니다.》(《**김정일**선집》 증보판 제11권 138～139페지)

선행연구[3, 4]에서는 주어진 화제의 대표적인 핵심본문에 기초하여 대규모본문코퍼스로부터 적응본문을 추출하는 방법들을 제안하였지만 실현가능성문제로 하여 일정한 제한이 있다.

론문에서는 자동표기를 핵심본문으로 리용하는 무감독방식의 적응본문선택에 기초한 조선어음성언어모형적응방법을 제안하였다.

1. 문 제 설 정

일반적으로 음성인식체계의 *n*-그람언어모형들은 화제들에 해당한 다량의 본문모임에 기초하여 학습되지만 주어진 특정의 화제에 해당한 음성을 인식하는데는 적합하지 않다. 이러한 문제를 해결하기 위하여 화제적응에 의하여 해당 화제에 부합되도록 기준언어모형의 *n*-그람확률들을 재추정하는 적응방법들이 널리 리용되고있다.

표준적인 화제적응방법[1]에서는 *n*-그람재추정에 필요한 특정화제의 자료를 추출하기 위하여 웨브자료기지를 코퍼스로 리용한다. 웨브자료기지에 기초한 적응처리에서는 목적화제를 대표하는 주어진 본문(핵심본문)으로부터 질문을 추출하고 웨브탐색엔진에서 이 질문들을 리용하여 웨브페지들을 추출한 다음 추출된 적응자료와 배경학습자료를 통합하여 적응모형을 구축한다.[2]

핵심본문은 수동적인 감독방식[3]과 인식결과를 리용하는 무감독방식[4]으로 작성된다. 그러나 이러한 본문을 수동적으로 생성하는데 요구되는 품이 많이 들므로 감독적응의 실현가능성은 핵심본문의 크기에 관계된다. 그러므로 이러한 공정의 자동화는 실천응용에서 특정화제의 언어모형개발에 필요한 비용과 노력을 크게 절약할수 있다.

론문에서는 무감독방식의 핵심본문추출에서 인식오유들이 음성인식정확도개선에 주는 영향과 핵심본문크기에 대한 의존성을 감독방식과 대비적으로 분석평가하였다.

2. 언어모형적응단계

적응공정을 다음과 같이 3단계로 구성하였다.

① 목적화제를 대표하는 주어진 핵심본문으로부터 질문들을 추출한다.

② 자료기지검색엔진에서 질문들을 리용하여 적응코퍼스구축에 필요한 본문자료들을 추출한다.

③ 기준언어모형학습에 리용된 배경본문모임과 적응코퍼스에 의하여 적응언어모형을 학습시킨다.

1) 핵심본문으로부터의 질문추출

질문추출방법의 원리[3]에서는 기준언어모형의 어떠한 n-그람들이 주어진 핵심본문에 대하여 충분히 모형화되였는가를 결정하고 n-그람들을 직접 질문으로 리용하였다.

핵심본문 T가 주어졌을 때 핵심본문에 대한 우도가 기준언어모형을 리용한 우도보다 더 큰 적응모형을 구하는것이므로 다음과 같이 형식화할수 있다.

$$P_A(T) > P_B(T) \tag{1}$$

여기서 P_A와 P_B는 각각 적응모형과 기준모형의 확률분포이다.

식 (1)은 우도를 T의 매 n-그람(h, w)에 관하여 다음과 같이 분해할수 있다.

$$P_A(w|h) > P_B(w|h), \forall(h, w) \tag{2}$$

여기서 w는 단어이고 h는 단어리력이다.

결국 질문추출문제는 T의 어느 n-그람들이 식 (2)를 만족시키는데 가장 적합한가를 찾는데 귀착된다.

실천에서는 기준언어모형학습과정에 관측되지 않은 핵심본문의 유일한 3-그람들 즉 확률이 back-off로 계산되는 3-그람들을 임의로 질문으로 선택할수 있다. 그러나 핵심본문 T의 크기에 따라 n-그람수가 증가되는것으로 하여 추출공정시간이 대단히 길어지게 된다. 그러므로 n-그람모임을 축소하기 위하여 최종적으로 시작기호와 끝기호, 정지단어를 포함하는 모든 n-그람들은 무시한다.

실험에서는 이러한 질문추출방법으로 주어진 핵심본문에 대하여 수백개의 질문들을 생성하였다.

2) 적응자료의 추출과 적응언어모형학습

목적화제의 적응자료들을 추출하기 위하여 검색엔진에 질문들을 제기한다. 이때 추출된 자료들을 적응코퍼스에 합치기 전에 정규화를 진행한다. 다음 화제적응모형을 학습시킨다. 구체적으로 배경코퍼스와 적응코퍼스를 리용하여 혼합언어모형들을 각각 학습시킨다. 다음 이 모형들을 선형보간하여 그 결합이 핵심본문에 대하여 분기수를 최소화하도록 하고 최종모형을 가지자르기한다.

3. 평 가 실 험

핵심본문은 특정화제의 질문들을 추출하는데 리용되며 특정화제의 n-그람확률들과 배경학습본문들로부터 구한 확률들을 선형결합할 때 적응자료의 중요도(보간무게)를 결정하는데 요구된다.

기준 3-그람언어모형(LM0)은 로동신문 3년분(2010－2012)코퍼스에 기초하여 학습하였다. 체육부문에 대한 적응코퍼스구축을 위하여 감독 및 무감독방식으로 체육신문 1개 호분(참조자료)에 해당한 핵심본문을 작성하고 그것으로부터 추출된 질문들을 리용하여 15개 분야의 대규모본문코퍼스로부터 적응자료(개발자료)들을 선택하였다. 검사자료로서 텔레비죤체육통로록화물 3개(대략 0.6h분)를 준비하였다.

실험에서는 질문추출에 미치는 핵심본문의 영향과 핵심본문크기에 대한 의존성을 연구분석하였다.

1) 질문추출에 미치는 핵심본문의 영향분석

무감독방식의 핵심본문추출에서 인식오유가 질문추출에 미치는 영향을 조사하기 위하여 참조자료와 음성인식자동표기를 리용하였을 때의 결과들을 비교하였다.

질문추출에 각이한 핵심본문들을 리용하였을 때의 모형분기수를 다음의 표에 보여주었다.

표. 질문추출에 각이한 핵심본문들을 리용하였을 때의 모형분기수

핵심본문의 류형	개발자료/개	검사자료/개
기준모형	29	31
참조자료(감독방식)	17	24
인식결과(무감독방식)	22	25

참조자료를 핵심본문으로 리용할 때(감독방식) 개발자료에 대하여 크게 개선(41%)되였다.

그러나 감독방식에서는 핵심본문이 학습본문과 류사해지는 인위적인 경향성이 있으므로 검사자료에 대하여 분기수의 개선(22.5%)이 상대적으로 적다. 한편 음성인식자동표기결과를 핵심본문으로 리용할 때에는(무감독방식) 개발자료와 검사자료사이의 분기수개선(24.1%, 19.3%)의 차이가 크게 나타나지 않았다.

2) 핵심본문의 크기에 대한 의존성분석

핵심본문의 크기에 따라 적응결과는 가변적이라고 생각할수 있으므로 여기서는 핵심본문의 크기가 화제적응에 주는 영향에 대하여 론의한다.

핵심본문의 크기에 따르는 모형분기수그라프를 다음의 그림에 보여주었다.

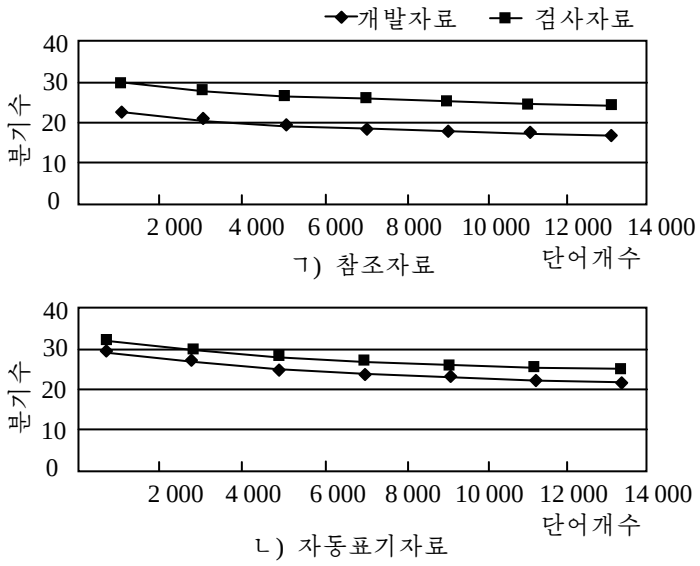

그림. 핵심본문의 크기에 따르는 모형분기수그라프

그림에서 곡선들이 류사한 경향성을 가지는것은 핵심본문의 크기가 언어모형적응에 강한 작용을 하지 않는다는것을 의미한다.

맺 는 말

무감독방식에서 자동표기에 내제된 인식오유들이 모형적응의 성능을 크게 저하시키지 않는다는것을 실험적으로 확증하고 조선어음성인식에 적용하였다.

참 고 문 헌

[1] I. Bulyko et al.; ACM Trans. On Speech and Language Processing, 5, 1, 1, 2007.
[2] T. Hain et al.; IEEE Trans. On Audio, Speech, and Language Processing, 20, 486, 2012.
[3] V. Wan, T. Hain; Proc. of ICASSP, 1520, 2006.
[4] A. Ito et al.; Proc. of ICALIP, 1412, 2008.

주체109(2020)년 2월 5일 원고접수

Comparative Analysis on Supervised and Unsupervised Adaptation Text Selection

Kim Chol, Ri Hyok Chol

In topic adaptation based on supervised and unsupervised adaptation text selection, we verify that unsupervised method is relatively stable in performance change over supervised one and that the size of seed text has little effects on improvements of perplexity.

Keywords: speech recognition, language model, topic adaptation