Guideline for English Lexical Normalisation Shared Task

Disclaimer: the tweets are randomly collected from public Streaming API. They may contain offensive messages.

This task focuses on context-sensitive lexical normalisation of English Twitter messages. Non-standard words (NSW) include spelling errors (*commitee* for *committee*), informal abbreviations (*tmrw* for *tomorrow*) , phonetic substitutions (*4eva* for *forever*).

Guideline:
Read tweets quickly and judge whether the tweet contains NSW. If a tweet doesn't contain any NSW, you then skip it; If a tweet contains NSW, please provide your corrections in the 2nd column as shown in the following example. For your convenience, your corrections can be case insensitive and all Out-of-Vocabulary (OOV) tokens are pre-marked with "#" in the spreadsheet.

Index:53	470862524326638000		
1	My	My	
2	frnd	friend	#
3	gave	gave	
4	me	me	
5	a	a	
6	white	white	
7	elephant	elephant	
8	and	and	
9	collcted	collected	#
10	it	it	
11	cus	because	#
12	I	I	
13	was	was	
14	tryin	trying	#
15	2	to	

Rules for annotations:

1. Non-standard words (NSWs) are normalised to one or more canonical English words relative to the given lexicon.
    1. One-to-one normalisation: *tmrw* is normalised to *tomorrow*
    2. Many-to-one normalisation: *l o v e* is normalised to *love*. In case of many-to-one annotation, provide correction in the first cell and leave the rest cell empty.

50	I	I
51	I	love
52	o	
53	v	
54	e	
55	u	you

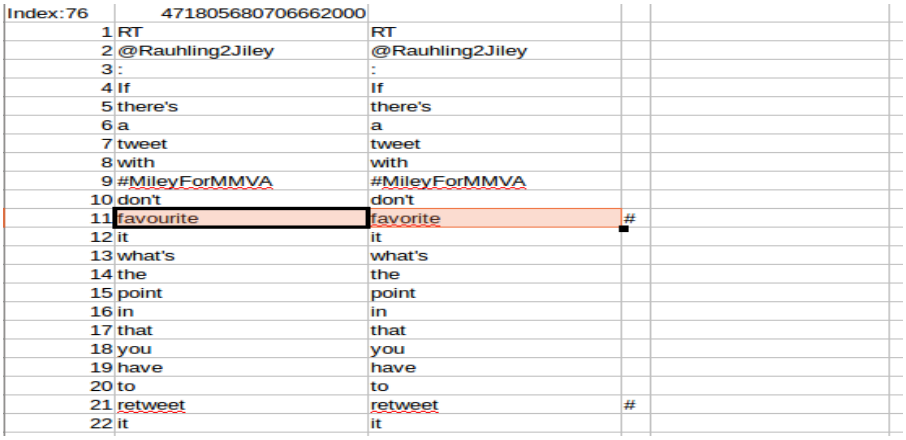
    3. One-to-many normalisation: *cu* is normalised to *see you*
    4. Many-to-many normalisation: *cu tmrw are normalised to see you tomorrow. Note that* syntactic reordering is not required, e.g., *c tmrw u* are normalised to *see tomorrow you*
    5. Additionally, *IBM* is kept untouched as it is in the lexicon and a informal *lol* shall be expanded to *laughing out loud*.
2. Non-standard words may be OOV tokens (e.g., *tmrw* for *tomorrow*). They could also be In-Vocabulary (IV) tokens (e.g., *wit* for *with* in *I will come wit you*).
3. Only alphanumeric tokens (e.g., *2*, *4eva* and *tmrw*) and apostrophes used in contraction forms (e.g., *yoou've* for *you've* and *cant* for *can't*) are considered for normalisation. Tokens

involved with hyphens, single quotes and other contractions are ignored. Domain specific entities are ignored even if they are in non-standard forms, e.g., *#ttyl*, *@nyc*

4. Proper nouns shall be kept untouched even if they are not in the given lexicon, e.g., *Twitter*.

Index:118	471785552245572000		
1	RT	RT	
2	@RichardBarrow	@RichardBarrow	
3	:	:	
4	#Thailand	#Thailand	
5	-	-	
6	RT	RT	
7	@iBBpim	@iBBpim	
8	:	:	
9	Patong	Patong	#
10	Beach	Beach	
11	,	,	
12	Phuket	Phuket	#
13	http://t.co/8uxrUfFLbh	http://t.co/8uxrUfFLbh	

5. Your normalisations shall be in American spelling, e.g., *tokenize* rather than *tokenise*.



Index:76	471805680706662000		
1	RT	RT	
2	@Rauhling2Jiley	@Rauhling2Jiley	
3	:	:	
4	If	If	
5	there's	there's	
6	a	a	
7	tweet	tweet	
8	with	with	
9	#MileyForMMVA	#MileyForMMVA	
10	don't	don't	
11	favourite	favorite	#
12	it	it	
13	what's	what's	
14	the	the	
15	point	point	
16	in	in	
17	that	that	
18	you	you	
19	have	have	
20	to	to	
21	retweet	retweet	#
22	it	it	

6. Exclamative words such as *haha* and *wooooooo* are left untouched. Usually these exclamative words are independent in a sentence.

7. If you are not sure whether a word is a NSW or not, or if you know a word is a NSW, but you cannot determine the correct normalisation then mark it with "?" in the 5th column.

Index:61	469980378317722000			
1	RT	RT		
2	@LilReese300	@LilReese300		
3	:	:		
4	@Dante_Grady	@Dante_Grady		
5	lol	laughing out loud	#	
6	dey	dey		?
7	on	on		
8	crack	crack		
9	asum	asum	#	?

8. If you know a word is a NSW, and you think you have multiple corrections, then list them all with comma as delimiters. Note that these cases are expected to be rare in the data, because tweet context often resolves the ambiguity.

9. Contraction normalisations shall be in their contracted form in the correction, e.g., *dont* is normalised as *don't*

10. If a tweet or a word in non-English, then leave them unchanged.

Index:51	469919909036954000		
1	vIERON	vIERON	#
2	QUE	QUE	
3	GREG	GREG	
4	HABLA	HABLA	#
5	IGUAL	IGUAL	#
6	A	A	
7	NIALL	NIALL	#