基于词频-逆文档频率的乌兹别克语文本摘要-

Khabibulla Madatov[1], Shukurla Bekchanov[2], **Jernej Vičič[3]**

[1]乌尔根奇州立大学，Kh. Alimdjan街14号，乌尔根奇市，邮编220100，乌兹别克斯坦 habi1972@mail.ru[2]乌尔根奇州立大学，Kh. ~~Alimdjan街14号~~，乌尔根奇市，邮编220100，乌兹别克斯坦shukurla15@gmail.com[3]斯洛文尼亚科学与艺术学院研究中心，Fran Ramovš研究所，Novi trg 2，1000卢布尔雅那，斯洛文尼亚。普里莫尔斯卡大学，FAMNIT，Glagoljaska 8，6000 科佩尔，斯洛文尼亚jernej.vicic@upr.si

摘要

随着互联网和电子信息服务的迅速发展，信息量正以惊人的速度增长。受时间限制，我们没有机会阅读所有这些信息。即使是分析与某一领域相关的文本数据，也需要投入大量工作。文本摘要有助于解决这些问题。本文介绍了一项针对乌兹别克语文本摘要任务的实验，其方法基于词频-逆文档频率算法进行文本抽象摘要。利用这一密度函数，可以提取文本中语义上重要的部分。我们对全文的重要部分应用 n元语法方法，从而对给定文本进行摘要。作者使用了一个专门手工构建的语料库，名为“School语料库”，以评估所提方法的性能。结果表明，该方法能够有效地从乌兹别克语文本中提取摘要，并有望应用于信息检索和自然语言处理等多个领域。总体而言，这项研究为资源匮乏语言文本摘要领域不断发展的研究成果作出了贡献。

关键词：文本摘要、乌兹别克语、词频-逆文档频率、School语料库。

1. 引言

文本摘要的任务是在保留最重要信息的同时，将给定文本压缩为更短的版本。互联网、网页、新闻、文章、状态更新、博客，以及每天都有新发现的科学家们的研究成果等，都是海量文本文件的来源。日常生活中生成和传播的信息数量十分庞大，这带来了许多问题：我们必须找到、使用并处理所需的文本信息。在这种情况下，我们面临着对文本数据进行分析的文本摘要问题。文本摘要是自然语言处理（NLP）最重要的应用之一。

自动文档摘要具有以下优势：

1）缩短数据学习时间；
2）加快文档分析过程；
3）便于在查阅文档时选择文档；

4）比传统的文本摘要提供更多摘要选项。

自动摘要的最低要求是：自动摘要所得文本的词数不超过原始文本词数的约 30%（Torres-Moreno，2014 年，第 33 页）。

自动文本摘要是一个复杂的过程，由多个阶段组成。第一阶段是对文本进行初始处理，通常包括以下步骤：

- 将文本拆分为部分、句子、段落等；- 对片段进行句法分析，将其拆分为单词或进行分词；

- 对单词进行标准化（词形还原、词干提取等）；

- 删除停用词。

初始处理是一项困难的任务，很大程度上取决于文本所使用的语言。例如，句子边界由标点符号标记，而标点符号的用法因语言而异。此外，并非所有语言中的单词都以空格分隔。用中文、日语或韩语等东方语言书写的文本，与用欧洲拉丁字母书写的文本之间存在很大差异。

这一问题也会出现在其他步骤中，例如词形还原或词干提取，以及根据语法词性标注对单词进行规范化。即使是移除停用词[1] (X. A. Madatov 等，2021)，也取决于自然语言，同样并非一项简单的任务。

本文探讨基于 TF-IDF 算法（Aizawa，2003）的乌兹别克语文本摘要。尽管资源丰富型语言的摘要问题已通过多种方法得到解决，但乌兹别克语等低资源语言的文本摘要问题仍未解决。这正是本研究开展基于 TF-IDF 的文本摘要研究的主要动机。在此过程中，我们从文本中包含的词语创建词典，并根据这些词语的 TF-IDF 值建立概率分布律。为了评估所提方法的结果，我们使用了名为“School语料库”的语料库，该语料库由免费提供的学校教材构建而成，例如《阅读书》《母语》和《文学》。

[1] 如果从文本中删除这些词不仅不会改变上下文含义，还能保留仍足以承载该如果考虑上下文的含义，那么这类词就可以称为本研究中的停用词。

乌兹别克语。乌兹别克语属于突厥语族语言，在乌兹别克斯坦及其周边的中亚国家有超过 3000 万人使用。它是乌兹别克斯坦的官方语言，使用拉丁字母书写。乌兹别克语历史悠久、文化丰富，其语言结构也深受阿拉伯语、波斯语和俄语等语言的影响。乌兹别克语有多种方言，主要包括卡拉卡尔帕克语、花剌子模方言和撒马尔罕方言等。尽管乌兹别克语十分重要，但针对该语言文本摘要的研究相对较少。我们的方法旨在填补这一空白，并提供一种实用的乌兹别克语文本摘要工具。

本文结构如下：首先，第 1 节介绍研究背景。第 2 节介绍相关研究。本文的主要贡献以方法论的形式呈现在第 3 节。随后，第 4 节介绍实验与结果，其中包含本研究所得结果的相关信息。最后，我们在第 5 节对本文进行讨论并作出结论。

2. 相关研究

自动文本摘要领域于 20 世纪 50 年代通过 Hans Peter Luhn 的研究逐步发展起来。Luhn 首次建立了科技文章摘要模型（Luhn，1958）。受其工作的启发，许多其他研究人员也取得了大量有价值的研究成果，例如开发了运用句子选择与排除技术的自动摘要系统，以及词汇控制列表（Edmundson 和 Wyllys，1961）；这些方法生成了高质量摘要，值得进行大规模测试（Rush 等，1971）。

Pollock 和 Zamora 研究了化学文摘社关于文本摘要的研究，该研究采用了经修改的 Rush-Salvador-Zamora 算法。他们发现，有些主题更适合自动抽取，并建议针对范围较窄的主题领域定制算法，以获得更好的结果；他们还探讨了自动抽取的可行性（Pollock & Zamora, 1975）。

近年来，研究人员提出了各种方法来提升文本摘要的效果，从将 TF-IDF算法与潜在狄利克雷分配（LDA）算法结合以生成信息量更丰富的摘要，到采用神经网络的方法，利用注意力机制捕捉文本的语义信息并提升摘要质量（Lehnert & Ringle, 2014）。

以下研究涉及乌兹别克语自然语言处理（K. Madatov, 2019; K. A. Madatov et al., 2022）。

关于乌兹别克语词语之间的关系，已有研究以 School语料库为例，探讨了填空任务以及乌兹别克语停用词的抽取，相关内容见（K. Madatov et al., 2022b）和（K. Madatov et al., 2022c）。乌兹别克语停用词检测的准确度见（K. Madatov et al., 2022a）；（Sharipov & Sobirov, 2022）讨论了乌兹别克语词干提取和词形还原。这些都是本研究采用的基础方法。

尤其是近年来，乌兹别克语领域涌现出越来越多与自然语言处理相关的研究成果和资源，其中包括机器转写工具（Salaev 等，2022a）、情感分析数据集以及分析模型（Kuriyozov 等，2022），以及语义评估数据集（Salaev 等，2022b）等。

3. 方法论

乌兹别克语因此大多数用于语言文本摘要的方法无法直接用于乌兹别克语。

本文的科学创新之处在于：为 School语料库创建一种基于词频-逆文档频率的乌兹别克语文本摘要方法（K. Madatov 等，2022c）。

文献（K. Madatov 等，2022a）研究并证明了这样一个问题：在 School语料库中寻找包含停用词的乌兹别克语文本片段。利用该方法，可以找到给定文本中的重要部分。下面，我们介绍乌兹别克语文本摘要方法的算法。算法1简要描述了所提出方法实施过程中采取的步骤，完整算法见附录1。

算法1：乌兹别克语文本摘要

1. 将给定文本（初始文本）分解为词，后文称为文本2。使用从School语料库中提取的数据集移除停用词。根据处理后的文本创建唯一词典，后文称为文本 UW ^{4} $_{i}$ $^{-}$。对于属于文本_UW 的每个 a，计算 w_i=词频-逆文档频率(a_i)。随后使用公式 $p_i=w_i/\Sigma w_i$，将 w_i 转换为 $p_i(a)$。为文本 UW 创建密度函数。此步骤包括以下计算过程：（K. Madatov 等，2022a，第4–5页）$E=\Sigma i\cdot p_i$——唯一词的数学期望；$D=\Sigma(i-E)^2\cdot p_i$——唯一词的离散度；D——唯一词的标准差；$E_k=\Sigma p_i\cdot i^k$ ——唯一词的$\mu_3=E_3-3\cdot E_1\cdot E_2+2\cdot E_1^3$ k——唯一词的三阶中心矩；$A_s=\mu_3/\sigma^3$——理论分布的偏度

5. 文本 UW^{-}的重要部分取决于 A_s。令 $k=E-\sigma$ ，且 $m=E+\sigma$。我们分别找到文本 UW 中处于 k 位和 m 位的词，以及这些词在初始文本中的$^{-}$位置

文本。令它们分别为 k_1 和 m_1 依次。6. a)如果 $A_s>0$，则将文本中从首词到第 k_1 个词的部分重新载入文本，并删除其余部分；b)如果 $A_s<0$，则将文本中从第 m_1 个词到末词的部分重新载入文本，并删除其余部分；c)如果 $A_s=0$，则将文本中从第 k_1 个词到第 m_1 个词的部分重新载入文本，并删除其余部分。7. 对文本应用 3-gram 方法。8. 输出结果。

4. 实验与结果

该方法介绍了如何将乌兹别克语文本摘要算法应用于给定文本，以解决基于词频-逆文档频率从给定文本中提取重要部分并进行摘要这一复杂问题。本文这一部分将展示乌兹别克语文本摘要算法的应用示例。

实验选取 Kh. Tokhtabayev 的历险小说《骑黄巨人》作为示例。这部杰作的部分内容收录于八年级文学教材中。

将乌兹别克语文本摘要算法应用于该文本的过程可按以下步骤说明：

1. 这部杰作由49,705个词元组成，其中13,740个为唯一词。2. 删除停用词（K. Madatov 等，2021）。结果得到文本。3. 创建文本的词典。该词典称为唯一词，这些词的集合记作Text_UW。在此情况下，len(Text_UW)=13740。4. 每个唯一词记作a_i；$i\epsilon[1...13740]$。对于每个a_i，计算w_i=词频-逆文档频率(a_i)。词 a_i 的概率可使用以下公式计算：$p_i=w_i/\sum w_i$

唯一词概率的数值列于表1。

表 -1：唯一词概率的数值

变量	Item
E	4379,22
D	16019213,55
Σ	4002,4
σ^3	64115315874
E_1	4379,22
E_2	35196780,35
E_3	3,40214E+11
μ_3	45776660238
A_s	0,714

5. K=377。6. 在我们的案例中 As>0。K 在文本 UW 中对应的词是“shavla”。我们找到_的首次出现文本中的单词“shavla”。我们从文本的第一个单词开始，截取至单词“Shavla”，然后重新载入文本。7. 我们对文本应用三元语法。8. 打印结果。图1

该方法如图1所示。

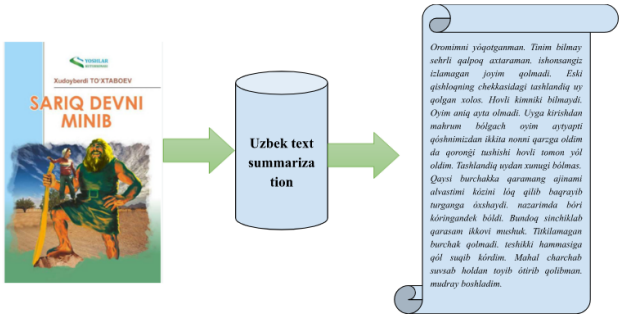

图1：应用乌兹别克语文本摘要算法所得的结果

5. 讨论与结论

实验中使用了 3-gram 对所取文本进行摘要。总体而言，将 n-gram 应用于乌兹别克语文本摘要问题以及选择最佳 n-gram 的问题仍未解决。尽管该问题尚无定论，但可以使用本文所提出的算法解决类似问题，这是本文的主要成果。为此，采用 i-gram 代替 3-gram。利用所有 i-gram 得到的结果，由专家选出最佳 n-gram。

本文提出了一种基于词频-逆文档频率的算法，用于解决乌兹别克语文本摘要问题。在应用该算法的过程中，我们使用先前创建的 School语料库移除了停用词。所得结果对后续研究的价值可概括如下：

1）解决乌兹别克语自然语言处理问题；2）选择基于词频-逆文档频率的乌兹别克语文本摘要问题中的最佳 n-gram；3）在解决乌兹别克语文本摘要问题时，与乌兹别克语文本摘要的主流方法相比，如何更有效地进行基于词频-逆文档频率的乌兹别克语文本摘要。此外，还有许多类似的应用场景。

致谢。作者谨向 Erasmus+ 计划以及 ELBA 项目（在中亚建立培训与研究中心并开发课程，以开展智能大数据分析；项目参考编号：610170-EPP-1-2019-1-ES-EPPKA2-CBHE-JP）表示衷心感谢。

本文作者还诚挚感谢乌尔根奇州立大学自然语言处理团队在模型创建和实验期间的讨论中给予的大力支持。

参考文献

Aizawa, A.（2003）。从信息论视角看 tf–idf 度量。《信息处理与管理》，39(1)，45–65。

Edmundson, H. P.，& Wyllys, R. E.（1961）。自动文摘与索引——综述与建议。《ACM 通讯》，4(5)，226–234。Knorz, G.，Krause, J.，& Womser-Hacker, C.（1993）。信息检索 '93：从建模到应用。《信息检索 '93》会议论文集，1，67–81。Kupiec, J.，Pedersen, J.，& Chen, F.（1995）。可训练的文档摘要系统。第18届 ACM SIGIR 信息检索研究与发展会议论文集，68–73。Kuriyozov, E.，Matlatipov, S.，Alonso, M. A.，& Gómez-Rodrguez, C.（2022）。低资源语言情感数据集的构建与评估：以乌兹别克语为例。《人类语言技术：计算机科学与语言学面临的挑战：第9届语言与技术会议，LTC 2019，波兹南，波兰，2019年5月17–19日，修订版精选论文》，13212，232–243。Lehnert, W. G.，& Ringle, M. H.（2014）。自然语言处理策略。心理学出版社。Luhn, H. P.（1958）。文学摘要的自动创建。《IBM 研究与开发期刊》，2(2)，159–165。Madatov, K.（2019）。乌兹别克语 WordNet 词条的 Prolog 格式。Madatov, K. A.，Khujamov, D. J.，& Boltayev, B. R.（2022）。基于土耳其语 WordNet 创建乌兹别克语 WordNet。《AIP 会议论文集》，2432(1)，60009。Madatov, K.，Bekchanov, S.，& Vičič, J.（2021）。乌兹别克语停用词列表。滨海大学，安德烈·马鲁希奇研究所。Madatov, K.，Bekchanov, S.，& Vičič, J.（2022a）。乌兹别克语停用词检测的准确性：基于“School语料库”的案例研究。arXiv 预印本，arXiv:2209.07053。Madatov, K., Bekchanov, S. 和 Vičič, J.（2022b）。针对乌兹别克语文本的停用词自动检测。Madatov, K., Bekchanov, S. 和 Vičič, J.（2022c）。从乌兹别克语文本中提取的停用词数据集。《数据简报》，43。https://doi.org/10.1016/j.dib.2022.108351 Madatov, X. A., Sharipov, M. S. 和 Bekchanov, S. K.（2021）。乌兹别克语文本中的非重要词。计算机语言学：问题、解决方案与展望 共和国第一届科学技术会议，1，156–162。Pollock, J. J. 和 Zamora, A.（1975）。化学文摘社的自动文摘研究。《化学信息与计算机科学期刊》，15(4)，226–232。Rush, J. E., Salvador, R. 和 Zamora, A.（1971）。自动文摘与索引。II. 通过应用语境推理和句法连贯性标准生成指示性摘要。《美国信息科学学会期刊》，22(4)，260–274。Salaev, U., Kuriyozov, E. 和 Gómez Rodr\'\iguez, C.（2022a）。乌兹别克语字母之间的机器转写工具。黏着语语言技术作为自然语言处理挑战国际会议（ALTNLP）。Salaev, U., Kuriyozov, E. 和 Gómez Rodr\'\iguez, C.（2022b）。SimRelUz：作为乌兹别克语语义评估数据集的相似度与相关性分数。LREC2022 资源匮乏语言特别兴趣小组（SIGUL 2022）研讨会。Sharipov, M. 和 Sobirov, O.（2022）。通过有限状态机为乌兹别克语开发基于规则的词形还原算法。arXiv 预印本，arXiv:2210.16006。Torres-Moreno, J.-M.（2014）。自动文本摘要。约翰·威立父子出版公司。

附录

附录 1：基于词频-逆文档频率的文本摘要合整算法（摘要化）

1: 输入（文本） 2: 分词（文本） 3: 删除（两个词的搭配（文本）） 4: 删除（一元语法（文本）） 5: 删除（基于规则（文本）） //{代词、副词、连词、插入语、状语词、助词、介词} 6: 词频-逆文档频率（文本） 7: p（文本_UW）= 词频-逆文档频率（文本_UW）/SUM（词频-逆文档频率（文本_UW）） // 文本_唯一词 - 文本 _UW 8: E=SUM （文本 _UW ） *p （文本 _UW ） 9: D=SUM （文本 _UW -E ） ^2*p（文本_UW） 10: SIGMA=SQR（D） 11: EI(m)=SUM（p（文本_UW））* 文本 _UW ^m// m – 1..len（文本_UW） 12: M1=E 13: M2= SUM（p（文本_UW））* 文本 _UW ^2 14: M3= SUM（p（文本_UW））* 文本_UW ^3 15: MIYU=M3-3*M1*M2+2 *M1^3 16: AS= MIYU/ SIGMA^3 17: 如果（AS>0）那么 文本2=词频-逆文档频率[0; E-SIGMA] 18: 如果（AS<0）那么 文本2=词频-逆文档频率[E-SIGMA- len（文本_UW）] 19: 如果（AS=0）那么 文本2=词频-逆文档频率[E-SIGMA； E+SIGMA] 20: 三元语法（文本2） 21: 分词（文本2） 22: 摘要文本：= P（分词（i）|分词（i+1）| 分词（i+2）） 23: 输出（摘要文本） 24: 结束