Computer Science and Application 计算机科学与应用, 2021, 11(7), 1923-1930
Published Online July 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.117197

Hans汉斯

面向严格对齐任务的文本自动生成：以招标技术范本为例

卢　爽

中国神华国际工程有限公司，北京
Email: 13707502@chnenergy.com.cn

收稿日期：2021年6月15日；录用日期：2021年7月12日；发布日期：2021年7月19日

摘　要

自动生成的严格对齐的文本，生活中更有常用，例如：自动生成对齐的招投标文件等。然而，自动生成对齐文本时，首先需要的是结构化数据。本文设计了基于历史招标文件的严格对齐文本自动生成模型。方法包括：基于正则匹配的数据清洗和结构化关键标签的抽取(例如：招标文件的技术参数等)；基于k-means的结构化关键标签聚类；基于word2vec计算词向量之间余弦距离的结构化关键标签去重；最后，基于结构化关键标签，预测出最终的编制范本。实验以专家手工标记的100篇招标文件技术范本为参照，文中算法不仅可以达到与专家人工编制范本之间80%以上的重合度，同时参数覆盖更全面，鲁棒性高，可以满足生产需求。

关键词

对齐文本自动生成，关键标签抽取，文本去重

Automatic Text Generation for Strictly Aligned Tasks: Taking the Tendering Technical Template as an Example

Shuang Lu

China Shenhua International Engineering Co., Beijing
Email: 13707502@chnenergy.com.cn

Received: Jun. 15th, 2021; accepted: Jul. 12th, 2021; published: Jul. 19th, 2021

文章引用：卢爽. 面向严格对齐任务的文本自动生成：以招标技术范本为例[J]. 计算机科学与应用, 2021, 11(7): 1923-1930. DOI: 10.12677/csa.2021.117197

Abstract

Automatic generation of strictly aligned text, is more commonly used in life, such as: automatic generation of aligned bidding documents, etc. However, when you automatically generate aligned text, you need structured data first. In this paper, a strict alignment text automatic generation model based on historical bidding documents is designed. The methods include: data cleaning based on regular matching and extraction of structured key labels (such as technical parameters of bidding documents); Structured key label clustering based on k-means; Structured key tag deduplication based on word2vec to calculate cosine distance between word vectors; Finally, based on the structured key label, the final compilation template is predicted. The experiment takes 100 technical templates of bidding documents manually marked by experts as reference, and the algorithm in this paper can not only achieve more than 80% coincidence degree with the manual templates compiled by experts, but also have more comprehensive parameter coverage and high robustness, which can meet the production requirements.

Keywords

Align Text Automatically Generated, Key Label Extraction, Text De-Duplication

1. 引言

对齐文本的自动生成，实际生活中有着更丰富的应用场景。例如：招标文件的编写，尤其是涉及到专业领域的技术部分，往往会耗费大量人力和时间，同时，还可能需要有该方向领域专家作把关。能够自动产生一份风格一致，主要相关标的物涉及的内容一致的招标文件，对招标公司与投标公司都是很重要的。然而，对齐的文本的产生，需要结构化的数据。如何面向历史招标文本自动化抽取出结构化数据，再基于结构化的数据，自动生成对齐文本，是研究的难点。此外，招投标文件技术部分通常拥有大量参数，参数的填写与要求的陈述是技术部分的主体部分。因此将参数作为文本抽取时关键的核心结构化数据，是非常重要的。

本自动生成是指计算机借助于自然语言处理技术以及语言知识，在潜在的非语言形式信息的基础上，自动地生成报告、新闻、摘要等文本信息[1] [2]。按照不同输入的划分，文本自动生成可包括文本到文本的生成(text-to-text generation)、意义到文本的生成(meaning-to-text generation)、数据到文本的生成(data-to-text generation)以及图像到文本的生成(image-to-text generation) [3]。文本到文本的生成主要包括文本摘要、文本复述等任务等。文本摘要生成目前主要是基于传统机器学习方法，如 TextRank 算法和 Seq2Seq 模型。由于 seq2seq 模型生成的文本摘要，具有不准确以及词句重复的缺陷，谷歌联合斯坦福提出了 Pointer-Generator Network 对上述问题进行了改进[4]。文本复述任务目前主要有基于机器翻译的复述生成方法，将成熟的统计机器学习模型和系统应用到复述生成问题上来[5] [6]。Ehud Reiter 提出了数据到文本生成系统的一般框架，分为信号处理、数据分析、文档规划及文本实现四个步骤[7]。

本文首先设计基于正则匹配的数据清洗和结构化关键标签数据的抽取(例如：招标文件的技术参数等)；基于 k-means 的结构化关键标签聚类；基于 word2vec 计算词向量之间余弦距离的结构化关键标签去

重；最后，基于结构化关键标签，预测出最终的文本。针对招标文本范本的自动编制的例子，首先设计正则表达式，提取可能的参数项；之后，对提取结果进行清洗，去重，得到最终的参数库(结构化数据)；最后自动生成范本；另外，对专家所遗漏的技术参数内容加以补充，调整输出阈值，提高对新添加的技术参数条目的敏感度，实现范本的增量更新。

论文结构如下：1 是文章的介绍；2 是文章的相关的理论背景部分；3 是本文的模型方法部分；4 是实验与结果分析；5 是全文小结。

2. 相关理论

2.1. 正则表达式

正则表达式是一种用于描述文本搜索符号串的语言，正则表达式要求有一个试图搜索的模式字符串以及搜索的目标字符串[8]。被常用来检索、替换符合指定模式字符串。而未来提取出招标文件技术部分中的特定的内容以及生成智能化的招标文件中的技术内容范本，正则表达式发挥了关键性的作用。

为了生成范本以及提取特定需要的内容，需要通过运用正则表达式来寻找它们之间的相同点，即运用正则表达式来概括各个文件特定内容的规律。

2.2. 余弦距离 (Cosine Distance)

余弦距离通过测量两个向量的余弦值来度量它们之间的距离，在信息检索等文本挖掘领域，有很多的应用[9]，同时也证明了它的有效性，计算公式如下所示：

$$\begin{aligned}\text{similarity} &= \frac{A \cdot B}{\|A\| \cdot \|B\|} \\ &= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}\end{aligned} \tag{1}$$

给定两个向量 A、B，Cosine 系数由 A、B 的向量的点积大小与 A、B 的长度乘积的比值计算求得。

2.3. word2vec

word2vec 是一种将文本转为向量表示的方案，包括 Skip-gram 和 CBoW 两种神经网络模型，可以在大量文本上进行高效训练，并得到特征词的向量表示形式，与其他文本转向量方案相比，word2vec 生成的词向量包含了文本上下文中大量的语义和语序信息，在文本相似度等上游任务中得到了大量应用[10]，且取得了不错的效果。

3. 面向对齐文本的自动生成模型

本文提出的面向对齐文本的自动生成方法可分为基于正则匹配的数据清洗、基于正则匹配的核心标签预提取、基于 k-means 的核心标签聚类、基于 word2vec 计算词向量之间余弦距离的核心标签去重、基于结构数据的文本生成五个步骤。(如下图 1 所示)

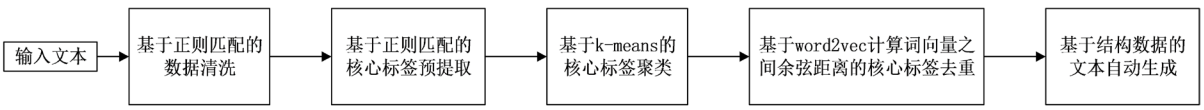

Figure 1. Automatic preparation process flow chart of bidding technical document templates
图 1. 招标技术文件范本自动编制方法流程图

3.1. 基于正则匹配的数据清洗

历史文本除了核心标签这一部分内容之外，还有许多与文本主题的相关的内容。而对核心标签提取无关的内容需要通过数据清洗来去掉[11]。详细步骤如下：

1) 数据格式统一。将 doc、docx 格式文件，借助 office 工具进行转换成 pdf 格式。

2) 例如时间等标签信息提取。标签信息可能携带时间等信息。例如：需要得到核心标签信息随时间更新的情况时，时间越靠前的标签，在自动生成的文本中，权重越高。例如：在招标文件的时间等参数的提取，该部分借助 PyPDF2，抽取 PDF 中招标公告内容，并进一步通过正则匹配公式(\d{4}-\d{2}-\d{2}\s\d{1,2}:\d{1,2}\)获取招标截止时间或开标时间内容作为参数时间戳信息，如图 2 所示。

```
{'Producer': 'Microsoft® Word 2013',
 'Title': '无标题文档',
 'Author': 'NTKO',
 'Creator': 'Microsoft® Word 2013',
 'CreationDate': "D:20190522112531+08'00'",
 'ModDate': "D:20190522112531+08'00'",
 'BidDate': '2019-06-12 09:30'}
```

Figure 2. Time information extraction content
图 2. 时间信息提取内容

3) 截取历史文本部分内容并进行数据清洗，去除页眉页脚内容，以文本格式存储，作为模型的训练样本。

3.2.基于正则匹配的核心标签的预提取

通过上述步骤得到了文本的核心部分之后，紧接着使用正则匹配，提取可能为核心标签的文本内容：

(1)匹配核心标签类别。例如：通过事先对历史招标文件文本的分析，和新标签常以“数字.参数类别”的格式出现。

(2)匹配核心标签内容。例如：招标文件的核心标签中以“参数名：参数值”格式出现，其中，由于编写规范的不同，会有“：”或“:”等类型，需要进行冗余匹配，除了以上情况，参数或要求还有可能以完整句子出现，该类型，我们以参数名保存整个句子，由下一步做进一步处理。

(3)以(参数类型、参数名、参数值、时间戳、抽取文件)为格式，对核心标签进行整理，保存为 json 格式，如图 3 所示。

```
"parameter_type": "采煤机主要参数",
"parameter_name": "生产能力",
"parameter_value": "不小于2500t/h",
"time": "2018年 12 月 04 日 09:00",
"parameter_belonging_to": "采煤机",
"filename": "00611434-ed85-42c0-a4af-59494754cbf7.pdf"
```

Figure 3. Example of core tag extraction results
图 3. 核心标签提取结果示例

3.3. 基于 k-means 的核心标签聚类

对上一步的处理结果核心标签结构数据集，首先对核心标签类型进行文本聚类[12]，抽取近似核心标签类型集合。例如：“招标人提出的特别技术要求”与“招标人提出的其它技术要求”在本算法中视作

同类参数(核心标签)，之后以参数类型(核心标签类型)为标准使用 k-means 算法进行聚类，形成多个参数簇(核心标签簇)。对每个参数簇进一步处理，主要包括同义词的替换，如“其他”与“其它”，“”与“”等，以及去除可以完全匹配的重复参数项，同时要记录其匹配次数。

3.4. 基于 word2vec 计算词向量之间余弦距离的核心标签去重

在实践过程中，观察得到因为编写习惯的不同，尤其是文本编写时间跨度较大的类型，如招标文件的采煤机等，普遍存在不同结构，但同义的参数文本。如果不进行去重，会造成参数的冗杂，影响输出效果。

我们设计了面向短文本的去重方案。分为两步：基于内容匹配去重，以及基于特征匹配去重。在去重过程中，同样要记录其匹配次数，用于后续参数输出期望计算步骤。

基于内容匹配去重，主要针对可以完全匹配的文本核心标签项或只有标点符号等停用词不同的内容。例如：招标文件中的“主要部件大修周期”与“主要部件的大修周期”等。

基于特征匹配去重，则是通过提取文本的结构特征或语义特征作为去重标准。本方法选取了 TF-IDF [13]结合余弦距离、word2vec 结合余弦距离、词袋模型[14]结合杰卡德距离[15]三种方法进行了实验对比，最终选取 word2vec 方法作为去重的方案，相比于其他两类方案，word2vec 不仅考虑了结构特征，同时蕴含了语义特征，因此去重效果较好。

3.5. 基于核心标签结构化数据的文本自动生成

通过上一步的匹配去重，最终得到了以核心标签类型为分割的多个核心标签簇。例如：在该步骤，以不同权重分别计算的提取时间，以及核心标签项在去重阶段统计的匹配次数的结果并求和；求和结果作为该项参数的可输出期望值；最后，将可输出期望值超出指定阈值的核心标签以 doc 文件格式进行输出。如图 4 所示。

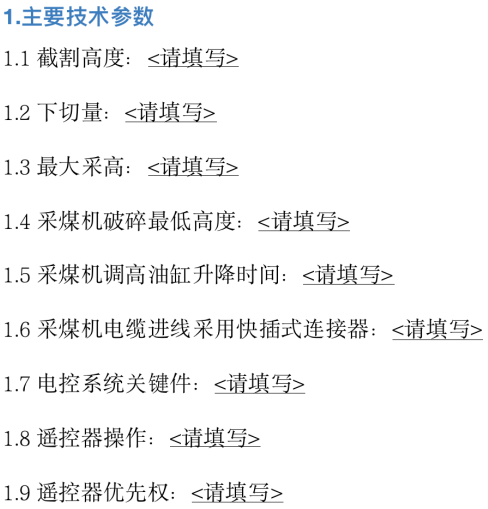

Figure 4. Screenshot of output result of shearer sample in bidding document
图 4. 例如招标文件采煤机范本输出结果部分截图

4. 实验验证

4.1. 实验设置

为了验证本方法的有效性和先进性，我们采用招标文件的自动生成作为案例分析。

本实验选取由 100 类原始招标文件集作为实验数据输入，其中有大量结构相同或语义相同的内容。实验分为三组，分别为 TF-IDF 结合余弦距离、word2vec 结合余弦距离、词袋模型结合杰卡德距离。实验参数设置如表 1 所示。

Table 1. Experimental parameter setting table
表 1. 实验参数设置表

参数名	参数值质量	参数说明
title_threshold	0.4	标题聚类相似度阈值
parameter_threshold	0.5	参数去重相似度阈值
xrate_importance	0.3	参数出现频次权重
date_importance	0.7	参数出现时间权重
num_features	100	特征向量维度

4.2. 实验指标

本实验以重合度、覆盖度、冗余度作为性能指标。

最终输出的内容与 100 篇对应的专家人工编制的范本进行对比，计算结果范本技术参数条数与专家编制范本技术参数条数的重合数量占比得到重合度 O，如公式 2 所示。重合度越高，表示自动编制的范本更加接近实际生产需求。

$$O=\frac{\sum_{f\in\{T,F\}}^{F}1}{F}\times 100\% \tag{2}$$

其中，f 代表结果范本中的技术参数条目，F 为结果范本技术参数条目集合，T 为专家编制范本技术参数集合。

通过观察原始数据集，对照结果范本内容，统计人工编制范本遗漏但结果范本包含的条目并计算其占总技术参数条目的比例得到覆盖度 C，如公式 3 所示。覆盖度指标越高，表示结果范本在对技术参数条目的覆盖上更加全面。

$$C=\frac{\sum_{f\in\{F,M\},f\notin T}^{F}1}{F}\times 100\% \tag{3}$$

其中，M 代表有效参数集合。

通过统计三种算法出现的重复条目、无效条目数所占总技术参数条目的比例作为冗余度 R，如公式 4 所示，冗余度越低，则代表该方案去重效果越好。

$$R=\frac{\sum_{f\in F,f\notin M}^{F}1}{F}\times 100\% \tag{4}$$

4.3. 实验结果

第一组实验，采用 TF-IDF 模型，首先将文本进行分词处理，得到各词的 TF-IDF 值，用得到的值将文本转为向量，并计算两段文本向量的余弦距离。

第二组实验采用由原标书抽取出的技术部分文本作为训练集的 word2vec 模型，将待比较文本中的词语转为向量，求取向量的平均值，作为整个文本的特征向量，并计算两个特征向量之间的余弦距离。

第三组实验先由两段文本得到词袋模型，以词袋模型将文本转为特征向量，并计算两个特征向量之间的杰卡德距离。

三组实验的结果如下表 2 所示。

Table 2. Experimental results
表 2. 实验结果

实验类型	平均重合度 O/%	平均覆盖度 C/%	平均冗余度 R/%
TF-IDF + Cosine distance	82.3%	8.6%	23.3%
word2vec + Cosine distance	81%	13.4%	14.6%
Bow + Jaccard distance	75%	10.2%	20.4%

4.4. 结果分析

从实验结果可以发现，基于 TF-IDF 结合余弦距离模型生成的技术范本，具有更高的平均重合度，说明其与专家人工编制的范本相同的技术参数条目比例更高，但其平均冗余度也是三类方案中最高的，通过观察，该方案生成的范本体积通常也是最大的，去重效果较差，包含了太多无效条目。同时，其平均覆盖度也是较低的，对参数较不敏感。

基于 Bow 结合杰卡德距离方案生成的技术范本与专家人工编制范本平均重合度较低，对 100 组实验观察，其重合度一直在 80%以下波动，而平均冗余度与 word2vec 模型方案相比较高，包含了较多的无效条目与重复条目，平均覆盖度也较低，虽然较 TF-IDF 模型而言更敏感，但与 word2vec 模型相比较差。

基于 word2vec 结合余弦距离模型生成的技术范本与 TF-IDF 模型方案相比，具有相近的平均重合度，通过对 100 组实验观察，TF-IDF 模型方案的重合度方差较大，鲁棒性较差，而 word2vec 模型方案则一直在 81%上下波动，鲁棒性较强，同时其平均覆盖度最高，对技术参数较敏感，平均冗余度最低，去重效果明显。

5. 结束语

本文提出一种面向对齐文本自动生成方法。为了验证方法的有效性、先进性和鲁棒性，以招标文件的范本自动生成为案例分析。通过对大量招标文件的分析，提取标的物技术部分参数内容，形成招标文本的结构化数据。本算法使用正则匹配对数据进行清洗以及初步提取，基于 word2vec 文本相似度模型对同类参数进行聚类，对同类章节参数，使用该模型进行去重及统计，最后以一定的权重系数，计算参数的可输出期望值，以该值为依据，输出最终的技术参数，实现招标文件范本的自动编制。本算法可以达到 80%以上的与专家人工编制范本的重合度，同时，可以对专家所遗漏的技术参数条目进行一定的补充，可以满足生产的需求。同时，通过 10 次和当前其他方法的比较试验，验证了方法的鲁棒性和先进性。

参考文献

[1] Reiter, E., Dale, R. and Feng, Z. (2000) Building Natural Language Generation Systems. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511519857

[2] 刘挺. 人机对话浪潮: 语音助手、聊天机器人、机器伴侣[J]. 中国计算机学会通讯, 2015, 11(10): 54-56.

[3] 万小军, 冯岩松, 孙薇薇. 文本自动生成研究进展与趋势[C]//中国计算机学会. CCF2014-2015 中国计算机科学技术发展报告会论文集. 2015: 298-323.

[4] See, A., Liu, P.J. and Manning, C.D. (2017) Get to the Point: Summarization with Pointer-Generator Networks. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Van-

couver, July 2017, 1073-1083. https://doi.org/10.18653/v1/P17-1099

[5] Finch, A., Hwang, Y.S. and Sumita, E. (2005) Using Machine Translation Evaluation Techniques to Determine Sentence-Level Semantic Equivalence. *Proceedings of the IWP*, Jeju, 14 October 2005, 17-24.

[6] Dan, F. (2002) On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering*, **6**, 15-28. https://doi.org/10.1017/S1351324900002370

[7] Reiter, E. (2007) An Architecture for Data-to-Text Systems. *Proceedings of the Eleventh European Workshop on Natural Language Generation*, Saarbrücken, June 2007, 97-104. https://doi.org/10.3115/1610163.1610180

[8] Jeong, W.S., Lee, C., Kim, K., *et al*. (2020) REACT: Scalable and High-Performance Regular Expression Pattern Matching Accelerator for In-Storage Processing. *IEEE Transactions on Parallel and Distributed Systems*, **31**, 1137-1151. https://doi.org/10.1109/TPDS.2019.2953646

[9] Gibrael, A. and Hadi, O. (2021) Using Residual Networks and Cosine Distance-Based K-NN Algorithm to Recognize On-Line Signatures. *IEEE Access*, **9**, 54962-54977. https://doi.org/10.1109/ACCESS.2021.3071479

[10] Mohamed, E.H. and El-Behaidy, W.H. (2021) An Ensemble Multi-Label Themes-Based Classification for Holy Qur'an Verses Using Word2Vec Embedding. *Arabian Journal for Science and Engineering*, **46**, 3519-3529. https://doi.org/10.1007/s13369-020-05184-0

[11] Bertossi, L., Kolahi, S. and Lakshmanan, L. (2013) Data Cleaning and Query Answering with Matching Dependencies and Matching Functions. *Theory of Computing Systems*, **52**, 441-482. https://doi.org/10.1007/s00224-012-9402-7

[12] Jabi, M., Pedersoli, M., Mitiche, A. and Ayed, I.B. (2021) Deep Clustering: On the Link between Discriminative Models and k-Means. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 1887-1896. https://doi.org/10.1109/TPAMI.2019.2962683

[13] Zhang, L. (2021) Research on Case Reasoning Method Based on TF-IDF. *International Journal of System Assurance Engineering and Management*, **12**, 608-615. https://doi.org/10.1007/s13198-021-01135-6

[14] Chen, H., Zhang, B., Sun, F., Huang, Y. and Yuan, J. (2020) Incremental Scene Detection in Outdoor Environment Based on Hierarchical Bag-of-Words Model. *Control Theory & Applications*, **37**, 1471-1480.

[15] Liu, X., Di, H., Yang, W., Lin, P. and Wang, S. (2020) Mosaic of Cultural Relics Fragments Based on SURF Feature Extraction Descriptor and Jaccard Distance. *Optics and Precision Engineering*, **28**, 963-972.