Joint Semantic Synthesis and Morphological Analysis of the Derived Word

석사과정 송지수

0. 무슨 논문?

- ACL 2017에 published 된 논문
- 영어/독일어 단어를 형태소로 Tokenize 한 다음, 형태소의 의미를 종합하여 원래 단어의 vector를 추정하는 결합확률 기반 모델

1. Introduction

Questionably → Question + able + ly

- 결합 확률 모델 of:
  - 단어 w를 구성 성분(형태소) 으로 structural decomposition
  - 분해된 형태소의 vector로부터 단어 w의 embedding vector 추측

unachievability

Segment

un achiev abli ity

unachievability

Restore

unachieveableity

Segment

un achieve able ity

unachievability

↓ Restore

unachieveableity

↓ Segment

un achieve able ity

} ***Canonicalization!***
(정규화)

핵심 개념 1: Canonicalization

- 같은 뜻의 segment라도, 어떤 단어에 붙어있었느냐에 따라 형태가 달라짐
- Canonicalization을 통해 각 형태소의 원래 형태를 복원하여 이용할 수 있음

unachievability

(Noun) 성취하지 못할 가능성/척도.

achievement

(Noun) 성취

underachiever

(Noun) (기대에 비해) 성취하지 못한 사람

achieves

(Third-person Verb) 성취하다

unachievability

achievement

underachiever

achieves

un achiev abil ity

achieve ment

under achiev er

achieve s

un achiev abil ity

achieve ment

under achiev er

achieve s

Same?

un	achiev	abil ity
	achieve	ment
under	achiev	er
	achieve	s

unachieveableleity

achievement

underachieveer

achieves

un achieve able ity

achieve ment

under achieve er

achieve s

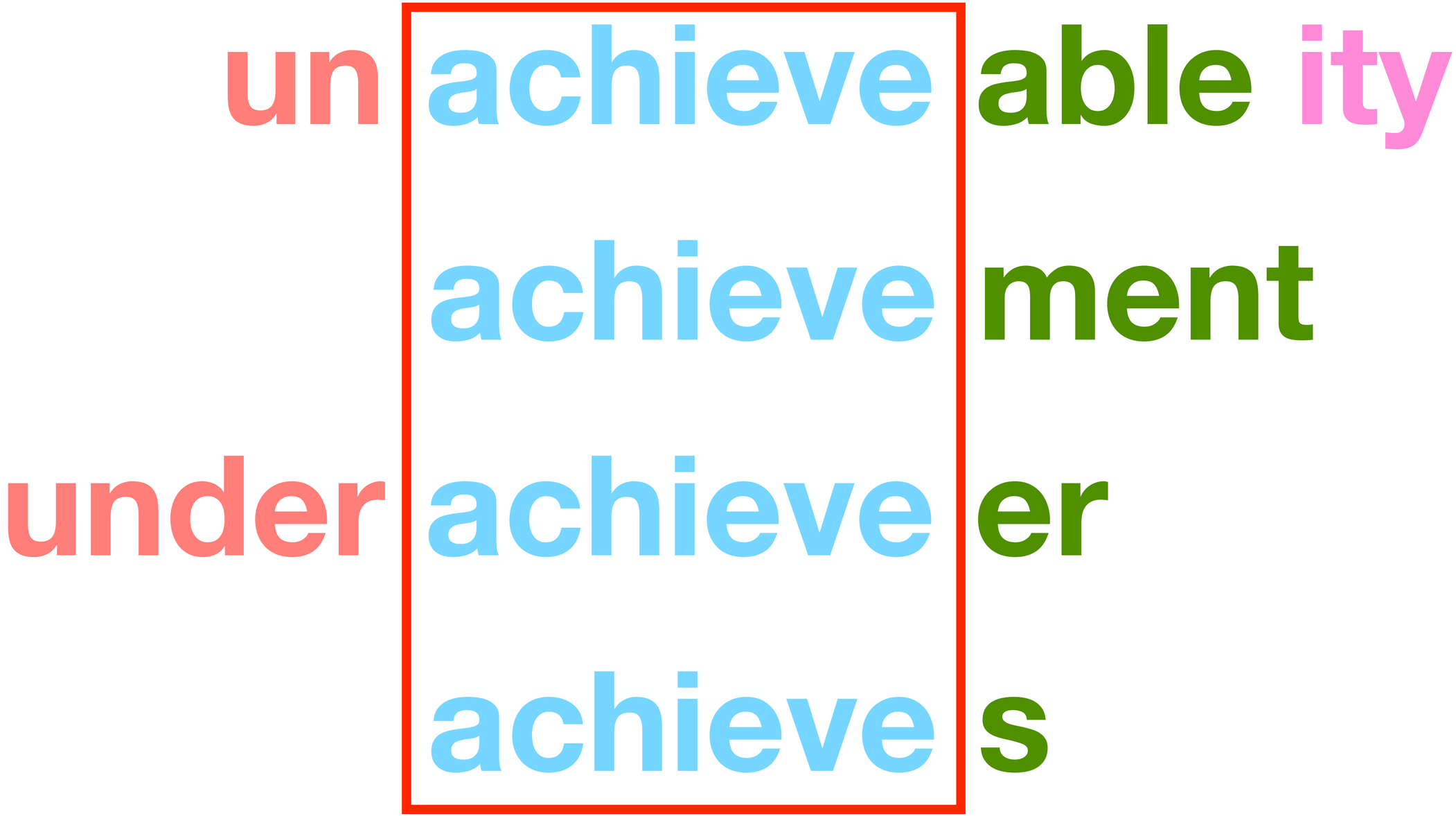
un achieve able ity

achieve ment

under achieve er

achieve s

Segmentations are *canonicalized!*



2. Derivational Morphology

Morphology

명사 : (언어) 형태론

Inflectional Morphology

- 어형과 어미의 변화로서 시제 혹은 단/복수 등
  문장 속에서 다른 단어와의 관계를 나타냄

- run -> runs, running, ran, …

Derivational Morphology

- 어두, 어미의 변화가 단어의 의미 자체를 변화시킴

- content > contented > discontented > discontentedness
  (명. 내용) (형. 만족해 하는) (형. 불만스러워 하는) (명. 불만을 품음)

<u>영어는 Derivational Morphology의 관점에서 매우 Complex한 언어!</u>

핵심 개념 2: Productivity

- Affix (Prefix, Suffix) 가 Productive 하다 =
  더욱 많은 단어에 붙을 수 있으며 자신의 역할을 수행할 수 있다

  - -ness : 거의 모든 형용사에 붙어 명사화 수행
    (red > redness, happy > happiness, …)

  - -th : 붙을 수 있는 단어가 비교적 제한적
    (stealth = steal + th (O), cheapth = cheap + th (X))

- Productivity를 정의하는 것은 쉽지 않다

핵심 개념 3: Semantic Coherence

- 어떤 단어가 Semantic Coherent 하다 =
  각 morpheme들의 의미를 합한 것이 합성된 단어의 본래 의미와 같다
    - Questionably = Question + able + ly (Semantic Coherent!)
    - Blackmail = Black + mail (Nope.)
- 같은 Morpheme이라도 Semantic Labeling에 따라 의미가 다르다
    - prefix “post” =/= stem “post”

3. A Joint Model

Hypercuriosity

Hyper + curious+ ity

- 정답

Hyper + cure + ous +ity

- by vice > vicious

어떻게 Segmentation을 할까 에 대한 문제는
단어 출현의 확률 분포에서 힌트를 찾을 수 있다!

(curious 와 cure 의 co-occurence)

$$p(v, s, \zeta, u \mid w)$$

단어 (string 형태)
unachievability

$$p(v, s, \ell, u \mid w)$$

Underlying Form

unachieveableity

$$p(v, s, \ell, u \mid w)$$

Semantic Labeling

{prefix, stem, suffix, suffix}

$$p(v, s, l, u \mid w)$$

Canonical Segmentation

un achieve able ity

$$p(v, s, \zeta, u \mid w)$$

Word Embedding of unachievability

$$p(s, l, u \mid w) \propto e^{(\boldsymbol{f}(s,l,u)^\top \boldsymbol{\eta} + \boldsymbol{g}(u,w)^\top \boldsymbol{\omega})}$$

$$p(s, l, u \mid w) \propto e^{(} \boxed{\boldsymbol{f}(s,l,u)^\top \boldsymbol{\eta}} + \boldsymbol{g}(u,w)^\top \boldsymbol{\omega})$$

Segmentation Factor

s = un achieve able ity, u = unachieveableity

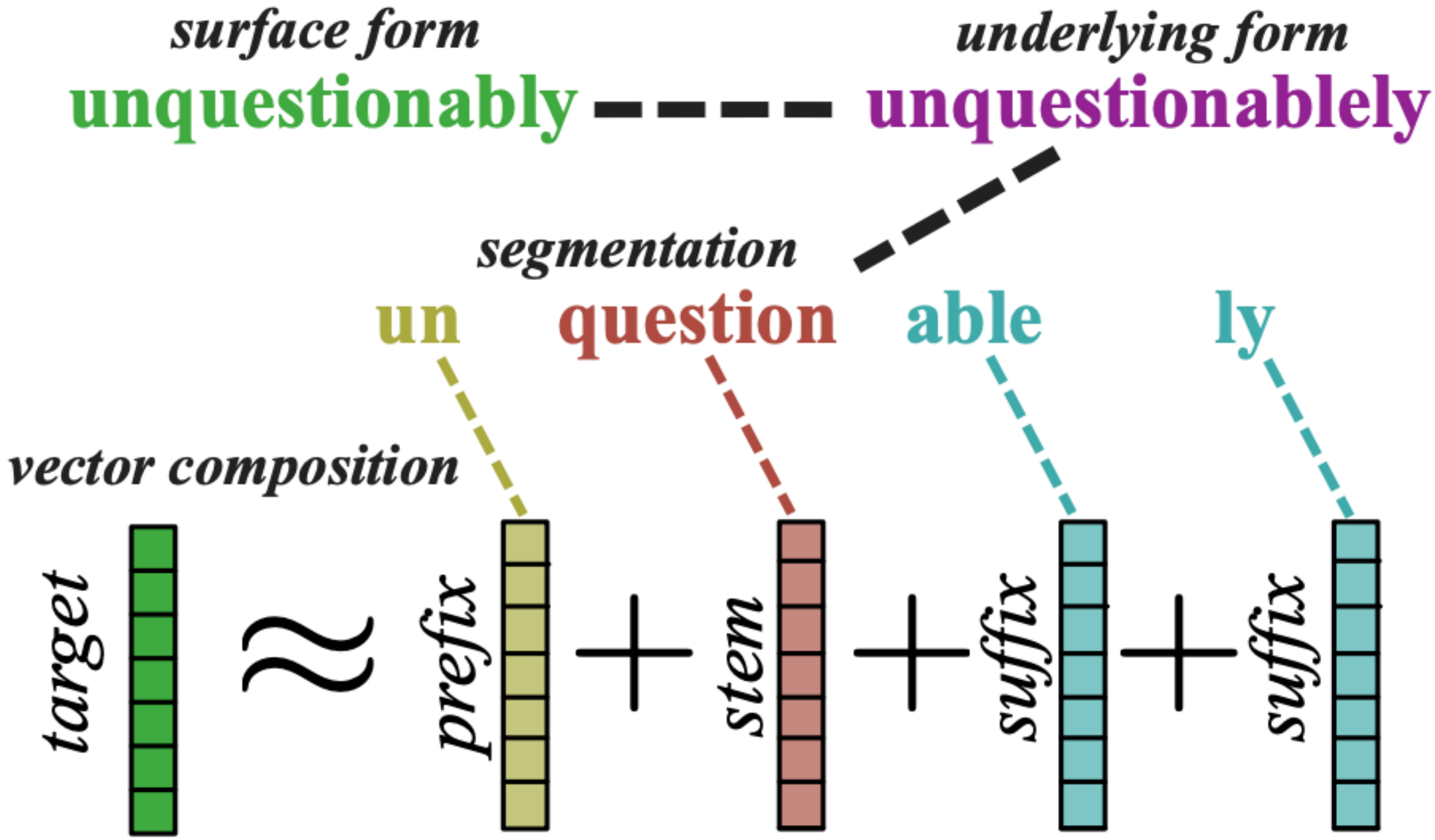
- **Underlying Form의 Segment를 Scoring**
  (Segmentation이 잘 되었는지)
- **Semantic Labeling도 고려**
  (prefix:post =/= stem:post)

$$p(s, l, u \mid w) \propto e^{(\boldsymbol{f}(s,l,u)^\top \boldsymbol{\eta} + \boldsymbol{g}(u,w)^\top \boldsymbol{\omega})}$$

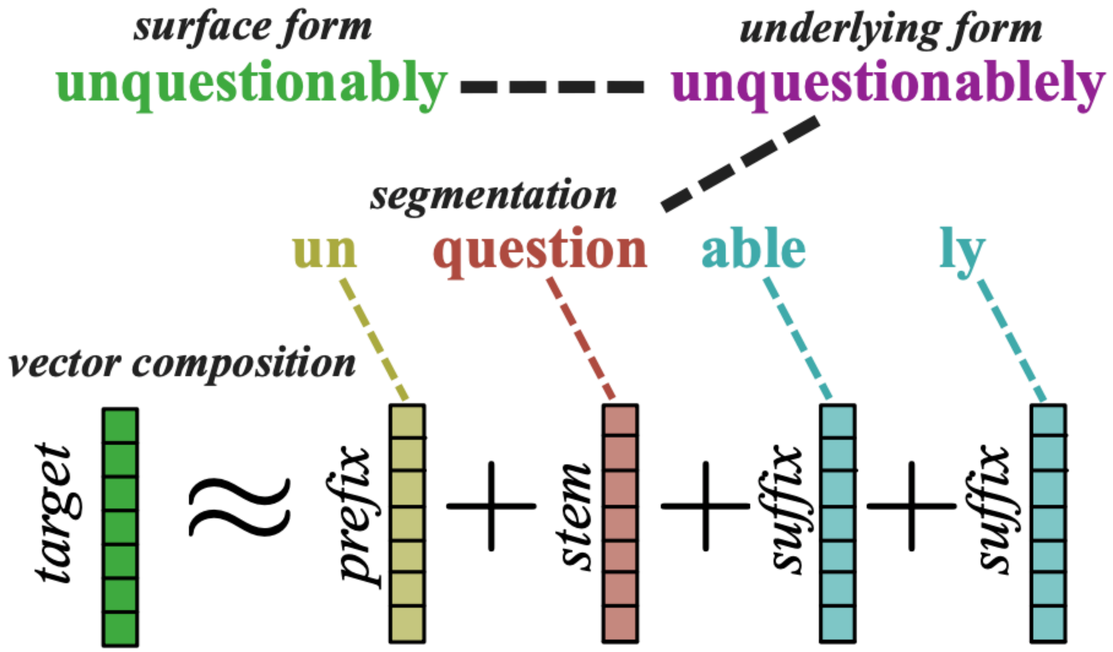
Transduction Factor

u = unachieveableity, w = unachievability

- **Underlying Form 과 Surface Form이 쌍이 맞는지 (U와 W가 좋은 pair일 때 Score 높게)**


surface form
underlying form
unquestionably
unquestionablely
segmentation
un
question
able
ly
vector composition
target
≈
prefix
+
stem
+
suffix
+
suffix

content > contented > discontented > discontentedness 의 경우에는? 먹힐까?



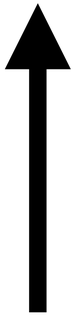
$$p(v, s, l, u \mid w) = p(v \mid s) \cdot p(s, l, u \mid w)$$

$$p(v \mid s) \propto e\hat{}\left(-\frac{1}{2\sigma^2} ||v - C_\beta(s, l)||_2^2 \right)$$

$$p(v, s, l, u \mid w) = p(v \mid s) \cdot p(s, l, u \mid w)$$

$$p(v \mid s) \propto e\hat{}\left(-\frac{1}{2\sigma^2} ||v - C_\beta(s, l)||_2^2 \right)$$

가우시안 분포



형태소 벡터들의 합성 함수

model	composition function		
stem	c	$=$	$\sum_{i=1}^{N} \mathbb{1}_{l_i=\text{stem}} m_{s_i}^{l_i}$
mult	c	$=$	$\bigodot_{i=1}^{N} m_{s_i}^{l_i}$
add	c	$=$	$\sum_{i=1}^{N} m_{s_i}^{l_i}$
wadd	c	$=$	$\sum_{i=1}^{N} \alpha_i m_{s_i}^{l_i}$
fulladd	c	$=$	$\sum_{i=1}^{N} U_i m_{s_i}^{l_i}$
LDS	h_i	$=$	$X h_{i-1} + U m_{s_i}^{l_i}$
RNN	h_i	$=$	$\tanh(X h_{i-1} + U m_{s_i}^{l_i})$

$$p(v, s, l, u \mid w) = p(v \mid s) \cdot p(s, l, u \mid w)$$

$$p(v \mid s) \propto e^{\wedge}\left(-\frac{1}{2\sigma^2} ||v - C_{\beta}(s, l)||_2^2 \right)$$

이런 방법의 이점:
OOV(Out of Vocabulary, 사전에 없는 단어) 의 Semantic Vector를 형태소 Vector 이용해 근사 가능!

4.1 Inference by Importance Sampling

어떤 string w가 주어졌을 때, w의 올바른 짝인
Underlying Form u, Segmentations s, Semantic Labeling l
을 찾아야 함

그러나 w로부터 나올 수 있는 모든 u와 s, l를 고려하는 것은
경우의 수가 엄청나게 많아지기 때문에 힘들다

어떤 string w가 주어졌을 때, w의 올바른 짝인
Underlying Form u, Segmentations s, Semantic Labeling l
을 찾아야 함

그러나 w로부터 나올 수 있는 모든 u와 s, l를 고려하는 것은
경우의 수가 엄청나게 많아지기 때문에 힘들다

Importance Sampling 을 사용하여
확률적으로 높은 u와 s, l만 고려하자!

$$\boldsymbol{h}(l, s, u) = \boldsymbol{f}(s, l, u) + \boldsymbol{g}(u, w).$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log Z &= \mathop{\mathbb{E}}_{(l,s,u)\sim p} [\boldsymbol{h}(l, s, u)] \\
&= \sum_{l,s,u} p(l, s, u) \boldsymbol{h}(l, s, u) \\
&= \sum_{l,s,u} \frac{q(l, s, u)}{q(l, s, u)} p(l, s, u) \boldsymbol{h}(l, s, u) \\
&= \mathop{\mathbb{E}}_{(l,s,u)\sim q} \left[\frac{p(l, s, u)}{q(l, s, u)} \boldsymbol{h}(l, s, u) \right],
\end{aligned}$$

p: 전체 l, s, u의 분포

q : Importance Sampling을 위해 뽑아낸 Sample Distribution

$$\frac{1}{\sum_{i=1}^{M} w^{(i)}} \sum_{i=1}^{M} w^{(i)} \boldsymbol{h}(l^{(i)}, s^{(i)}, u^{(i)}),$$

$$w^{(i)} = \frac{\bar{p}(l^{(i)}, s^{(i)}, u^{(i)})}{q(l^{(i)}, s^{(i)}, u^{(i)})}.$$

Sampled Importance Distribution q로부터 l, s, u 의 weight를 구함

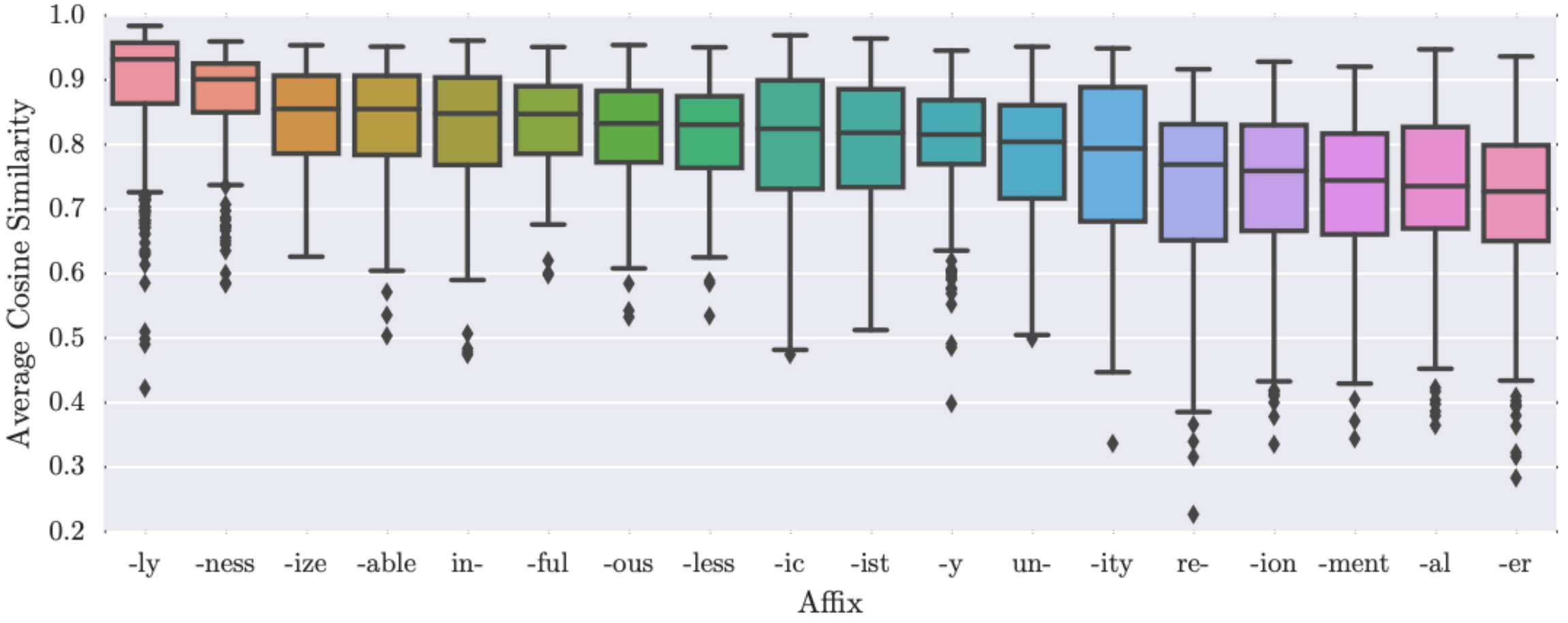
4.2 Learning

$$\nabla_{\boldsymbol{\theta}} \log p(v, s, l, u \mid w) = \boldsymbol{f}(s, l, u)^\top + \boldsymbol{g}(u, w)^\top$$
$$- \frac{1}{\sigma^2}(v - \mathcal{C}_{\boldsymbol{\beta}}(s, l))\nabla_{\boldsymbol{\theta}}\mathcal{C}_{\boldsymbol{\beta}}(s, l)$$
$$- \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}(w), \qquad (9)$$

앞서 소개한 Factor들을
log-likelihood optimization을 이용하여 학습

5. Experiments and Results

		EN						DE	
		BOW2		BOW5		DEPs		SG	
		dev	test	dev	test	dev	test	dev	test
oracle	stem	.403	.402	.374	.376	.422	.422	.400	.405
	add	.635	.635	.541	.542	.787	.785	.712	.711
	LDS	.660	.660	.566	.568	.806	.804	**.717**	**.718**
	RNN	.660	.660	.565	.567	**.807**	**.806**	.707	.712
joint	stem	.399	.400	.371	.372	.411	.412	.394	.398
	add	.625	.625	.524	.525	.782	.781	.705	.704
	LDS	.648	.648	.547	.547	.799	.797	**.712**	**.711**
	RNN	.649	.647	.547	.546	**.801**	**.799**	.706	.708
char	GRU	.586	.585	.452	.452	.769	.768	.675	.667
	LSTM	.586	.586	.455	.455	.768	.767	.677	.666


Average Cosine Similarity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
-ly
-ness
-ize
-able
in-
-ful
-ous
-less
-ic
-ist
-y
un-
-ity
re-
-ion
-ment
-al
-er
Affix

감사합니다!