实验三: 机器翻译

一、数据集

- 地址: http://www.statmt.org/wmt16/translation-task.html (WMT16)
- 选择**新闻领域数据集**:

Shared Task: Machine Translation of News

[HOME] [SCHEDULE] [PAPERS] [AUTHORS] [RESULTS]
TRANSLATION TASKS: [NEWS] [IT-DOMAIN] [BIOMEDICAL] [MULTIMODAL] [PRONOUN]
EVALUATION TASKS: [METRICS] [QUALITY ESTIMATION] [TUNING]
OTHER TASKS: [AUTOMATIC POST-EDITING] [BILINGUAL DOCUMENT ALIGNMENT]

• 选择英语->德语的翻译任务:

The recurring translation task of the WMT workshops focuses on news text and European language pairs. For 2016 the language pairs are:

- Czech-English
- German-English
- Romanian-English
- Finnish-English
- Russian-English
- Turkish-English
- 训练数据不受限制, 自取:
- DOWNLOAD

Parallel data:								
File	Size	CS-EN	DE-EN	FI-EN	RO-EN	RU-EN	TR-EN	Notes
Europarl v7	628MB	V	V					same as previous year, corpus home page
Europarl v8	215MB			✓	√			ro-en is new for this year, corpus home page
Common Crawl corpus	876MB	V	V			✓		Same as last year
News Commentary v11	72MB	4	1			✓		updated
CzEng 1.6pre	3.1GB	V						New for 2016. Register and download CzEng 1.6pre.
Yandex Corpus	121MB					✓		ru-en
Wiki Headlines	9.1MB			1		√		Provided by CMU
CETTA IECO	22 4 40	$\overline{}$		$\overline{}$		-		DI LI LI LI DOUG

Monolingual language model training data:

Corpus	cs	DE	EN	FI	RO	RU	TR	All languages combined	Notes
Europari v7/v8	32MB	107MB	99MB	95MB					
News Commentary	13MB	17MB	20MB			17MB		65MB	Updated
Common Crawl	10.5GB	102GB	103 GB	5.3GB	11.3GB	42GB	18GB		New for 2016, Deduplicated with development and evaluation sentences removed. English was updated 31 January 2016 to remove bad UTF-8. Downloads can be verified with SHAS12 checksums. More English is available for unconstrained participants.
News Crawl: articles from 2007	3.7MB	92MB	198MB					302MB	
News Crawl: articles from 2008	191MB	313MB	672MB			2.3MB		1.5GB	
News Crawl: articles from 2009	194MB	296MB	757MB			5.1MB		1.6GB	
News Crawl: articles from 2010	107MB	135MB	345MB			2.5MB		727MB	
News Crawl: articles from 2011	389MB	746MB	784MB			564MB		3.1GB	News Crawl
News Crawl: articles from 2012	337MB	946MB	751MB			568MB		3.1GB	Extracted article text from various online news publications.
News Crawl: articles from 2013	395MB	1.6GB	1.1GB			730MB		4.3GB	The data sets from 2007-2014, and news-discuss, are the same as last year's.
News Crawl: articles from 2014	380MB	2.1GB	1.4GB	52MB		801MB		5.3GB	<u> </u>
News Discussions, Version 1 from 2014/15			<u>1.7GB</u>						
News Crawl: articles from 2015	360MB	2.2GB	1.3GB	203MB	125MB	608MB		4.8G	

THE COMMON CHAIR MONORING DATA IS NOSEED by PARISON AREO SERVICES as a <u>power, used Sep</u>. The dispersing 35 Oric is \$257 Met supplies assistant used

- 验证集和测试集必须用下列给定数据集:
 - Development sets (22 MB)
 - Updated Romanian dev, with fixed diacritics (0.5 MB)
 - NEW: <u>Test sets</u> (3.4 MB) Test sets (source and references)

二、方法 (Seq2Seq)

- Word2vec、Glove、FastText、BERT、RoBERTa、XLNet
- GRU、LSTM、BiLSTM
- Transformer

三、评价指标

- BLEU Score: https://blog.csdn.net/guolindonggld/article/details/56966200
- NLTK模块: <u>nltk.align.bleu score</u>
- BLEU值不得低于0.25

四、参考论文

- 19 ICLR MULTI-AGENT DUAL LEARNING
- 16 ACL Edinburgh Neural Machine Translation Systems for WMT 16
- <u>21 Lessons on Parameter Sharing across Layers in Transformers</u>