Overview
Parallel Data Extraction
Machine Translation
Weibo Corpus
Twitter Corpus
Twitter Gold Corpus
Terms Of Use
Acknowledgements

Machine Translation in Social Media

While Machine Translation on Microblogs and Social Media is a hot topic, the actual work that has been published in this area is still very limited. We believe that one big factor is the lack of parallel data to train, tune and test MT systems.

However, this is an interesting problem, since text in this domain is radically different from commonly translated domains, such as news or parliament data. Here are some examples of parallel sentences that exemplify frequent phenomena in this domain:

	Source (English)	Target (Mandarin)
Abbreviations	She love my tattoos ain't got no room for her name, but **imma** make room -	她喜欢我的纹身，那上面没有纹她名字的地方了，不过我会弄出空余空间
Orthographic Errors	happy singles day in China - sorry I won't be **celebratin witchu**, I have my love... -	中国的粉丝们，光棍节快乐 - 抱歉我不能和你们一起庆祝节日了，我有我可爱的老婆S...
Syntactic Errors	Less guilty of some wrong, will be able to talk less "I'm sorry " to themselves or others are worth celebrating.	少犯一些错，就能少说“对不起”，这对自己或对别人都是值得庆幸的。
Emoticons	So excited to reveal the title of my new album on the new KellyRowland.com **:)** -	非常激动要在 KellyRowland.com 上揭晓我新专辑的名字了:)

Social Media Machine Translation Toolkit

To promote research in this direction, we present our toolkit SMMTT (Social Media Machine Translation Toolkit). You can check out from https://github.com/wlin12/SMMTT or download it here. If you are pursuing research on Machine Translations in Social Media or Microblog data, we recommend you to use this toolkit as a starting point, as it provides a baseline for your experiments.

Data - The toolkit provides 8000 training, 1250 development and 1250 test sentence pairs for the Mandarin-English language pair. These were carefully extracted from the full 3M sentence pairs extracted from Sina Weibo using heuristics (filtering duplicates, removing frequent alignment errors and finding users that frequently post parallel messages). You can find this corpus in the "./data" directory.

Modeling - It also provides scripts to automatically build a translation system using the training and development sets, and evaluate the results using the test set by running the "./scripts/runExperiment.sh" script.

Baseline - Results were computed and presented in the MT marathon 2013. You can use these as baseline for your experiments.

Datasets used	BLEU	Experiment Description
µtopia (8k)	14.33	Relatively small in domain data
FBIS (300k)	12.84	Relatively large out of domain data
µtopia (8k)+FBIS (300k)	16.28	Combining in domain and out of domain data

For now, you can use this article as a reference to the toolkit.

The content in this website was created by Wang Ling. Feel free to mail me ideas, feedback or suggestions and I will do my best to accommodate them.