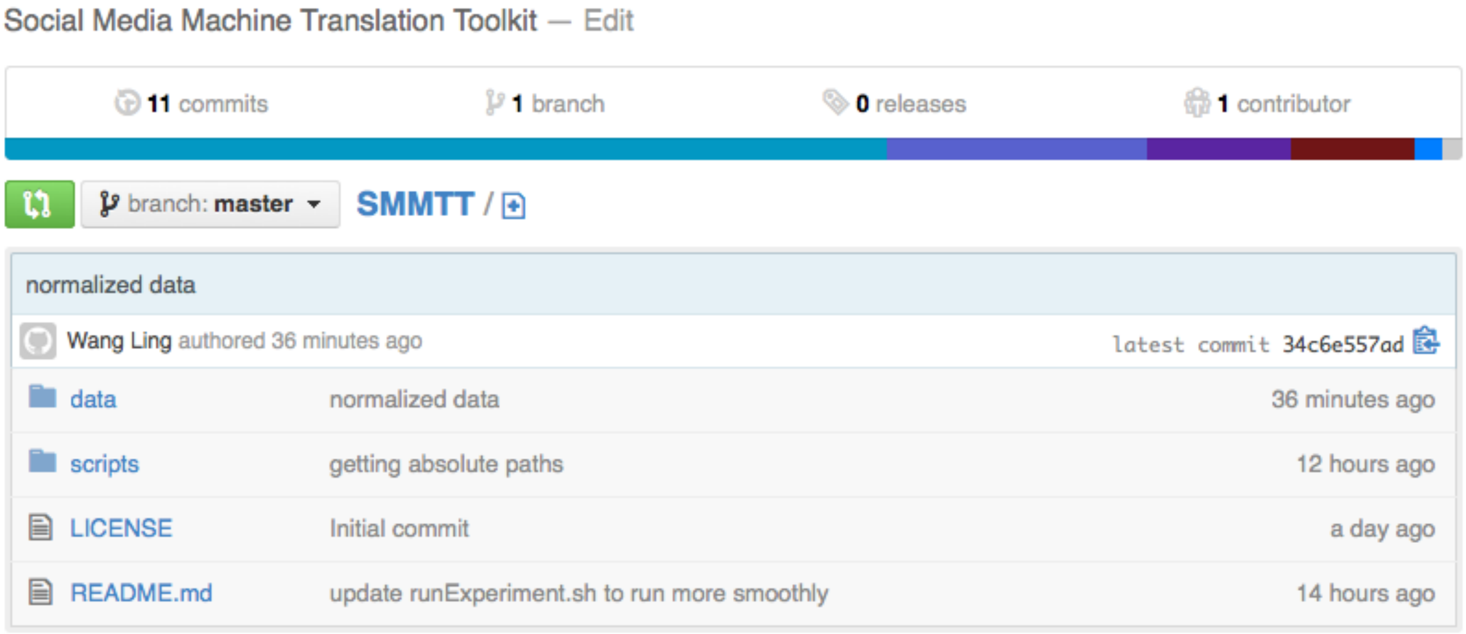
Social Media Machine Translation Toolkit (SMMTT)

Wang Ling
Carolin Haas
Chris Dyer
Adam Lopez

Social Media Machine Translation Toolkit (SMMTT)

- **Toolkit Available at https://github.com/wlin12/SMMTT**
  - Microblog Data
  - Tokenizer
  - Normalization



Social Media Machine Translation Toolkit — Edit

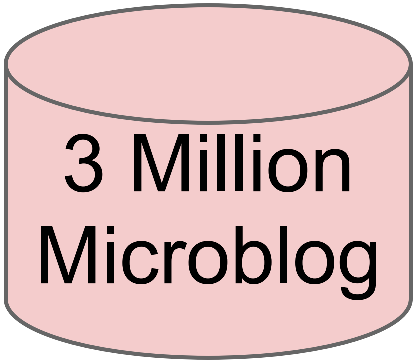
normalized data

Wang Ling authored 36 minutes ago — latest commit 34c6e557ad

data	normalized data	36 minutes ago
scripts	getting absolute paths	12 hours ago
LICENSE	Initial commit	a day ago
README.md	update runExperiment.sh to run more smoothly	14 hours ago

Social Media Machine Translation Toolkit (SMMTT)

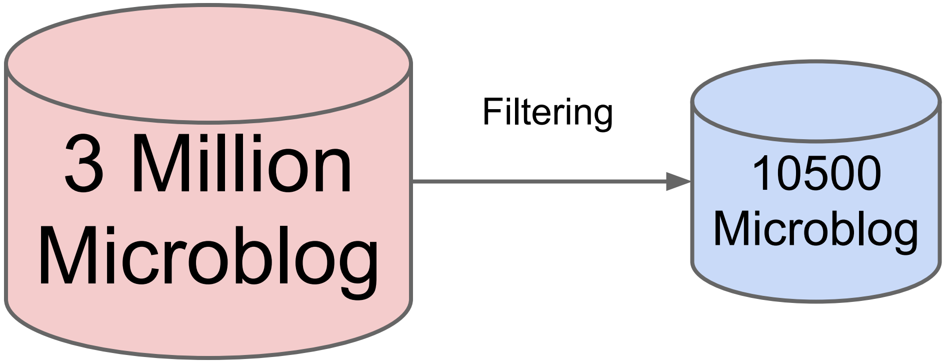
- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization



3 Million Microblog

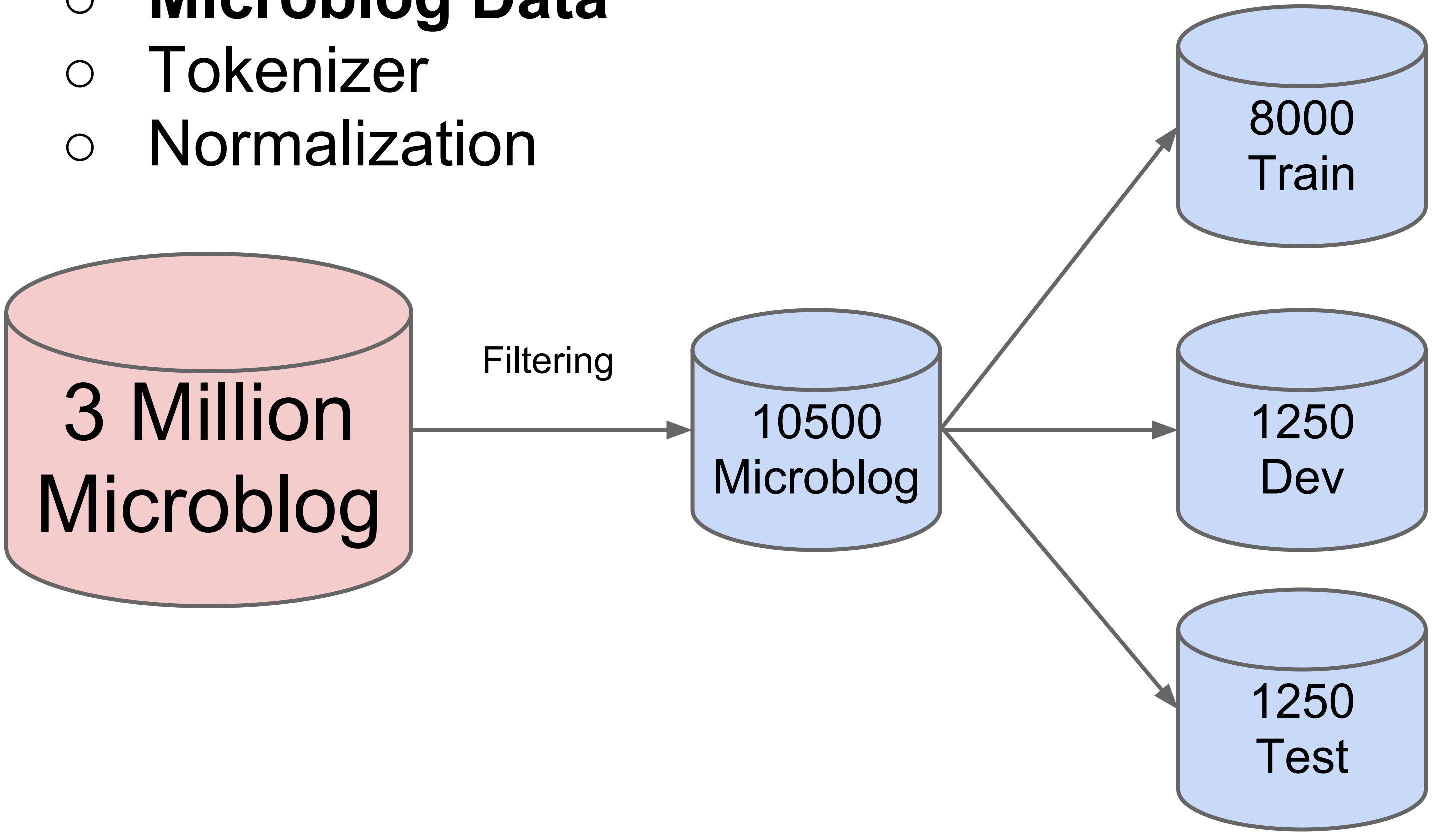
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization



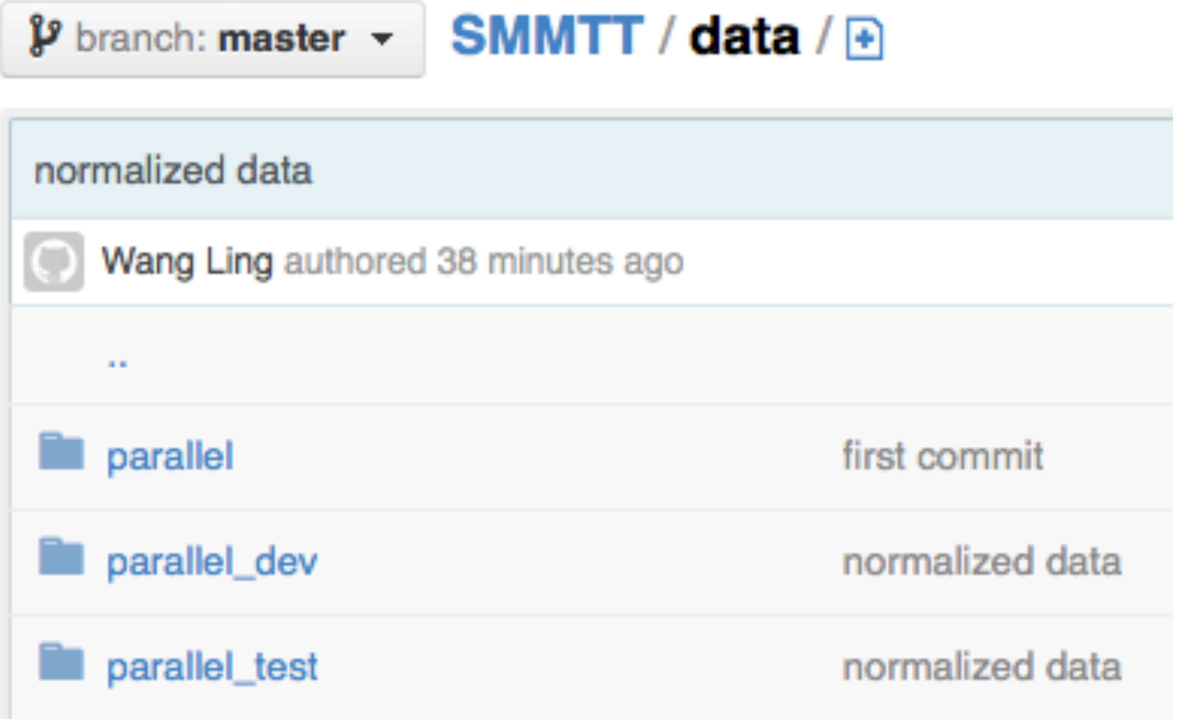
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization



Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
    - **Microblog Data**
    - Tokenizer
    - Normalization



SMMTT / data /

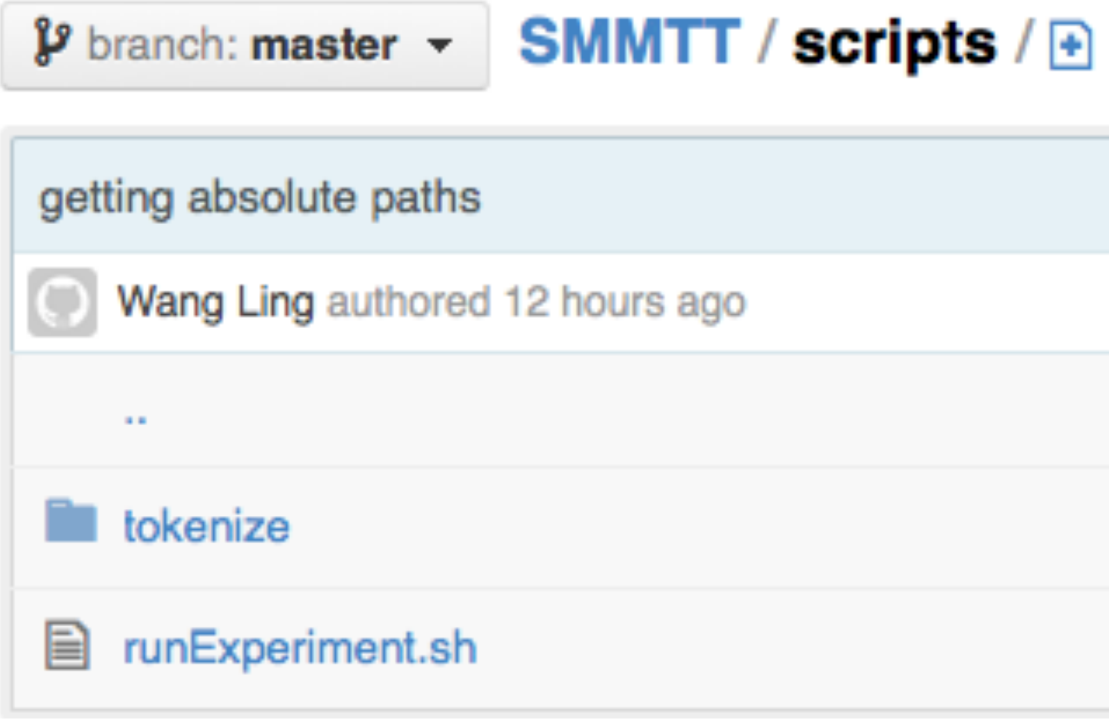
normalized data

Wang Ling authored 38 minutes ago

..	
parallel	first commit
parallel_dev	normalized data
parallel_test	normalized data

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
    - **Microblog Data**
    - Tokenizer
    - Normalization



getting absolute paths

Wang Ling authored 12 hours ago

..

tokenize

runExperiment.sh

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization

Dataset	BLEU
Weibo (8K)	**14.33**

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization

Dataset	BLEU
Weibo (8K)	14.33
FBIS (300K)	12.84

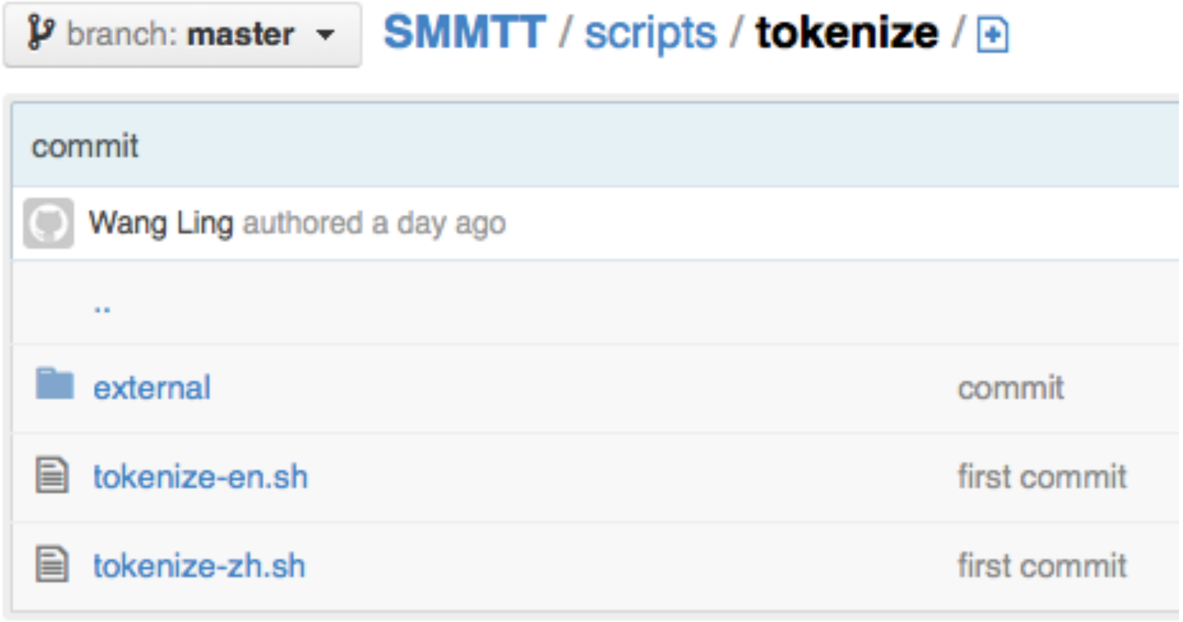
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - **Microblog Data**
  - Tokenizer
  - Normalization

Dataset	BLEU
Weibo (8K)	14.33
FBIS (300K)	12.84
Weibo+FBIS (308K)	**16.28**

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - **Tokenizer**
  - Normalization



SMMTT / scripts / **tokenize** /

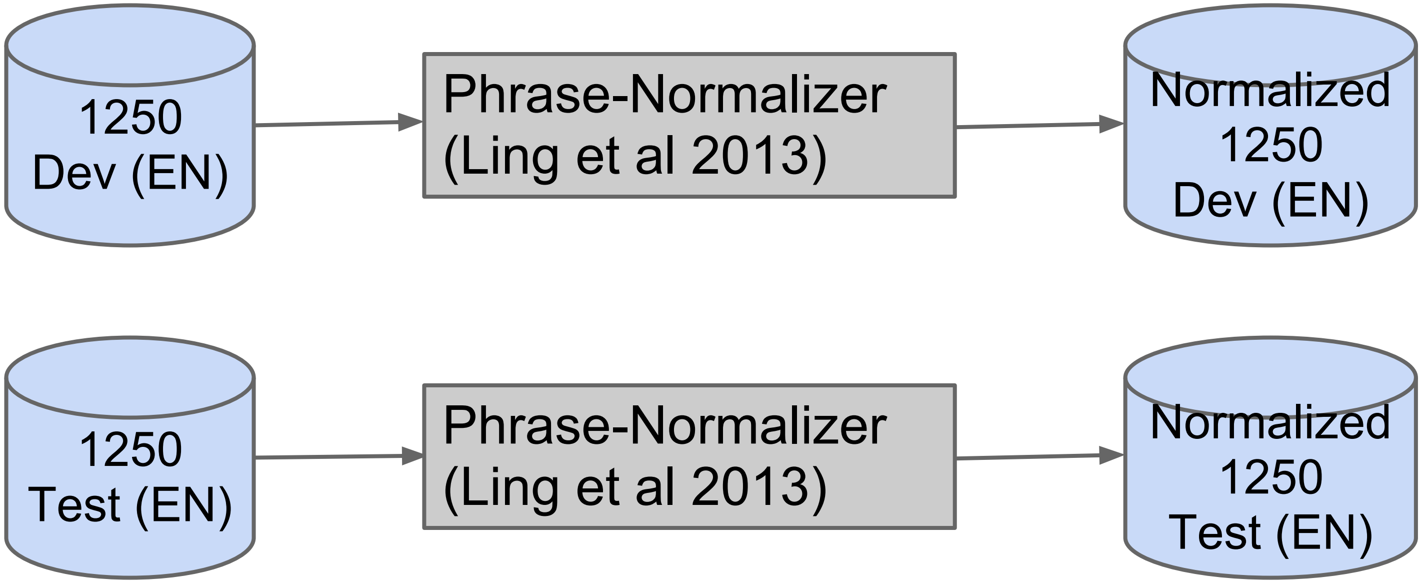
commit

Wang Ling authored a day ago

..	
external	commit
tokenize-en.sh	first commit
tokenize-zh.sh	first commit

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - **Normalization**



1250 Dev (EN)
Phrase-Normalizer (Ling et al 2013)
Normalized 1250 Dev (EN)
1250 Test (EN)
Phrase-Normalizer (Ling et al 2013)
Normalized 1250 Test (EN)

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - **Normalization**

Original	Normalized
To Unknown Nick: haha. You pick up the food; **I'll** cook it **brotha**!	to unknown nick: haha. you pick up the food; **i will** cook it**, brother** !

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - **Normalization**

Original	Normalized
To Unknown Nick: haha. You pick up the food; **I'll** cook it **brotha**!	to unknown nick: haha. you pick up the food; **i will** cook it**, brother** !
u guys are awesome ! cheers !	**you are the best** ! cheers !

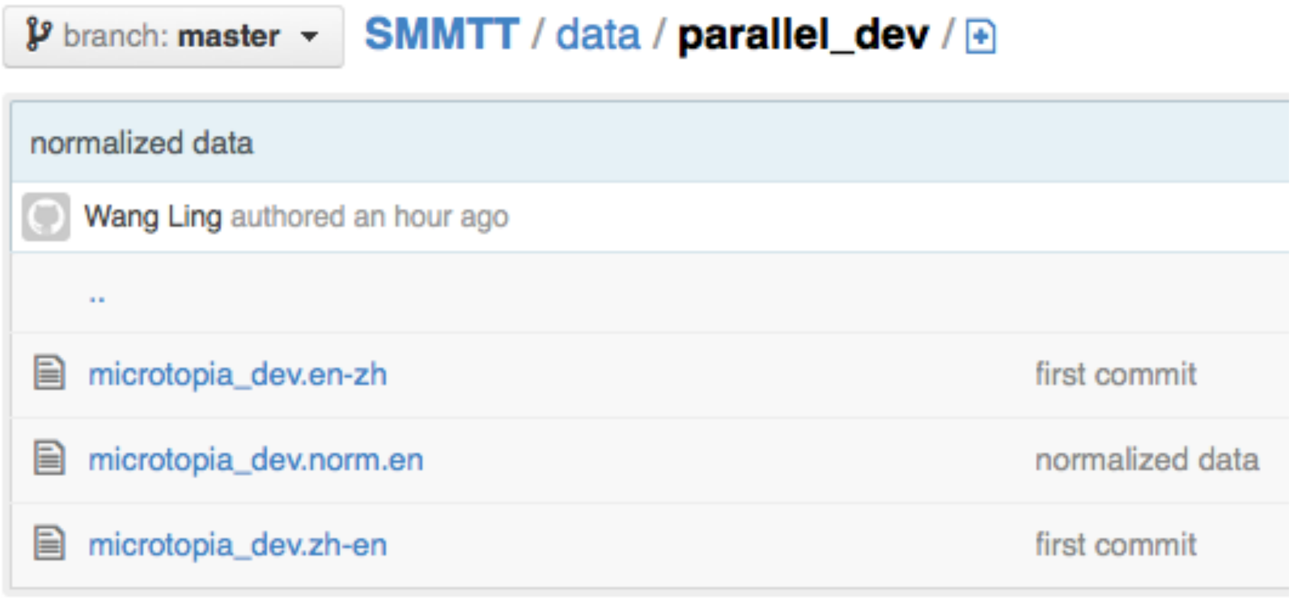
Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - **Normalization**

Original	Normalized
To Unknown Nick: haha. You pick up the food; **I'll** cook it **brotha**!	to unknown nick: haha. you pick up the food; **i will** cook it**, brother** !
u guys are awesome ! cheers !	**you are the best** ! cheers !
To **Fuzzy** Mannerz, ha! You**'re** cute for **sayin** that	to **vague** mannerz, ha! you **are** cute for **saying** that

Social Media Machine Translation Toolkit (SMMTT)

- Toolkit Available at https://github.com/wlin12/SMMTT
  - Microblog Data
  - Tokenizer
  - **Normalization**



branch: **master** ▾ SMMTT / data / **parallel_dev** /

normalized data

Wang Ling authored an hour ago

..

microtopia_dev.en-zh	first commit
microtopia_dev.norm.en	normalized data
microtopia_dev.zh-en	first commit

Thanks guys!!!!!!