一些关键词，缩写

> XLM: cross-lingual Language Model
>
> 平行文本（parallel text）是与译文并行放置的文本。
>
> 平行文本的大集合被称为平行语料库（parallel corpora). 在翻译过程中，翻译人员可以对句子进行拆分、合并、删除、插入或重新排列。这使得对齐任务变得异常重要。
>
> zero-shot learning: 零样本学习
>
> BERT：pretrained language model architecture
>
> NLI: natural language inference
>
> speech recognition: 语音识别
>
> word alignments: natural language processing task of identifying translation relationships among the words (or more rarely multiword units) in a bitext, resulting in a bipartite graph between the two sides of the bitext, with an arc between two words if and only if they are translations of one another. Word alignment is typically done after sentence alignment has already identified pairs of sentences that are translations of one another.
>
> RA：Reorder Augmentation
>
> SA：Semantic Augmentation
>
> FastAlign: 词对齐模型
>
> PGD: projected gradient descent

主要内容

本论文主要涉及到供给跨语言的NLP模型的源语言数据扩增方法的构建，和用来增强模型鲁棒性的对抗文本的生成方法。其后作者还介绍了他们实验的成果。

abstract中提到，除去基本的单一语言的NLP模型，现在的跨语言的NLP可以做到在只提供源语言数据集作为训练集的基础上做出对特定目标的新语言的分析处理。这在我初次读到时觉得不可思议，因为机器之前甚至没有得到目标语言的数据集标签，怎么能做出分析呢？后来想到，一些不同的语言其实非常类似，比如英语和德语都是印欧语系-日耳曼语族下的语言，发展至今肯定会有许多类似或者互相借鉴的地方，如果模型能抓住共同的特征，实现跨语言分析也不是不可能。况且，所有的语言都存在一些共同点任何语言都一定有表示人或事物的名词、表示动作行为的动词、表示性质状态的形容词和表示称代的代词等词类；任何语言的句子结构都有一定的层次规则和语序规律；任何语言都有表示“时、体、态”等语法意义的方法等等。 有学者根据语言学研究的结果列举了人类语言的一些普遍现象：所有语言的词汇量都可增加；所有语言中音义结合的方式都是任意的；所有语言都有大致相同的词类；所有语言的语法规则都有递归性等等。

作者认为，跨语言分析的NLP模型的主要目的就是解决某些语言可能存在的标记数据集太少的问题（scarce labeled data）,因此，他们提出了一种新的数据扩增策略，能丰富数据集，从语义层面反映多样性。要如此，有二法。其后，他们又利用对抗训练作为补充增强鲁棒性。通过对最后结果的 分析，他们在跨语言推导方面发现了很大的突破！

数据集扩增，解决的就是特定领域的平行语料库过于稀疏无法训练模型的问题。

作者要实现的一个跨语言的模型系统就是实现

$$L_S(trained) \rightarrow L_T(target)$$

之前对跨语言的处理方法： In the past, such systems typically drew on **translation dictionaries**, **lexical knowledge graphs**, or **parallel corpora**, to build a cross-lingual model that exploits simple connections between words and phrases across different languages

现在的一些预训练语言模型架构能在一个共享的词汇库下，用自监督学习结合的多语言展现形式。但是，产出的结果还是很依赖于目标语言与源语言之间相似度的大小。 比如“some more distant language pairs ”就不理想（substaintial drops)。

作者主要聚焦于NLI，即推断预定句子与假设句子之间的近似关系，是相矛盾还是相一致还是中立的。难点在于不仅要推断出隐晦的字面意义上的差异， 还要能分析出语言环境的前提假设和隐含语义。

我们的数据扩增，要在目标语言的未标记的数据上做文章。这样才能让模型更好解释目标语言的特性。，不然调得再好也是只能针对源语言。一个自然的方法是标准的半监督方法让模型自己预测未标记数据集输入。问题是，这些预测可能包含了太多噪音，不准，做不了训练的信号。作者的思想是，正如在计算机视觉和语音识别中常见的那样，“generate new training data from existing labeled data”。直接性的处理手段是对句子序列做操作，包括 同义词替换（synonym replacement）, 随机插入（random insertion）,随机调换（ random swapping），随机删除（random deletion）等。相应的问题是合成数据噪音大、不可信；语义会发生偏移。（重要！！！ NLI主要就是对句子隐晦差异的推理）

数据扩增 Data Augmentation

> One serves to encourage language adaptation by means of reordering source language words based on word alignments to better cope with typological divergency between languages, denoted as Reorder Augmentation (RA). Another seeks to enrich the set of semantic relationships between a premise and pertinent hypotheses, denoted as Semantic Augmentation (SA).

RA简单说就是利用词对齐重新排序源语言句子，来处理两种语言之间的拓扑差异问题。

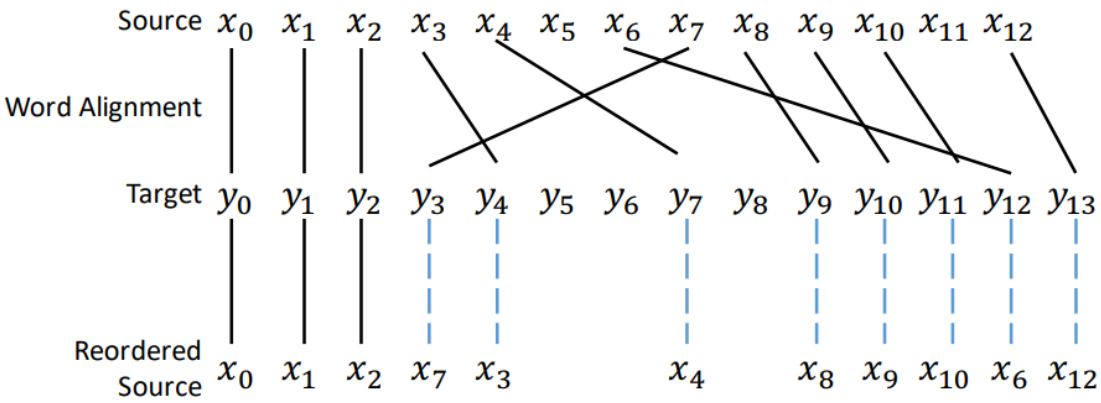
SA简单说就是丰富前提语句和假设语句的语义关系。

RA

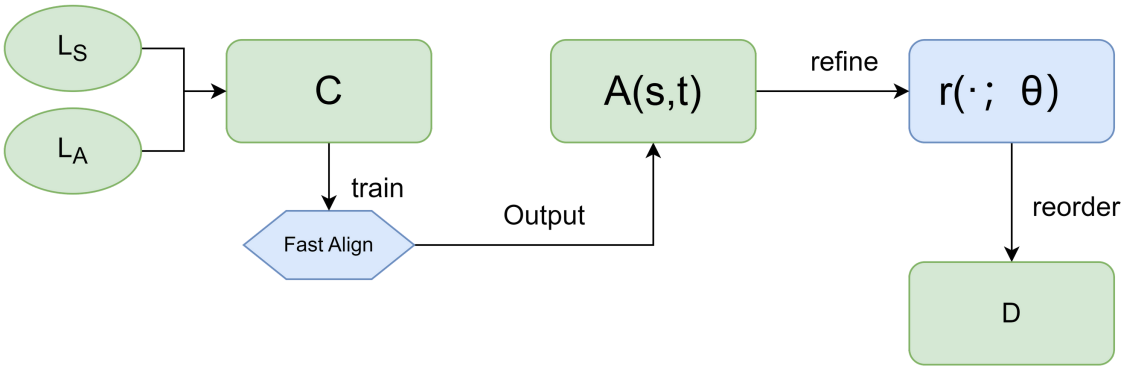
Reorder augmentation is based on the intuition of making a model more robust with respect to differences in word order typology.

举例：如果L_S中的句子全是英语那样的**subject–verb–object** (SVO)格式，模型就很难发现与遵行**subject–object–verb** (SOV)格式的目标语言之间的细微语义差异。

1. 首先，对没有标注的双语平行语句L_S和作为”auxiliary language"的L_A实现词对齐，L_A不要求和L_T一样。
2. 基于对齐重排序源语言句子，来匹配L_A，然后根据这一新生成的平行语料库调整原来的Seq2Seq模型，然后利用调优之后的Seq2Seq模型对原来的NLI标注训练集里的句子重新排序，优化训练集，实现数据扩增。



下面是我自己对RA的理解：



- LA: Auxiliary Language
- C: unannotated bilingual parallel sentence pairs
- r(·； θ）:Seq2Seq Model
- A(s,t): word pair table for each sentence pair
- D:Labeled training dataset

: Model

: Dataset

Made by xlm

SA

需要训练一个可控模型使得给一个句子和一个期待关系标签，能产出一个符合描述的句子。

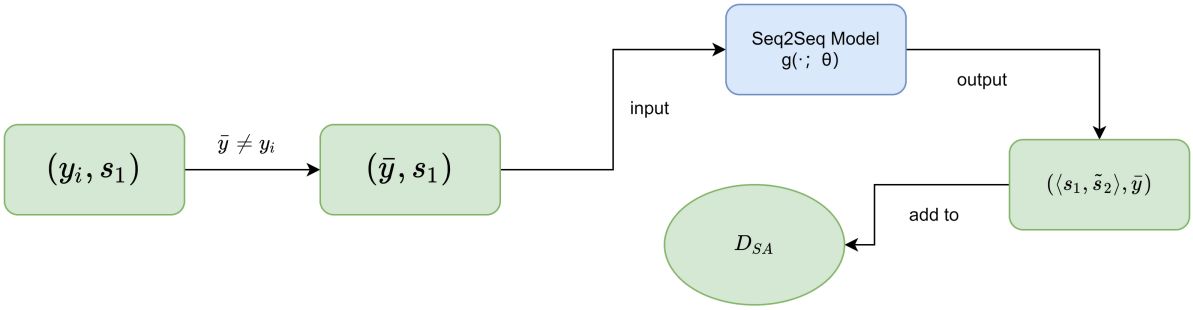
$$< s_1, \bar{y} > \rightarrow \tilde{s_2}$$

这样一来，针对已有的训练语句对，就能激发模型自动生成许多合适的不同新语句对。考虑到这样产出的结果都带噪音，我们再加入$Teachermodel$来精修这些合成样例。

Generation

作者把requested label 放在作为输入给Seq2Seq模型的文本语句的前面做前缀。

下面是我对SA的理解。



Made by xlm

Label Rectification

我们得到的$\tilde{s}_2$不能保证总是"have the desired relationship $\bar{y}$ to s_1",所以要加入纠正策略。我们需要一个$TeacherModel$。使用开始的标记数据集D,先通过K次迭代得到一个"教学网络"（Teacher Network）：$h(\cdot;\theta)$。这个网络就用来自动验证和纠错我们之数据扩增得到的数据集。给定$\tilde{x}_i$,模型会给出$(\tilde{y}_i, c)$,其中$\tilde{y}_i$是预测的可能关系，$c \in [0, 1]$是置信度得分。

我们有两个策略决定最终标签。

- Teacher Strategy：

  $$D_r = \{(\tilde{X}_i \tilde{y}_i) | (\tilde{X}_i \tilde{y}_i) \in D_{\tilde{a}}, (\tilde{y}_i, c) = h(\tilde{X}_i), c > T\}$$

  就是说当置信度高于T时，我们就有理由相信，Teacher model 能确保一个可靠的标签，否则直接丢弃

- TR Strategy：

  $$D_r = \{(\tilde{X}_i, \Phi(y_i, \tilde{y}_i, c)) | (\tilde{X}_i \tilde{y}_i) \in D_{\tilde{a}}, (\tilde{y}_i, c) = h(\tilde{X}_i)\}, where$$
  $$\Phi(y_i, \tilde{y}_i, c) = \begin{cases} \tilde{y}_i, c > T \\ y_i, otherwise \end{cases}$$

  这里，只有当模型足够确信时，才改用模型给的标签否则就保持原来的标签。

对抗训练 Adversarial Training

经过上面的步骤，得到合成数据集合D_r,其可以被整合进原有数据集D,从而得到最终训练集$D_a = D_r \cup D$,有了这些，我们就能训练出一个新模型$f(\cdot;\theta)$,来进行最后的跨语言句子对分类。不过为了"minimize the maximal loss incurred by label-preserving adversarial perturbations, thereby promising to make the model more robust"，还得加入一个对抗训练。这部分细节没有深入挖掘。大致含义就是，要寻找到最优的参数θ,使得能对抗在一定范围内的任意扰动r.涉及到求损失函数。作者利用了projected gradient descent，投影梯度下降来进行迭代计算。

结果 Main Results

Table 2: Accuracy (in %) on XNLI with augmented examples used for cross-lingual transfer. The number of augmented examples from EA, RA and SA are 80k, 20k, 80k. EA (Wei and Zou, 2019) is Easy Data Augmentation The best cross-lingual transfer results under XLM-R are given in boldface.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
In-Language Supervised Learning (Translate–Train)																
RoBERTa	88.2															
mBERT	73.3	65.2	69.0	66.5	66.5	64.8	61.7	57.7	56.3	65.8	63.4	49.3	61.5	66.9	59.3	63.1
XLM-R	77.7	70.6	73.0	68.1	72.8	70.6	67.4	61.8	60.5	73.2	71.0	68.9	69.3	70.2	64.9	69.3
Zero-Shot Cross-Lingual Transfer																
XLM-R	77.7	71.7	72.6	69.5	72.7	70.2	67.7	60.7	61.0	72.0	70.2	67.4	69.0	71..0	64.9	69.1
+PGD	78.9	71.8	74.5	70.2	73.5	71.1	67.3	60.7	62.0	72.9	71.3	68.7	69.2	71.3	64.9	69.9
+EA(80k)	77.8	70.3	73.1	69.2	72.9	70.3	67.5	61.6	63.5	72.1	70.1	68.1	68.7	69.5	65.1	69.3
+RA(20k)	78.4	71.0	73.1	67.3	73.0	70.2	67.1	61.5	61.1	71.9	70.3	65.5	67.5	69.5	64.7	68.8
+SA(80k)	79.5	72.0	74.4	69.6	74.1	71.9	67.5	63.6	62.7	73.6	71.9	69.0	69.2	71.0	66.1	70.4
+EA+PGD	77.9	71.9	74.4	71.1	73.5	71.5	68.8	63.3	64.4	74.1	68.3	69.5	68.9	70.4	66.9	70.3
+RA+PGD	78.9	72.5	74.7	71.1	74.5	72.0	68.6	63.1	63.6	73.3	72	69.0	69.9	71.7	65.9	70.7
+SA+PGD	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	71.5
+EA+SA+PGD	80.0	74.0	76.1	73.0	75.5	73.9	**70.2**	63.7	65.5	75.4	73.3	70.5	71.4	72.9	68.0	72.2
+RA+SA+PGD	**80.8**	**74.5**	**77.3**	**73.6**	**75.8**	**74.9**	70.0	**64.8**	**65.7**	**76.3**	**74.9**	**71.6**	**71.4**	**74.5**	**68.5**	**73.0**

作者利用控制变量法，先在单一语言内进行训练，然后使用跨语言训练，逐步增加对抗训练、数据扩增的条件来进行对比。注意EA是传统的NLP的扩增方法。可以发现，加入对抗训练后准确率得到提升，数据扩增后同样有所帮助，但是XLM-R几乎不能从RA中得到什么帮助，作为对比，SA有很大帮助。

Table 3: Accuracy (in %) on XLNI with different rectifying strategies, training on XLM-R with SA and PGD. T is the threshold. p denotes the percentage of initial augmented examples retained for training.

Approach	p	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
Teacher ($T = 0$)	100%	79.7	72.8	75.6	71.7	73.9	73.0	69.3	64.5	63.8	74.0	72.6	69.8	70.0	71.8	66.5	**71.3**
Teacher ($T = 0.8$)	94%	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.6**
TR ($T = 0.8$)	100%	79.1	72.9	75.3	71.4	74.1	73.1	68.8	64.1	63.6	73.9	73.1	70.4	70.4	72.0	66.6	**71.3**
Agreement	66%	78.7	71.3	74.5	70.8	72.7	71.7	68.7	63.8	62.6	73.0	72.0	69.7	69.4	71.1	65.9	**70.4**
Requested	100%	75.4	67.5	70.1	69.0	68.0	69.2	65.7	61.1	61.6	70.5	68.3	65.9	68.3	70.6	64.1	**67.7**

table 3表明，teacher model 的加入带来一定收益。

Table 4: Accuracy (in %) on XNLI experiments with different amounts of training and augmentation data, and different adversarial training methods.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
XLM-R (10k)	74.5	68.0	70.3	65.5	70.8	68.0	64.2	61.1	60.2	69.9	68.9	65.0	66.9	68.4	61.5	**66.9**
+SA +FGM	77.5	70.9	73.6	68.3	73.1	70.7	67.3	62.2	62.2	72.8	70.5	68.4	67.3	70.2	64.9	**69.3**
+SA +PGD	78.2	71.4	73.7	70.8	73.1	71.2	68.1	62.3	63.2	73.6	71.8	68.9	69.2	71.1	65.6	**70.1**
+RA +SA +PGD	79.1	73.0	75.2	72.3	73.6	72.5	69.2	64.5	63.7	74.5	72.3	70.0	70.7	72.7	67.2	**71.4**
Improvement(%)	6.2	7.4	7.0	10.0	4.0	6.6	7.8	5.6	5.8	6.6	4.9	7.7	5.7	6.3	9.3	**6.7**
XLM-R (20k)	77.7	70.0	72.5	69.2	72.7	70.6	66.9	61.6	60.8	72.0	70.2	66.7	68.7	70.6	64.9	**69.0**
+SA +FGM	79.3	72.4	74.7	70.6	73.7	71.8	67.6	63.5	63.0	72.9	71.9	68.3	69.3	71.6	66.6	**70.5**
+SA +PGD	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.6**
+RA +SA +PGD	80.8	74.5	77.3	73.6	75.8	74.9	70.0	64.8	65.7	76.3	74.9	71.6	71.4	74.5	68.5	**73.0**
Improvement(%)	4.0	6.4	6.6	6.4	4.3	6.1	4.6	5.2	8.1	6.0	6.7	7.3	3.9	5.5	5.5	**5.8**

table 4是FGM的对比。结果表明，对同样的训练数据集，PGD的准确率要高于FGM，说明PGD能提供更好的对抗扰动来增加模型鲁棒性。

Table 5: Accuracy (in %) on XNLI experiments trained using 20k vs. 80k augmentation data from EA, RA, SA, with and without PGD.

Approach	*en*	*de*	*es*	*zh*	*fr*	*ru*	*ar*	*sw*	*ur*	*bg*	*el*	*th*	*tr*	*vi*	*hi*	**avg**
XLM-R (20k)	77.7	71.7	72.6	69.5	72.7	70.2	67.7	60.7	61.0	72.0	70.2	67.4	69.0	71..0	64.9	69.1
+EA (20k)	77.4	69.1	71.9	67.5	71.6	69.3	65.5	61.0	61.5	71.1	69.2	67.1	67.1	68.8	63.9	68.1
+EA (80k)	77.8	70.3	73.1	69.2	72.9	70.3	67.5	61.6	63.5	72.1	70.1	68.1	68.7	69.5	65.1	**69.3**
+RA (20k)	78.4	71.0	73.1	67.3	73.0	70.2	67.1	61.5	61.1	71.9	70.3	65.5	67.5	69.5	64.7	**68.8**
+RA (80k)	77.5	70.8	73.3	68.1	72.2	70.3	66.8	60.7	60.3	72.5	70.5	66.0	67.6	69.3	63.3	68.6
+SA (20k)	78.2	70.6	72.8	67.3	72.6	70.3	66.5	61.4	60.4	71.8	69.6	66.9	67.6	69.5	64.0	68.6
+SA (80k)	79.5	72.0	74.4	69.6	74.1	71.9	67.5	63.6	62.7	73.6	71.9	69.0	69.2	71.0	66.1	**70.4**
+PGD	78.9	71.8	74.5	70.2	73.5	71.1	67.3	60.7	62.0	72.9	71.3	68.7	69.2	71.3	64.9	69.9
+EA +PGD (20k)	77.6	70.9	73.9	69.8	73.0	71.1	67.1	62.4	63.8	73.0	71.3	68.9	69.1	71.2	65.8	69.9
+EA +PGD (80k)	77.9	71.9	74.4	71.1	73.5	71.5	68.8	63.3	64.4	74.1	68.3	69.5	68.9	70.4	66.9	**70.3**
+RA +PGD (20k)	78.9	72.5	74.7	71.1	74.5	72.0	68.6	63.1	63.6	73.3	72	69.0	69.9	71.7	65.9	**70.7**
+RA +PGD (80k)	78.4	71.9	74.9	71.0	73.7	71.9	68.7	62.6	64.0	73.4	72.1	68.9	69.9	71.9	66.4	70.4
+SA +PGD (20k)	79.3	73.3	74.0	69.4	73.3	71.0	67.6	62.7	62.4	73.7	71.7	68.3	69.28	71.1	65.6	70.2
+SA +PGD (80k)	80.4	73.4	75.7	71.8	74.0	73.1	69.3	64.5	63.7	74.5	73.2	70.3	70.2	72.3	66.9	**71.5**

table 5展示了扩增数据数量对准确率的影响。80k是扩增影响的上限，20k是RA的上限。相对EA（EA肯定不行啊），RA等扩增带来的效果更明显，EA要80k才能打败用了PGD的XLM-R，RA只要20k就够。

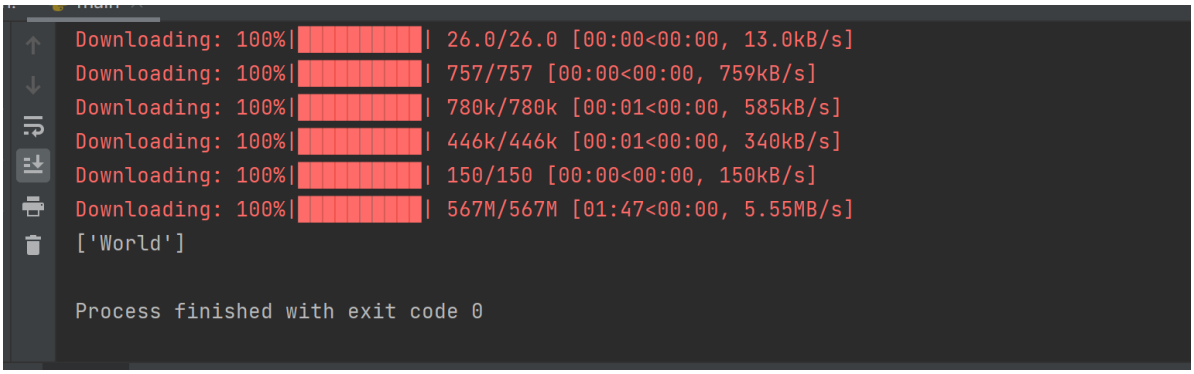
工具复现

借鉴了text2text

下载了facebook的一个多语言翻译模型，能实现100种语言的互相翻译，作为测试。

学习参考了text2text的实现过程。包括基本的分词，向量化，tfidf计算等等。

总的来说只能跑通和使用给的一些函数调用，自己实现还存在困难。



> https://zhuanlan.zhihu.com/p/368226087

其他

论文主要是text augmentation啊，不是image augmentation?本来想着选过数字图像处理的课能帮上点忙，没想到直接寄😭😭

ɔust Unsupervised Data Augmentation Framework for Zero-Resource Cross-Lingual NLP	Record & Replay
Data Augmentation with Adversarial Training for Cross-Lingual NLI	Image Augmentation
AdvAug: Robust Adversarial Augmentation for Neural Machine Translation	Image Augmentation