一种利用命名实体知识的神经机器翻译数据增强方法

张敏，杨浩*，陶仕敏，王明涵，郭嘉鑫，陈一萌，

苏畅，杜纯宁，王宇侠，秦璎

（华为技术有限公司 2012 实验室，北京，100095）

摘要：神经机器翻译对数据非常依赖，数据增强方法因此广泛应用其中。本文是针对文献[1]的改进，从命名实体知识角度出发，提出了一种基于命名实体加权的 Fuzzy Matching 方法（Named Entity Weighted Fuzzy Matching, NEWFM）用于计算源语言相似句；同时针对文献[1]中词标签依赖外部处理程序可能带来不准确的问题，提出了一种简洁且有效的词标签生成方式。在本文提出的 NEWFM 和词标签基础上，给出了应用于神经机器翻译数据增强的整体流程，其有效性在英法医药领域数据上得到实验验证。

关键词：神经机器翻译，数据增强，命名实体，Fuzzy Matching，词标签

中图分类号：TP391.2　　**文献标志码：**A

1. 引言

利用翻译记忆库（Translation Memory, TM）来提升翻译质量，因其直观性和可行性，从统计机器翻译时代开始，在学术界和工业界就被广泛关注[2,3]。随着神经机器翻译技术的突破和发展[4-7]，翻译记忆库 TM 的应用可以分成 3 类：

1. 直接用于训练模型[8]：将 TM 数据作为训练数据的一部分，或作为领域数据进行模型微调（Fine-Tuning）
2. 术语约束[9]：将 TM 数据作为或挖掘出平行术语库，在神经机器翻译中作为术语约束引入，一般通过受限解码实现[10]。
3. 数据增强[11]：在 TM 数据中找出与源语言句相似的句子，将其译文作为源语言句的额外信息引入到模型训练中，让模型学习如何使用这些额外信息，类似人类的翻译过程。文献[1]提出了一种利用 TM 中相似翻译的数据增强方法，在其框架下 Fuzzy Matching 是翻译效果最好的相似度计算方法，并为此设计了一种词标签方式，翻译效果得到进一步提升，实验结果表明该数据增强方法不仅具备目标语言词复制能力（*copy effect*），还具备一定的上下文能力（*context effect*）。

本文提出的方法，是针对文献[1]数据增强方法的改进，该文提出了基于相似句译文和词标签的数据增强方法，但在相似度计算中没有考虑不同词之间的差异性，而且词标签计算依赖外部程序 Fast Align[12]的效果，很大概率会带来不准确的问题。在领域翻译场景中，领域命名实体翻译质量的重要性不言而喻，会直接决定整个句子的翻译质量。以此为出发点，我们对源语言句子和 TM 源语言句子都进行领域命名实体识别，在相似度计算 Fuzzy Matching 中引入命名实体知识，找到与源语言句子领域更相关的句子译文进行数据增强，获得了更好的翻译效果。另外，我们设计了一种简洁且有效的词标签生成方式，不依赖外部程序，并取得了类似文献[1]中的提升效果。

本文以下部分的组织方式如下：在第 2 节中介绍基于命名实体加权的匹配度计算方法 NEWFM；在第 3 节中介绍新的词标签生成方法，并给出本文的数据增强方法；第 4 节给出英法医药领域数据集上的实验对比结果；最后在第 5 节对全文进行总结。

2. 基于命名实体加权的 Fuzzy Matching

* 通讯作者：yanghao30@huawei.com

2.1 Fuzzy Matching

Fuzzy Matching[1]的定义：从词匹配维度计算两个句子的匹配度。对给定的两个句子S_i和S_j，匹配度的计算如下：

$$FM(S_i, S_j) = 1 - \frac{ED(S_i, S_j)}{max(|S_i|, |S_j|)} \quad (1)$$

其中，$ED(S_i, S_j)$表示两个句子S_i和S_j在词级别上的编辑距离，$|S_i|$和$|S_j|$分别表示句子S_i和S_j所包含的词个数。

需要指出的是，尽管计算两个句子匹配度的方法很多，如 N-gram Matching、向量化表示等[1]，但从神经机器翻译的数据增强角度来看，Fuzzy Matching 更合适一些，文献[1]的实验结果也表明了这点（在上述提到的匹配度计算方法中，通过 Fuzzy Matching 进行数据增强，翻译效果提升最显著[1]），这也是我们对其改进优化的出发点。

2.2 NEWFM：Named Entity Weighted Fuzzy Matching

文献[1]在计算两个句子的编辑距离$ED(S_i, S_j)$时，认为每个词插入、删除、替换的代价是一样的，而我们知道在领域翻译任务中，领域命名实体的翻译对整体翻译至关重要，因此设计了一种基于命名实体加权的 Fuzzy Matching 即 NEWFM，以保证领域命名实体词插入、删除、替换的代价不同于其它词，具体计算流程如图 1。

输入：句子S和T，词个数分别为m和n

输出：NEWFM 匹配度

步骤：

1. 对输入句子进行领域命名实体识别，句子中的每个词属于实体词或非实体词
2. 定义实体词和非实体词插入、删除、替换的代价函数$cost(word_type, op_type)$
   其中$word_type$取值实体或非实体，op_type取值$insert$、$delete$、$substitute$之一
3. 计算S和T的 NEWFM 匹配度

```
d[0,0] = 0
for i from 1 to m:
    d[i, 0]+= cost(word_type(S[i]), delete)
for j from 1 to n:
    d[0, j]+= cost(word_type(T[j]), insert)
for j from 1 to n:
    for i from 1 to m:
        if S[i] == T[j] :
            subs = 0
        else:
            subs = max(cost(word_type(S[i]), substitute), cost(word_type(T[j]), substitute))
        d[i, j] = min(
                        d[i − 1, j] + cost(word_type(S[i]), delete),
                        d[i, j − 1] + cost(word_type(T[j]), insert)
                        d[i − 1, j − 1] + subs
                        )
return 1 − d[m, n]/max(d[m, 0], d[0, n])
```

图 1 NEWFM 计算流程

Fig. 1 The computation of NEWFM

需要指出的是，NEWFM 在计算复杂度上与 Fuzzy Matching 是相同的，本文实验中对代价函数$cost$的定义为：

$$cost(word\_type, op\_type) = \begin{cases} 2, if\ word\_type = \text{实体} \\ 1, otherwise \end{cases} \quad (2)$$

需要说明的是，本文实验中的代价函数没有按操作类型（插入、删除、替换）做进一步区分，而图 1 给出的 NEWFM 计算流程是可以支持的。

表 1 给出了使用 FM 和 NEWFM 计算匹配度的例子，句子S和T均来自本文实验数据。

表 1 FM 和 NEWFM 计算匹配度示例（实体词字体加粗）

Tab. 1 Examples of similarity calculated by FM and NEWFM (Named entity words are with bold fonts)

编号	句子S	句子T	FM	NEWFM
1.	**abilify** 5 mg tablets **aripiprazole**	**abilify** 5 mg tablets	0.80	0.71
2.	**abilify** 5 mg tablets **aripiprazole**	**abilify** 10 mg tablets **aripiprazole**	0.80	0.86
3.	**sulfobutylether -cyclodextrin** (sbecd) **tartaric acid sodium hydroxide** water for injection	the other ingredients are **sulfobutylether -cyclodextrin** (sbecd) , **tartaric acid** , **sodium hydroxide** , and water for injection .	0.57	0.67

从表 1 的示例中可以看到，对编号 1 和编号 2 的例子，FM 给出了相同的匹配度，而 NEWFM 通过对实体词加权给出了差异度较大的匹配度（0.71 vs. 0.86），而且高分的句子T也确实具有更好的匹配度；对编号 3 的例子而言，因两个句子中匹配的实体词较多，NEWFM 给出了更高的匹配度，根据文献[1]实验匹配度的选择阈值 0.6，FM 会造成句子S得不到该数据增强，而 NEWFM 则可以。

3. 词标签生成和数据增强

3.1 词标签生成

文献[1]使用 Fuzzy Matching 在 TM 库中找到与源语言句最匹配的句子后，会将匹配句对应的译文拼接在源语言句子后。由于匹配句译文会包含一些与源语言句无关词汇，文献[1]通过对比源语言句和匹配句的匹配词、匹配词在其译文句中的对齐词，给出了一种词标签生成方式，以进一步提升翻译效果，具体如图 2 示例。

源语言句： ***How long does*** *a cold* ***last ?***

匹配句： ***How long does*** *the flight* ***last ?***

匹配句译文：***Combien de temps dure*** *le vol* ***?***

源语言句的数据增强：

文本：***How long does*** *a cold* ***last ? \|\| Combien de temps dure*** *le vol* ***?***

标签：S S S S S S SR T T T T RR T

图 2 文献[1]中的词标签示例（匹配词及对齐的译文词字体加粗）

Fig.2 An example of word label in [1] (Matched words and their aligned target words are with bold fonts)

其中，源语言句和匹配句中加粗的词是两个句子最长公共子序列部分，译文句中加粗的词是匹配句中加粗词通过 Fast Align[12]找到的对齐词。在对源语言句进行数据增强时，文本部分是将源语言句和匹配句译文通过连接符（||[1]）拼接起来，词标签分 3 种类型：S 表示源语言句中的词，T 表示与匹配句加粗词对齐的译文词，R 表示其它译文词和连接符。每个词的向量化表示由其文本和标签

[1] 需要保证该连接符没有在词表中出现过。

分别向量化后拼接组成。

不难看出，文献[1]对其词标签的设计初衷是：**通过标签告诉模型哪些译文词可能是有用的**（可直接 *Copy*）。但这种词标签方式需要通过外部程序 Fast Align 来计算，而我们知道，Fast Align 是通过数据的统计信息来计算对齐特征，如果数据不足或与训练数据分布存在偏差，则其效果很可能达不到预期，Song et al.[13]指出其对齐错误率在英罗、英德语种上有 20%~40%，从而影响词标签的准确性，进而会影响数据增强的效果。

因此，能否对词标签重新设计，不依赖外部程序，同时也能满足之前的设计初衷。

对此，我们需要换个角度来看设计初衷，从目标语言调整到源语言，即**通过标签告诉模型哪些源语言词的翻译可能就在拼接的译文中**。本质上也是在给模型提供 *Copy* 的信号，但不需要依赖外部程序，更加简洁，示例如图 3 所示。

源语言句： ***How long does*** *a cold* ***last ?***

匹配句： ***How long does*** *the flight* ***last ?***

匹配句译文：***Combien de temps dure*** *le vol* ***?***

源语言句的数据增强：

文本：***How long does*** *a cold* ***last ?*** **||** ***Combien de temps dure*** *le vol* ***?***

标签：M M M U U M M S T T T T T T

图 3 本文词标签示例（匹配词及对齐的译文词字体加粗）

Fig. 3 Example of word label proposed in this paper (Matched words and their aligned target words are with bold fonts)

其中，加粗词的意义与图 2 相同，词标签分 4 种类型：M 表示源语言句的加粗词（即源语言句和匹配句的最长公共子序列），U 表示源语言句中的其它词，S 表示连接符（||），T 表示匹配句译文词。另外，翻译中亚词（sub-word）的标签直接继承被切分词的标签。

3.2 数据增强

至此，基于上文提出的 NEWFM 和新的词标签生成方式，我们给出应用于神经机器翻译的数据增强流程，具体如图 4 所示。

输入：翻译记忆库 TM、源语言句 S

输出：对 S 数据增强后的结果

步骤：

1. 用 NEWFM 在 TM 中找出与源语言句 S 匹配度最高的句子 M
2. 若两个句子匹配度低于阈值，则返回源语言句 S 和词标签（所有词标签为 U）
3. 若两个句子匹配度高于阈值，则按 3.1 节图 3 生成文本和词标签，并返回结果

图 4 应用于神经机器翻译的数据增强流程

Fig. 4 Process of data augmentation for Neural Machine Translation

4. 实验

4.1 数据集

本文的实验数据采用了文献[1]效果最显著的英法数据集 EMEA （Documents from the European Medicines Agency，可从 http://opus.nlpl.eu 下载），该数据集的基本信息如下：平行句对数 33.7 万；英文句子平均词数 16.8，词汇量 6.3 万；法文句子平均词数 20.3，法文词汇量 6.9 万。数据集划分的

方法与文献[1]类似：随机抽取 2000 句对和 1000 句对分别作为验证集和测试集，剩下句对作为训练集。英文和法文数据分词处理采用 Moses 工具集[14]中的 *tokenizer.perl* 脚本，并使用字节对编码[15]（byte-pair encoding, BPE）在英法数据集上训练了一个联合词表，规模大小为 3.2 万。

翻译记忆库TM采用训练集中的句对，与文献[1]中的方式一样。另外，NEWFM在选择匹配句时的阈值设置为0.6，且过滤与源语言句子完全匹配的匹配句。

数据集EMEA属于医药领域，为此我们设计了基于词典方式的医药命名实体识别，为保证高质量，词典建立采用机器挖掘和人工标注方式进行，命名实体词典规模为1598，在EMEA数据集上实体识别的准确率和召回率都在90%以上。

4.2 实验设置和评价指标

与文献[1]相同，本文的神经机器翻译模型采用了 OpenNMT-tf 工具集[16]中实现的经典 Transformer 框架[7]，参数设置也与文献[1]保持一致，具体如表 2 所示。

表 2. Transformer 模型参数设置

Tab. 2 Parameter settings for Transformer model

参数名	参数值
Word Embedding Size	512
Hidden Layers Size	512
Inner Feed Forward Layer Size	2048
Heads Number	8
Layers Number	6
Beam Size	5
Batch Size	4096

需要说明的是，在采用词标签特征时，最终的词向量是由 508 维文本向量和 4 维标签向量拼接起来的 512 维向量[1]。

我们采用 Lazy Adam 来优化模型，4000 次迭代进行预热且每 8 次迭代更新学习率，最大迭代次数为 10 万。另外，目标语言句子最大长度限制为 100 个词，源语言句子最大长度在有无数据增强时分别限制为 200 个词和 100 个词。

本文采用 BLEU 值作为翻译效果评价指标，使用 Moses 工具集[14]中 *multi-bleu.pl* 脚本进行计算。

4.3 实验结果

我们在 OpenNMT-tf 工具集 Transformer 框架上对比了无数据增强、文献[1] FM 数据增强方法和本文数据增强方法，具体对比模型如表 3 所示（FM+和 NEWFM+中的词标签分别表示文献[1]和本文的词标签生成方法），模型在测试集上的效果如表 4 所示。

表 3. 实验对比模型介绍

Tab. 3 Description of models in experiment

模型名	模型介绍
base	无数据增强 Transformer
FM	文献[1]FM 文本数据增强的 Transformer
FM+	文献[1]FM 文本+词标签数据增强的 Transformer
NEWFM	本文 NEWFM 文本数据增强的 Transformer
NEWFM+	本文 NEWFM 文本+词标签数据增强的 Transformer

需要说明的是，在表 4 中，列“增强比率”表示测试集中可进行数据增强的源语言句子占比，列

“无增强 BLEU 值”表示测试集中无数据增强部分的 BLEU 值，“有增强 BLEU 值”表示测试集中有数据增强部分的 BLEU 值。

表 4. 模型在测试集上的实验效果（加粗表示最好的 BLEU 分数）

Tab 4. Experimental results of models on test set (The best BLEU score is with bold)

模型名	BLEU 值	增强比率	无增强 BLEU 值	有增强 BLEU 值
base	56.99	0%	56.99	0.00
FM	64.21	55.4%	49.02	76.53
FM+	65.02	55.4%	50.08	76.98
NEWFM	64.94	54.6%	51.19	76.54
NEWFM+	**65.54**	54.6%	51.81	77.10

从表 4 中的实验结果来看，通过数据增强对 base 模型的 BLEU 值提升明显，这与文献[1]中的结论一致。从 FM 与 NEWFM 的结果对比来看，我们通过对医药领域实体词加权，获得了更好的结果，BLEU 值有 0.73 的提升；在 NEWFM 基础上，加入本文设计的词标签特征，在 BLEU 值上进一步提升了 0.60，表明了该标签对提升翻译效果的有效性。从 FM+与 NEWFM+的结果对比来看，二者在测试集的增强比率上基本相当（NEWFM+略低），但 NEWFM+在测试集上取得了更好的 BLEU 值，而且词标签的计算相比 FM+更加简洁，没有外部程序依赖，体现出该方法具备一定的优势。

另外，我们还对比了 FM 和 NEWFM 在训练集（随机抽取 1 万句）、验证集和测试集上增强数据的差异性，即这 2 个方法找到的最匹配句子有多少是不同的（选择阈值都设置为 0.6），具体结果如表 5 所示。

表 5. FM 和 NEWFM 在不同数据集上的差异性

Tab. 5 Match differences of FM and NEWFM on different data sets

数据集名称	数据集规模	FM 增强数量	NEWFM 增强数量	增强差异数量
训练集（随机 1 万）	10000	5714	5627	629
验证集	2000	1130	1109	116
测试集	1000	554	546	63

从表 5 中可以看到，NEWFM 数据增强的数量略低于 FM，这是因为在源语言句子或 TM 句子包含实体词却没有匹配的情况下，NEWFM 得到的匹配度会低于 FM，相同选择阈值下，NEWFM 召回的数量会少于 FM，但从各个数据集上的增强比率来看，基本相当。NEWFM 和 FM 在各个数据集上的增强差异，约占增强数量的 11%，若引入更好的实体识别技术和更大规模的 TM 数据，NEWFM 相比 FM 则会有更大的差异，提升效果也将更加显著。

5. 结论

神经机器翻译对数据的依赖性很强，在数据量有限的情况下，通过数据增强可进一步提升翻译效果。本文从文献[1]的 Fuzzy Matching 方法出发，基于领域命名实体知识，提出了一种基于命名实体加权的匹配度计算方法 NEWFM，以加强领域命名实体在数据增强中的权重；同时针对文献[1]数据增强中词标签依赖外部程序可能存在准确率不足的问题，将思考角度从“可参考哪些目标语言词”转变为“哪些源语言词有参考”，提出了一种简洁且有效的词标签生成方式。我们在文献[1]效果最显著的英法医药数据集 EMEA 上进行了实验对比，结果表明用本文提出的 NEWFM 和词标签进行数据增强，取得了更好的翻译效果。

如何从非结构化的翻译语料数据中提取出结构化知识，通过知识来提升神经机器翻译的效果，是一个值得探索的方向。本文尝试将命名实体知识应用到数据增强中，取得了一定的结果，未来我们希望将知识应用到模型网络设计中，进一步提升翻译效果。

6. 参考文献

[1] Xu J, Crego J, Senellart J. Boosting Neural Machine Translation with Similar Translations[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5 - 10, 2020: Association for Computational Linguistics, 2020: 1580-1590.

[2] Koehn P, Senellart J.Convergence of Translation Memory and Statistical Machine Translation[C]. Proceedings of AMTA Workshop on MT Research and the Translation Industry, 2010, Denver CO: 21-31.

[3] Plitt M, Masselot F. A productivity test of statistical machine translation post-editing in a typical localisation context[J]. The Prague bulletin of mathematical linguistics, 2010, 93: 7-16.

[4] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014: 3104–3112.

[5] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation[J]. CoRR, 2016, abs/1609.08144.

[6] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning[C]. Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August, 2017: 1243–1252.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA: 6000–6010.

[8] Chu C, Wang R. A survey of domain adaptation for neural machine translation[C]. In Proceedings of the 27th International Conference on Computational Linguistics, 2018, Santa Fe, New Mexico, USA. Association for Computational Linguistics: 1304-1319

[9] Gu J, Wang Y, Cho K, et al. Search engine guided neural machine translation[C]. In Sheila A. McIlraith and Kilian Q. Weinberger, AAAI, 2018: 5133–5140

[10] Hokamp C, Liu Q. Lexically constrained decoding for sequence generation using grid beam search[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, Vancouver, Canada: Association for Computational Linguistics: 1535–1546

[11] Bult´e B, Tezcan A. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation[C]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, Florence, Italy: Association for Computational Linguistics: 1800–1809

[12] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2[C]. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia. Association for Computational Linguistics: 644–648

[13] Song K, Wang K, Yu H, et al. Alignment-Enhanced Transformer for Constraining NMT with Pre-Specified Translations[C]. The 34th AAAI Conference on Artificial Intelligence, February 7-12, 2020, NY, USA: 8886-8893

[14] Koehn P, Hoang H, Birch A, et al. Moses: Open source toolkit for statistical machine translation [C]. ACL 2007,Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech, 2007: 177-180

[15] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[C]. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, 2016: 1715–1725

[16] Klein G, Kim Y, Deng Y, et al. OpenNMT: Opensource toolkit for neural machine translation[C]. Proceedings of ACL 2017, System Demonstrations, July, 2017, Vancouver, Canada, Association for Computational Linguistics: 67-72.

A Named Entity Knowledge Data Augmentation Method for Neutral Machine Translation

ZHANG Min, YANG Hao*, TAO Shimin, WANG Minghan, GUO Jiaxin,

CHEN Yimeng, SU Chang, DU Chunning, WANG Yuxia, QIN Ying

(2012 Lab, Huawei, Beijing, 100095)

Abstract: As Neutral Machine Translation (NMT) relies on training data seriously, various data augmentation methods are proposed to make better use of the data. In this paper, from the Named Entity Knowledge aspect, based on the method in [1], a Named Entity Weighted Fuzzy Matching (NEWFM) method is proposed for the sentence similarity calculation, which is a very important part in data augmentation of NMT. Meanwhile, as the word label method in [1] relies on the external program and may produce incorrect results, a simple yet effective word label method is given. Based on the NEWFM and the word label method proposed in this paper, the process of the data augmentation method for NMT is given. Experimental results on the English-French medicine corpora shows the effectiveness of our method.

Keywords: neural machine translation; data augmentation; named entity; fuzzy matching; word label