# GreekBART：首个希腊语序列到序列预训练模型

**Iakovos Evdaimon**[1] **Hadi Abdine**[1] **Christos Xypolopoulos**[1,2] **Stamatis Outsios**[3]

**Michalis Vazirgiannis**[1,4] **Giorgos Stamou**[2]

[1]巴黎综合理工学院，[2]雅典国立理工大学，[3]雅典经济与商业大学，[4]瑞典皇家理工学院

arXiv:2304.00869v1 [cs.CL] 3 Apr 2023

## 摘要

迁移学习时代彻底改变了计算机视觉和自然语言处理领域，带来了性能卓越、适用于多种任务的强大预训练模型。具体而言，自然语言处理任务一直由基于 Transformer 的语言模型主导。在自然语言推理和自然语言生成任务中，BERT 模型及其变体，以及 GPT 模型及其后续模型，都展现出了出色的性能。然而，这些模型中的大多数主要针对英语或多语言语料库进行预训练和评估。本文介绍 GreekBART，这是首个基于 BART-base 架构并在大规模希腊语语料库上预训练的序列到序列模型。我们在多种判别任务上评估并比较 GreekBART、BART-random、Greek-BERT 和 XLM-R。此外，我们还考察了该模型在 GreekSUM 的两项自然语言生成任务上的表现。GreekSUM 是一个新推出的希腊语摘要数据集。该模型、代码和新的摘要数据集将公开发布。

## 1 引言与相关工作

随着迁移学习的建立，机器学习领域进入了一个新时代，尤其为计算机视觉（Krizhevsky 等人，2017年）和自然语言处理领域带来了新的可能性。迁移学习已成为一种新趋势，因此从头开始训练计算机视觉或自然语言处理任务模型的情况已不常见，这也解决了真实世界机器学习应用中训练数据不足的问题。通过重用基于海量数据训练的预训练模型来解决任务，最终模型已达到最先进的性能。基于 Transformer（Vaswani等人，2017年）的预训练模型，例如 BERT（Devlin 等人，2019年）及其变体广泛应用于自然语言处理，并已被证明在许多任务中表现出色。

BART（Lewis 等人，2020年）是一种用于预训练序列到序列模型的去噪自编码器。它通过使用任意噪声函数破坏文本，并训练模型重建原始文本来完成训练。它采用标准的基于 Transformer 的神经机器翻译架构，以及一种标准的序列到序列架构，其中包含双向编码器（类似于 BERT）和从左到右的解码器（类似于 GPT（Radford等人，2018年））。这意味着编码器的注意力掩码是完全可见的，类似于 BERT；而解码器的注意力掩码具有因果性，类似于 GPT-2（Radford等人，2019年）。无监督预训练的 BART 能够学习语言模型，从而使我们可以将其适配到特定的自然语言处理任务。因此，微调不需要大规模的标注数据集。这类模型适用于机器翻译、问答，尤其是文本摘要任务，但这并不意味着 BART 不适用于序列分类任务；恰恰相反，它在这类任务中同样相当有效。

过去几年里，除了英语之外，研究者针对其他语言开展了大量研究。例如，法语模型 CamemBERT（Mar-tin et al.，2020年）和 BARThez（Kamal Eddine等人，2021年），阿拉伯语模型 CAMeLBERT（Inoueet al.，2021年）和 AraBART（Eddine et al.，2022年），日语模型 BART（Kim and Komachi，2021年），西班牙语和加泰罗尼亚语模型 BETO（Cañete et al.，2020年）和 NASes（Ahuir et al.，2021年），以及越南语模型 BARTpho（Tranet al.，2021年）。近来，人们提出了多种多语言语言模型，这些模型在多种语言的大规模语料库上进行预训练，以覆盖多种语言，并尝试同时学习多种语言的语言模型。值得注意的是，M-BERT（Devlin 等人，2019年）是一种多语言预训练

语言模型，由 BERT 的多语言版本构成，并在维基百科规模最大的 100 种语言上进行预训练。另一种广受欢迎的多语模型是 XLM（Conneau and Lample，2019年），这是一种基于 Transformer 的多语言语言模型，在 15 种语言的维基百科上进行预训练。该模型通过两项辅助任务进行训练：掩码语言建模和翻译语言建模。训练跨语言语言模型对低资源语言大有裨益，因为所有语言都使用相同的共享词汇表进行处理。Conneau 等人 2020年提出了 XLM-R，这是一种基于 RoBERTa 模型改进而来的 XLM。该模型采用跨语言掩码语言建模目标，在来自 Common Crawl 的 100 种语言、共 2.5 TB 的数据上进行训练（Wenzek et al.，2020年；Conneau 等人，2020年），使低资源语言可用的训练数据平均增加了两个数量级。最后，mBART（Liu 等人，2020年）是 BART 的多语言版本，在与 XLM-R 相同的数据集中选取的 25 种语言上进行预训练。在 mBART 中，我们使用其 250K SentencePiece（Kudo and Richardson，2018年）模型；该模型使用 XLM-R 中 100 种语言的单语数据进行训练，因此支持 mBART 最初训练所用的 25 种语言之外的其他语言。mBART25 的参数量约为 6.1 亿。后来，有人提出了 mBART 的扩展版本，新增 25 种语言（例如，总计 50 种语言），即 mBART50（Tang 等人，2020年），将参数量增加到约 6.8 亿。除了 mBART 和 mBART50 之外，前述所有其他多语言模型都支持希腊语。mBART25 和 mBART50 并未在现代希腊语上进行预训练，但其词汇表中包含现代希腊语。尽管如此，在大多数自然语言处理任务中，多语言模型的表现仍不及单语模型。近几个月来，另一个与 BART 相关、备受自然语言处理研究领域关注的模型是 ChatGPT [1]。ChatGPT 基于 GPT-3 架构（Brown et al.，2020年）构建，因此它是一种基于 Transformer 的语言模型，经过海量文本数据预训练，并针对对话式人工智能应用进行了微调。与 BART 一样，ChatGPT 能够生成高质量的文本序列，因此适用于文本摘要和问答等任务。然而，与 BART 不同的是，ChatGPT 专为对话应用而设计，因此非常适用于聊天机器人和其他对话系统。此外，ChatGPT 的架构是单向的，这意味着它能够按照从左到右的顺序生成文本，因此更适合语言生成和对话等任务。

[1] https://openai.com/blog/chatgpt

与广泛使用的语言相比，希腊语可用的语言资源较少。尤其是，针对希腊语的深度学习模型研究仍十分欠缺。不过，目前已有一些工作致力于开发用于希腊语自然语言处理的数据集、模型、知识库和框架。Outsios 等，2018 介绍了希腊语词嵌入的生成方法。他们使用了一个约 50GB 的大型语料库（包含 1.2 亿个句子），这些数据是从约 2000 万个 URL 抓取而来。随后，Lioudakis 等，2020提出了一种集成方法——连续跳字袋（Continuous Bag-of-Skip-grams），用于提取希腊语词表示。最近，Koutsikakis 等，2020采用了 Greek-BERT，这是首个基于 BERT 的希腊语 Transformer 语言模型。该模型在一个 29GB 的数据集上进行了预训练，并在多项希腊语自然语言处理任务中取得了最先进的性能。值得注意的是，Papantoniou 和 Tzitzikas，2020 对希腊语自然语言处理领域已开展的工作进行了全面综述。

在本文中，我们尝试解决多语模型不足以与单语模型竞争，以及可用于希腊语的深度学习模型有限这一问题。因此，我们提出了首个用于希腊语的预训练序列到序列单语模型。由于我们在大型单语希腊语语料库上对 BART-base 架构进行了预训练，该模型被称为 GreekBART。尽管已经存在 Greek-BERT（Koutsikakis 等人，2020年），我们的模型仍超越了 Greek-BERT 的能力，专注于生成任务。GreekBART 在两项不同的生成任务和四项判别任务上进行了评估。我们的主要贡献如下：

- 我们介绍一种基于 BART-base 架构、面向希腊语的预训练序列到序列模型（Lewis 等人，2020年），并在一个规模为 87.6 GB 的大型语料库上进行预训练。我们考察了该模型在四项判别任务（即两项分类任务、一项情感分析任务和一项自然语言推理任务）以及两项

生成任务中的表现。

• 我们推出首个希腊语摘要数据集 GreekSUM，并通过处理该数据集引入两项生成任务和一项分类任务。

• 我们将 GreekBART 与流行的语言模型进行比较，这些模型可能已经针对希腊语进行过预训练，也可能没有。在判别任务中，我们将本模型与 BART-random 模型、Greek-BERT（Koutsikakis 等人，2020年）和 XLM-R（Conneau 等人，2020年）进行比较。我们还考察了 GreekBART（即我们的模型）、BART-random 模型、mBART25（Liu 等人，2020年）和 mBART50（Tang 等人，2020年）在两项新型生成任务上的性能差异。

• 我们将发布代码和模型[2]，供所有希望进一步拓展我们工作应用范围或利用我们的贡献开展自身工作的人使用。

## 2 GreekBART

我们提出的模型基于 BART（Lewis 等人，2020年），这是一种去噪自编码器。我们采用*BASE* 架构，包含 6 层编码器和 6 层解码器。此外，编码器和解码器均使用 768 维隐藏层、12 个注意力头，并在编码器和解码器顶部各添加一个归一化层（Liu 等人，2020年）。添加这些层的目的是在采用 FP16 精度（Micikevicius 等人，2017年）时稳定训练。使用 FP16 精度可以加快模型的预训练。我们的模型总计约有 1.81 亿个参数。总体而言，我们遵循了与Kamal Eddine 等人 2021类似的方法：按照 BART（Lewis 等人，2020年）和 mBART（Liu 等人，2020年）的方法，对英语以外的其他语言的单语模型进行预训练。

### 2.1 预训练语料库

预训练语料库由以下语料库构成：（a）维基百科中的希腊语部分[3]；（b）欧洲议会会议平行语料库（EuroParl）中的希腊语部分[4] (Koehn，2005）；（c）OSCAR 中的希腊语部分[5]（Abadji 等人，2022年），这是 CommonCrawl 的一个清理版本[6]；（d）希腊语网络语料库，抓取自约 2000 万个希腊语 URL[7]（Outsios 等人，2018年）。具体而言，我们使用与 Greek-BERT（Koutsikakis 等人，2020年）模型相同的数据集，并纳入Outsios 等，2018的数据集，以获得更大的语料库，更适合用于 BART 模型的预训练。此外，选择这些数据集使我们涵盖了希腊语的广泛领域，包括正式和非正式文本、新闻文章、百科全书信息以及政治对话。如此多样的文本类型有助于确保 BART 模型的预训练具备稳健性，并能够处理希腊语使用中的不同风格和语域。总而言之，这些数据集的选择有助于确保希腊语 BART 模型能够胜任希腊语中的各种自然语言处理任务。

我们对上述每个语料库进行了预处理，删除了 URL、表情符号、标签和井号标签。此外，我们还删除了评论，以及一些观察到的、未提供额外上下文含义的噪声句子。不同数据集中的噪声句子各不相同，因此我们必须“手动”识别它们。此外，除维基百科数据集外，我们还删除了字符数少于一千的文档。对于维基百科，我们删除了字符数少于三十的文档。总体而言，我们没有删除非希腊语字符，因为我们认为这些字符数量微不足道，不会妨碍 GreekBART 理解语言模型。我们对每个语料库进行了去重，然后将它们合并为一个语料库。之后，我们又对合并后的数据集进行了最后一次去重。去重过程使用 runiq 软件包完成[8]。为了生成词汇表，我们使用了 SentencePiece[9] (Kudo 和 Richardson，2018年)，该工具实现了字节对编码（BPE）（Sennrich 等人，2016）。因此，无需进行任何预分词。我们将词汇表大小设为 50K 个子词，并使用预训练语料库中的 20GB 随机样本训练 SentencePiece 模型。我们将字符覆盖率设为 99。95%。语料库总大小为 76.9/87.6GB

[2] https://github.com/iakovosevdaimon/GreekBART
[3] https://dumps.wikimedia.org/elwiki/
[4] https://www.statmt.org/europarl/
[5] https://oscar-corpus.com/
[6] https://commoncrawl.org/
[7] http://nlp.polytechnique.fr/resources-greek
[8] https://github.com/whitfin/runiq
[9] https://github.com/google/sentencepiece

| 语料库 | 去重前大小<br>去重 | 去重后大小<br>去重 |
|---|---|---|
| OSCAR | 51.7 | 44.6 |
| 希腊语网络语料库 | 38.4 | 30.9 |
| 维基百科 | 0.9 | 0.9 |
| EuroParl | 0.5 | 0.5 |
| 总计 | 91.5 | 76.9 |

表 1：构成 GreekBART 预训练语料库的数据集（单位：GB，清理和去重前后）。

SentencePiece 分词前/后。

### 2.2 训练详情

我们遵循与 BART 相同的预训练流程。因此，GreekBART 通过最小化解码器输出与原始输入之间的交叉熵损失，尝试重建被破坏的输入。输入文本采用了两种噪声处理方式。首先，我们使用文本填充技术，将若干文本片段替换为一个称为[MASK], 屏蔽30% 文本的特殊词元。使用泊松分布($lambda = 3.5$) 来确定片段的长度。第二种扰动方法是句子置换，即随机打乱输入文档中的句子。我们在 Jean Zay 上对 GreekBART 进行预训练，每个图形处理器的批次大小设为 768000 个词元，并将更新频率设为 128。我们使用 Adam 优化器（Kingma and Ba，2015年），采用 $\epsilon = 10^{-6}$, $\beta_1 = 0.9$和 $\beta_2 = 0.999$，学习率从$6.10^{-4}$ 开始，并随训练步数线性递减。总训练步数的 6% 用于预热。前 12 个训练轮次中，丢弃率固定为 0.1；第 12 至第 16 个训练轮次中，丢弃率降至 0.05；最后，第 16 至第 20 个训练轮次中，丢弃率设为 0。所有实验均使用 Fairseq 库完成[10] (Ott等人，2019年)。

## 3 GreekSUM

基于Transformer的序列到序列模型（包括 BART）不仅能执行抽取式摘要生成，也能执行抽象式摘要生成。这类摘要生成是自然语言处理中最核心且最具挑战性的评测任务之一。然而，目前尚无适用于希腊语的摘要数据集。因此，我们创建了首个适用于抽象式摘要任务的希腊语数据集。

[10]https://github.com/facebookresearch/fairseq

### 3.1 动机

我们的主要目标是创建一个与OrangeSum数据集等效的希腊语版本[11] （Kamal Eddine等人，2021年）以及XSum数据集（Narayan 等，2018年）。OrangeSum通过抓取“Orange Actu”网站上的文章、单句标题及简短摘要而构建[12]。每篇文章的标题和摘要均由文章作者撰写。在OrangeSum和XSum数据集上表现良好的模型都需要具备很高的抽象性程度。

### 3.2 数据收集

我们采用了类似的方法，抓取了“News24/7”网站[13]。News24/7 是希腊领先的新闻网站之一，隶属于 24 ME- DIA 数字出版集团[14]。我们收集了从 2007年10月 至 2022年6月 的网页数据，涵盖五个主要类别：政治、社会、经济、文化和世界。每篇文章都包含一个单句标题和一段简洁的摘要。我们提取了这些内容，由此得到两项摘要任务：GreekSUM 标题和 GreekSUM 摘要。这两项新任务的 Gold 摘要平均长度分别为 9.95 词和 24.55 词（见表 2）。

### 3.3 后处理

首先，我们对抓取页面进行了筛选，删除所有空文章，以及标题少于 2 个词或摘要少于 5 个词的文章。其次，我们筛除了重复文章（即正文相同、标题相同或摘要相同的文章），因为一篇文章可能属于多个类别，因此可能被多次抓取。最后，我们注意到，有些摘要看起来更像是引言，而非对文章的实际总结。因此，我们剔除了摘要中新颖一元词比例最高的 10 % 文章。这对应于 46.7% 的新颖一元词阈值。对于提出的两项摘要任务，我们分别留出 1 万对用于测试集、1 万对用于验证集，其余所有配对均用于训练集。使用我们的代码[15]即可复现已发布的 GreekSUM 数据集。

[11]https://github.com/Tixierae/OrangeSum[12]
https://actu.orange.fr/[13]https://www.news247.gr/[14]
https://www.24media.gr/[15]
https://github.com/iakovosevdaimon/GreekSUM

| 数据集 | 训练集/验证集/测试集 | 平均文档长度 | | 平均摘要长度 | | 词汇量 | |
|---|---|---|---|---|---|---|---|
| | | 词 | 句子数 | 词 | 句子数 | docs | 摘要 |
| CNN | 90.3/1.22/1.09 | 760.50 | 33.98 | 45.70 | 3.58 | 34 | 89 |
| DailyMail | 197/12.15/10.40 | 653.33 | 29.33 | 54.65 | 3.86 | 564 | 180 |
| 《纽约时报》 | 590/32.73/32.73 | 800.04 | 35.55 | 45.54 | 2.44 | 1233 | 293 |
| XSum | 204/11.33/11.33 | 431.07 | 19.77 | 23.26 | 1.00 | 399 | 81 |
| OrangeSum 标题 | 30.6/1.5/1.5 | 315.31 | 10.87 | 11.42 | 1.00 | 483 | 43 |
| OrangeSum 摘要 | 21.4/1.5/1.5 | 350 | 12.06 | 32.12 | 1.43 | 420 | 71 |
| GreekSUM 标题 | 146.046/10/10 | 355.49 | 14.26 | 9.95 | 1.05 | 663 | 91 |
| GreekSUM 摘要 | 129.159/10/10 | 368.97 | 14.76 | 24.55 | 1.46 | 629 | 127 |

表 2：规模（第 2 列）以数千份文档为单位。文档和摘要的长度以词计，词汇量以数千词元计

| 数据集 | 标准摘要中新颖n元组的百分比 | | | | LEAD | | | EXT-ORACLE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 一元组 | 二元组 | 三元组 | 四元组 | R-1 | R-2 | R-L | R-1 | R-2 | R-L |
| CNN | 16.75 | 54.33 | 72.42 | 80.37 | 29.15 | 11.13 | 25.95 | 50.38 | 28.55 | 46.58 |
| DailyMail | 17.03 | 53.78 | 72.14 | 80.28 | 40.68 | 18.36 | 37.25 | 55.12 | 30.55 | 51.24 |
| 《纽约时报》 | 22.64 | 55.59 | 71.93 | 80.16 | 31.85 | 15.86 | 23.75 | 52.08 | 31.59 | 46.72 |
| XSum | 35.76 | 83.45 | 95.50 | 98.49 | 16.30 | 1.61 | 11.95 | 29.79 | 8.81 | 22.65 |
| OrangeSum 标题 | 26.54 | 66.70 | 84.18 | 91.12 | 19.84 | 08.11 | 16.13 | 31.62 | 17.06 | 28.26 |
| OrangeSum 摘要 | 30.03 | 67.15 | 81.94 | 88.3 | 22.21 | 07.00 | 15.48 | 38.36 | 20.87 | 31.08 |
| GreekSUM 标题 | 26.7 | 67.9 | 84.5 | 91.4 | 14.68 | 04.46 | 14.37 | 23.36 | 07.39 | 23.12 |
| GreekSUM 摘要 | 20.6 | 50.8 | 65.3 | 73.0 | 17.11 | 06.17 | 16.69 | 34.18 | 14.17 | 33.93 |

表 3：GreekSUM 与其他数据集的抽象性程度。图中显示，GreekSUM 与 OrangeSum 一样，遵循 XSum 的特点，比传统摘要数据集更具抽象性。

### 3.4 分析

表2 对比了 GreekSUM 与 Or- angeSum、XSum，以及广为人知的 CNN、DailyMail 和《纽约时报》数据集（Hermann 等人，2015）。我们可以看到，就文档和摘要的平均长度而言，GreekSUM 与 Or- angeSum 数据集非常相近。此外，GreekSUM 的规模与 XSum 相似。查看表 3可以发现，抽取式方法（即 LEAD 和 EXT-ORACLE）在 GreekSUM 上的表现并不理想，因此我们的数据集对抽取式模型的偏向较弱。鉴于这两种抽取式方法表现不佳，GreekSUM 似乎比传统摘要数据集（即 CNN、DailyMail、《纽约时报》）更偏向抽象式摘要生成。不过，GreekSUM 的摘要和标题并未像 OrangeSum 和 XSum 那样展现出很高的新颖度。在 GreekSUM 数据集中，摘要中有 20.6% 的新颖一元词，标题中有 26.7% 的新颖一元词；相比之下，OrangeSum 摘要中为 30%，OrangeSum 标题中为 26.5%，而 XSum 中为 35.7%。因此，我们可以得出结论：GreekSUM 的摘要并没有我们期望的那么偏向抽象式生成。

## 4 实验

本节展示所有实验的结果。基本上，我们研究了两类下游任务：判别任务和摘要生成任务。对于判别任务，我们将 GreekBART 与 BART-random、GreekBERT（Koutsikakis 等人，2020年）以及 XLM-R 模型（Conneau 等人，2020年）进行比较。除 BART-random 外，其他模型均已在希腊语上进行预训练。因此，我们将本模型的表现与当前仅针对希腊语预训练的最先进单语模型，以及广泛使用的多语模型进行比较。我们对上述所有模型进行了微调，以用于下游任务。

对于摘要任务，我们将 GreekBART、BART-random 和两个版本的 mBART 并列比较（Liu 等人，2020年；Tang 等人，2020年）。mBART25 和 mBART50 基于 BART 的 *LARGE* 架构构建，分别在 25 种和 50 种语言上进行预训练，但不包括希腊语。因此，我们在摘要任务中进行了零样本学习。另一方面，BART-random 模型采用与 Greek-BART 相同的架构和词汇表，但它是在下游任务上从头训练的。

### 4.1 判别任务

除了生成任务之外，BART 模型在判别任务中也取得了显著成果（Lewis 等人，2020年）。对于序列分类，我们在模型顶部添加一个分类头，并将输入同时送入编码器和解码器。新增的多类线性分类器使用解码器最后一个词元的表示。我们考察了各模型（即Greek-BERT、XLM-R、BART-random、GreekBART）在四项判别任务上的表现。具体而言，我们在两项分类任务、一项情感分析任务和一项自然语言推理（NLI）任务上评估模型。

#### 4.1.1 训练详情

在所有实验中，我们根据最佳验证分数，从 $\{10^{-4}$、$5.10^{-5}$、$10^{-5}\}$ 中选择学习率对模型进行微调。每项实验使用不同的随机种子重复 3 次，并记录模型在上述各项任务的测试集上的准确性均值和标准差。

#### 4.1.2 新闻类别分类任务（新闻类别分类任务）

对于第一个分类任务，我们使用了从新闻网站 News24/7 抓取的新型摘要数据集（GreekSum，参见第 3 节）。我们考虑了文章可能归属的五个不同主题：政治、社会、经济、文化和世界。这些类别作为标签，用于训练模型执行分类任务。本质上，模型接收文章内容，并学习预测其所属类别（即主题）。我们对所有受测模型进行了 5 个训练轮次的微调，批次大小设为 32。对于 XLM-R 模型，我们将学习率设为 $5.10^{-5}$，其余模型的学习率则设为 $10^{-4}$。训练集包含 146,046 个样本，而验证集和测试集均恰好包含 10,000 个实例，与两个摘要数据集（即GreekSUM 摘要和 GreekSUM 标题）相同。

在第二项分类任务中，我们使用了 Lioudakis 等，2020提出的希腊语分类数据集，该数据集由 Makedonianewspaper 的文章构建而成。该数据集包含 8005 篇文章，涵盖 18 个不同类别：体育、报道、经济、政治、国际、电视、艺术与文化、读者来信、观点、访谈、天气、社会、广告、传记、其他、文章、警方和星座。我们将 70% 的数据集用于训练，其余 30% 用于验证和测试。因此，训练集包含 5610 个样本，测试集和验证集分别包含 1191 个和 1204 个样本。所有模型均以 $5.10^{-5}$ 的学习率和 16 的批次大小进行微调，共训练 20 轮。由于数据集规模较小，我们使用了更多训练轮次和更小的批次大小来训练模型。

#### 4.1.3 自然语言推理

跨语言自然语言推理语料库（XNLI）（Conneau 等人，2018年）包含成对的句子。此任务旨在判断第一个句子（也称为前提）是否蕴含、与第二个句子矛盾，或与之中立；第二个句子称为候选文本。XNLI 语料库包含 5,000 对测试集数据、2,500 对验证集数据，以及来自 MultiNLI 语料库的 34 万对训练集数据（Williams 等人，2018年）。该数据集已从英语翻译成包括希腊语在内的 14 种语言。遗憾的是，大量训练集样本质量极差，因为它们是通过机器翻译生成的。这种情况可能会影响模型的性能。我们进行了 5 个训练轮次的微调，批次大小为 32，学习率为 $5.10^{-5}$。

#### 4.1.4 情感分析任务

我们使用了一个公开的希腊语电影评论情感分析数据集[17]。预处理数据集时，我们主要移除了表情符号和话题标签。每个样本由一条评论和一个评分组成。为了区分正面评论和负面评论，我们设定了 5 分制中的 3 分为阈值。评分高于此阈值的评论归类为正面评论，而评分不高于 3 分的评论则归类为负面评论。为了创建一个平衡的数据集，我们力求纳入数量相近的正面评论和负面评论。针对我们的任务，我们仅保留了评论和评分，舍弃了所有其他信息。我们将数据集划分为训练集、验证集和测试集。训练集包含 104,157 个样本，而验证集和测试集

[16] https://www.news247.gr/

[17] https://www.kaggle.com/datasets/nikosfragkis/ 希腊电影数据集

分别包含 22,320 个和 22,318 个样本。我们将学习率和批次大小分别设为 $5.10^{-5}$ 和 16。我们对模型进行了 5 个训练轮次的微调。

#### 4.1.5 结果

表4 报告了四项不同任务的测试集准确率。我们将模型与 GreekBERT（Koutsikakis 等人，2020年）、XLM-R（Con- neau 等人，2020年）以及 BART-random 进行比较。所有模型均采用相应的*BASE* 架构。在这些模型中，我们发现 GreekBART 在几乎所有判别任务中表现最佳，情感分析任务除外；在该任务中，GreekBERT 的表现最佳。通常，BERT 模型在这类任务中的表现优于 BART 模型。我们的模型（即 GreekBART）的表现印证了 BART 论文（Lewis 等人，2020年）中的结果：基于该架构的模型在生成任务和判别任务中均表现出色。

### 4.2 摘要生成

我们在两种不同的摘要生成任务中评估模型：模型根据文章相应的内容学习预测标题和摘要。在这两种生成任务中，GreekBART 均经过 30 个训练轮次的微调，学习率为 $5.10^{-5}$，该学习率在 6% 的训练步数内逐渐升温，随后线性衰减至 0。我们使用与 GreekBART 相同的超参数训练 mBART25 和 mBART50。至于 BART-random，我们训练了 60 个训练轮次。为了生成测试集的摘要，我们使用 ROUGE-L（Lin，2004）选择与最佳验证分数对应的检查点。此外，我们采用了两种抽取式技术作为基线：EXT-ORACLE 和 LEAD（Narayan 等人，2018年）。LEAD 技术通过抽取文档的前 $N$ 个句子来生成摘要，其中在我们的实验中 $N$ 设为 1。另一方面，EXT-ORACLE 从文档中选择一组句子，使特定得分最大化；在我们的实现中，所用得分为 ROUGE-L。具体而言，我们抽取了文档中 ROUGE-L 得分最高的一个句子。在表 5中，我们报告了 ROUGE-1、ROUGE-2、ROUGE-L 得分（Lin，2004）以及两种不同的 BERTScore （ Zhang 等人，2019 年），使用 M-BERT（Devlin 等人，2019年）模型和 Greek-BERT 模型来计算上下文嵌入。BERTScore 是一种近期提出的指标，它利用预测句子和参考句子的上下文表示。BERTScore 关注参考文本与候选文本中词元之间的语义相似性，旨在理解生成内容的含义，以及原本预期生成的内容。我们报告 BERTScore，是因为 ROUGE 主要捕捉 n 元语法重叠，这对于抽象式摘要任务而言并不充分。附录 A、B 部分提供了一些生成摘要的示例。

#### 4.2.1 定量结果

在表 5 中，我们比较了在摘要生成任务上微调的模型性能。尽管 GreekBART 是一个 BART-*BASE* 模型，却与 BART-*LARGE* 模型进行比较，但它在 GreekSUM 摘要任务上的表现优于其他所有模型。使用 M-BERT 模型评估时，只有 mBART50 的 BERTScore 略高于 GreekBART。另一方面，在 GreekSUM 标题任务上，两个 mBART 模型的表现都超过了我们的模型。不过，即使在该任务中，GreekBART 在 ROUGE 和 BERTScore 方面的表现也与两个 mBART 模型之一相当。我们的评估表明，mBART50 和 GreekBART 是这两项摘要生成任务中最有前景的模型。具体而言，mBART50 在两项生成任务中的整体表现更好：根据 ROUGE 和 BERTScore，它在 GreekSUM 标题任务中排名第一，在 GreekSUM 摘要任务中排名第二。另一方面，GreekBART 在 GreekSUM 摘要任务中表现出色，但在 GreekSUM 标题任务中排名第三。总体而言，值得注意的是，两个未针对希腊语进行预训练的 mBART 模型，凭借规模超过 10 万个训练样本的 GreekSUM 数据集，仍取得了良好的表现。显然，BART-random 的表现明显最差。最后，有趣的是，mBART50 在 ROUGE 和 BERTScore 方面的表现都优于 mBART25，而两者唯一的区别在于预训练所用语言的数量。这一现象值得进一步研究，因为 mBART50 支持的另外 25 种语言中，可能有些语言与希腊语同源，从而有助于提升该语言模型对希腊语的理解。

| 模型 | NCC | | 情感分析 | XNLI |
|---|---|---|---|---|
| | News24/7（我们的） | Makedonia（Lioudakis 等人，2020年） | | |
| Greek-BERT | $92.61^{\pm 0.19}$ | $89.45^{\pm 0.84}$ | $\mathbf{86.39}^{\pm 0.06}$ | $78.6^{\pm 0.62}$ |
| XLM-R | $93.1^{\pm 0.51}$ | $89.6^{\pm 0.29}$ | $85.43^{\pm 0.05}$ | $78.2^{\pm 0.59}$ |
| BART-random | $91.33^{\pm 0.17}$ | $80.17^{\pm 0.09}$ | $80.87^{\pm 0.12}$ | $60.1^{\pm 0.43}$ |
| GreekBART（我们的方法） | $\mathbf{93.2}^{\pm 0.29}$ | $\mathbf{91.1}^{\pm 0.43}$ | $85.43^{\pm 0.19}$ | $\mathbf{78.67}^{\pm 0.25}$ |

Tabl表 4：判别任务的结果。我们给出平均准确率及标准差 .

| | | GreekSUM 摘要 | | | | GreekSUM 标题 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | BERTScore | R-1 | R-2 | R-L | BERTScore |
| | LEAD | 17.11 | 06.17 | 16.69 | 72.61/63.56 | 14.68 | 04.46 | 14.37 | 70/57.13 |
| | EXT-ORACLE | 34.18 | 14.17 | 33.93 | 73.89/65.43 | 23.36 | 07.39 | 23.12 | 70.02/57.33 |
| BASE | BART-random | 13.85 | 04.47 | 13.65 | 72.44/63.27 | 11.55 | 03.27 | 11.42 | 74.47/62.22 |
| BASE | GreekBART（我们的） | **16.5** | **06.13** | **16.21** | 73.03/**64.46** | 15.35 | 05.02 | 15.18 | 75.78/63.98 |
| LARGE | mBART25 | 15.07 | 05.8 | 14.82 | 72.75/64.08 | 16.09 | 05.58 | 15.93 | **76.81/65.38** |
| LARGE | mBART50 | 15.53 | 06. | 15.31 | **73.07**/64.43 | **16.1** | **05.59** | **15.96** | **76.81/65.38** |

表5：希腊SUM上的结果。除 ROUGE 外，我们还提供 BERTScore。左侧的 BERTScore 使用 M-BERT 模型计算（Devlin 等人，2019年），右侧的则使用 Greek-BERT（Koutsikakis 等人，2020年）

#### 4.2.2 定性结果

如表6所示，与两个 mBART 模型相比，GreekBART 的摘要生成方式更偏向抽象式，其生成的摘要包含更多新颖 n元语法。总体而言，没有任何模型在 ROUGE 得分方面超过 LEAD 方法。此外，各模型的 ROUGE 得分表明，机器生成的摘要倾向于采用抽取式，因为 Gold 摘要本身也主要是抽取式的。这一点从新颖 n元语法的占比中也得到了证实（表 6）：GreekSUM 的 Gold 摘要中引入的新词很少，因此影响了所考察模型的训练，促使它们生成更偏向抽取式的摘要。此外，表6 显示，所有生成摘要的长度都与真实摘要的长度相当接近。根据表7，mBART50 生成的摘要中重复内容所占比例最低，GreekBART 次之。mBART50 摘要中的重复词比例接近真实摘要。最后，我们注意到，BART-random 虽然引入了许多新词，但这些词并不相关。

#### 4.2.3 人工评估

为了进一步理解并验证定量结果，我们采用最佳-最差量表法开展了一项人工评估研究（Louviere 等人，2015）。我们选取了 11 位年龄各异、介于 18 至 60 岁之间的希腊语母语者，教育背景和水平各不相同。遵循 Narayan 等人 2018年 的方法，我们从 GreekSUM 摘要的测试集中随机选取了 14 篇文档，并针对每篇文档生成了由人工撰写（Gold）、GreekBART、BART-random、mBART25 和 mBART50 摘要构成的所有可能配对，共计 140 对。因此，每对摘要都由两个不同模型生成。志愿者会看到一篇文档和一对摘要，并根据准确性（摘要是否包含准确事实？）、信息量（是否涵盖了重要信息？）和流畅度（摘要是否以通顺的希腊语撰写？）判断哪篇摘要最好、哪篇最差。每对摘要随机分配给三位参与者；系统的得分通过计算其被选为 最佳 摘要的比例，减去其被选为 最差摘要的比例得出。因此，模型的最高得分为 100，最低得分则为 $-100$。人工评估研究的结果见表 8。Gold 排名第一，其次是 mBART50 和 GreekBART。评估者认为，Gold 是最受青睐的摘要，远远领先于其他摘要；而 mBART50 的得分明显高于 GreekBART，这验证了我们根据定量结果作出的假设。最后，BART-random 的负分较高，表明其摘要在大多数情况下被认为

| | | GreekSUM 摘要 | | | | | GreekSUM 标题 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 一元组 | 二元组 | 三元组 | 四元组 | 长度 | 一元组 | 二元组 | 三元组 | 四元组 | 长度 |
| | Gold | 20.6 | 50.8 | 65.3 | 73.0 | 24.55 | 26.7 | 67.9 | 84.5 | 91.4 | 9.95 |
| BASE | BART-random | 9.6 | 43.0 | 64.5 | 76.8 | 20.27 | 21.6 | 69.4 | 89.1 | 95.8 | 9.37 |
| BASE | GreekBART（我们的模型） | **7.4** | **23.5** | **34.5** | **42.2** | 23.63 | **14.9** | **50.1** | **69.3** | **79.9** | **9.78** |
| LARGE | mBART25 | 6.2 | 20.0 | 29.4 | 36.0 | 26.22 | 12.8 | 46.6 | 65.6 | 76.2 | 10.67 |
| LARGE | mBART50 | 6.5 | 21.8 | 32.3 | 39.7 | 23.95 | 12.8 | 46.6 | 65.6 | 76.2 | 10.67 |

表6：生成的摘要中新颖n元组的比例。此外，还给出了生成的摘要长度（词数）

| | | **重复（%）** |
|---|---|---|
| Abstract | Gold | 7.77 |
| Abstract | BART-random | 28.12 |
| Abstract | GreekBART（我们的方法） | 12.19 |
| Abstract | mBART25 | 12.7 |
| Abstract | mBART50 | 10.03 |
| | | **重复（%）** |
| Title | Gold | 0.91 |
| Title | BART-random | 8.76 |
| Title | GreekBART（我们的方法） | 3.62 |
| Title | mBART25 | 2.52 |
| Title | mBART50 | 2.52 |

表7：摘要中重复词的比例

| | **系统** | **得分** |
|---|---|---|
| | Gold | 45.24 |
| BASE | BART-random | −72.62 |
| BASE | GreekBART（我们的方法） | 10.71 |
| LARGE | mBART25 | −03.57 |
| LARGE | mBART50 | **20.24** |

表8：人工评估研究的结果

在大多数情况下被认为更差。

## 5 结论

我们实现了专门针对希腊语的首个预训练序列到序列模型 GreekBART。此外，我们还创建了首个希腊语摘要数据集。我们的模型在 4 项判别任务中的 3 项上优于此前最先进的模型，在摘要任务上的表现则与 BART-*LARGE* 模型相当。此外，我们展示了零样本学习的能力：尽管一个多语言 BART 模型并未在希腊语上进行预训练，我们仍从头开始对其进行摘要任务训练。未来，我们可以考虑创建一个抽象程度更高的摘要数据集，并研究希腊语与 mBART50 的 25 种额外语言中的一种或多种之间是否存在相关性。最后，尝试以经济可行的语言自适应阶段提升 mBART50 在摘要任务上的表现，也会很有意思通过应用经济可行的语言自适应阶段，使其在合理数量的训练轮次中进一步接受希腊语预训练。

## 伦理声明

GreekSUM 数据集的收集通过 Python 爬虫完成，该爬虫遵守了*robots.txt* 文件 http://www.news247.gr的规定。由于该数据集仅用于评估，因此其内容遵循网页上列出的法律要求。

训练 GreekBART 时，我们使用了一个由 GPU 组成的集群，其中包括 2 个 NVIDIA V100 GPU，训练持续了 20 天。与大多数基于 BART 架构的语言模型一样，目前模型预训练所需的能源资源非常庞大，这一问题亟待解决（Strubell 等人，2019年）。

## 局限性

我们用于评估模型的 GreekSUM 数据集仅包含来自一个网页的新闻文章。因此，GreekBART 的抽象式摘要能力仅在一个领域内进行了评估。这是因为目前缺乏非英语基准测试和任务。判别任务也存在同样的情况：希腊语目前仅有句子分类或自然语言推理任务可用。希腊语的其他评估数据集尚不存在（例如词义消歧数据集），或未向公众开放（例如命名实体识别数据集）。

另一方面，GreekBART 仅与抽取式摘要方法或用于摘要任务的大型多语言语言模型进行了比较。由于它是该语言的首个基础模型，且基础版 mBART 模型并未公开，因此无法对 GreekBART 与其他摘要系统进行公平、深入的比较。

## 致谢

本研究得到了 ANR chair AML/HELAS（ANR-CHIA-0020-01）的资助。

本研究获准使用 IDRIS 的高性能计算资源，资源分配编号为 2022年- AD011013750，由 GENCI 提供。

我们衷心感谢所有参与这项人工评估研究的参与者。您花时间和精力完成问卷并参与研究，对我们收集有意义的数据至关重要。

您愿意分享自己的经历、见解和观点，这对启发我们的研究发挥了重要作用，我们也感谢您对我们的信任。您的贡献帮助我们加深了对研究主题的理解，并有望对未来的研究与实践产生重大影响。

我们还要强调，在所有参与者参加研究之前，获取其知情同意至关重要。您的参与完全出于自愿，我们感谢您愿意参与这项研究。

再次衷心感谢所有参与者为本研究做出的宝贵贡献。

## 参考文本

Julien Abadji、Pedro Ortiz Suarez、Laurent Romary 和 Benoît Sagot。2022年。迈向更清洁、面向文档的多语言网络爬取语料库。载于第十三届语言资源与评估会议论文集，第4344–4355页，法国马赛。欧洲语言资源协会。

Vicent Ahuir、Lluís-F. Hurtado、José Ángel González 和 Encarna Segarra。2021年。Nasca 和 nases：用于加泰罗尼亚语和西班牙语抽象式摘要的两种单语预训练模型。应用科学，11(21)。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901。

José Cañete、Gabriel Chaperon、Rodrigo Fuentes、Jou-Hui Ho、Hojin Kang 和 Jorge Pérez。2020年。西班牙语预训练 BERT 模型及评估数据。载于*ICLR 2020* 的 *PML4DC*。

Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guillaume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettle- moyer 和 Veselin Stoyanov。2020年。大规模无监督跨语言表征学习。载于计算语言学协会第*58*届年会论文集，第8440–8451页，线上举行。计算语言学协会。

Alexis Conneau 和 Guillaume Lample。2019年。跨语言语言模型预训练。神经信息处理系统进展，32。

Alexis Conneau、Ruty Rinott、Guillaume Lample、Ad- ina Williams、Samuel Bowman、Holger Schwenk 和 Veselin Stoyanov。2018年。XNLI：评估跨语言句子表示。载于*2018*年自然语言处理实证方法会议论文集，第2475–2485页，比利时布鲁塞尔。计算语言学协会。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会会议：人类语言技术，第*1*卷（长篇与短篇论文）论文集，第4171–4186页，美国明尼苏达州明尼阿波利斯。计算语言学协会。

Moussa Kamal Eddine、Nadi Tomeh、Nizar Habash、Joseph Le Roux 和 Michalis Vazirgiannis。2022年。AraBART：用于抽象式摘要的预训练阿拉伯语序列到序列模型。

Karl MoritzHermann、Tomas Kocisky、Edward Grefen- stette、Lasse Espeholt、Will Kay、Mustafa Suleyman 和 Phil Blunsom。2015年。教机器阅读与理解。载于神经信息处理系统进展，第28卷。Curran Associates, Inc.。

Go Inoue、Bashar Alhafni、Nurpeiis Baimukan、Houda Bouamor 和 Nizar Habash。2021年。阿拉伯语预训练语言模型中变体、规模与任务类型的相互作用。载于第六届阿拉伯语自然语言处理研讨会论文集，第92–104页，乌克兰基辅（线上）。计算语言学协会。

Moussa Kamal Eddine、Antoine Tixier 和 Michalis Vazirgiannis。2021年。BARThez：一种高性能的预训练法语序列到序列模型。载于*2021*年自然语言处理实证方法会议论文集，第9369–9390页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Hwichan Kim 和 Mamoru Komachi。2021年。面向专利

WAT 2021 的任务。载于第八届亚洲翻译研讨会（*WAT 2021*）论文集，第 133–137 页，线上。计算语言学协会。

Diederik P. Kingma 和 Jimmy Ba。2015年。Adam：一种随机优化方法。载于 第*3*届学习表示国际会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集。

Philipp Koehn。2005年。 EuroParl：用于统计机器翻译的平行语料库。载于 机器翻译峰会 *X*论文集，第79–86页，泰国普吉岛。

John Koutsikakis、Ilias Chalkidis、Prodromos Malaka- siotis 和 Ion Androutsopoulos。2020年。Greek-BERT：希腊人拜访芝麻街。载于第 *11* 届希腊人工智能会议，SETN 2020，第110–117页，美国纽约州纽约市。计算机协会。

Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E. Hinton。2017年。 基于深度卷积神经网络的ImageNet分类。 *Commun.ACM*，60(6):84–90。

Taku Kudo 和 John Richardson。2018年。SentencePiece：一种用于神经文本处理的简单、语言无关的子词分词器和反分词器。载于*2018*年自然语言处理实证方法会议：系统演示论文集，第 66–71 页，比利时布鲁塞尔。计算语言学协会。

Mike Lewis、Yinhan Liu、Naman Goyal、Mar-jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于计算语言学协会第 *58* 届年会论文集，第 7871–7880 页，线上举行。计算语言学协会。

Chin-Yew Lin。2004年。ROUGE：用于自动评估摘要的软件包。载于文本摘要的新方向，第 74–81 页，西班牙巴塞罗那。计算语言学协会。

Michalis Lioudakis、Stamatis Outsios 和 Michalis Vazirgiannis。2020年。 一种侧重希腊语词表示生成的集成方法。载于 第三届低资源语言机器翻译技术研讨会论文集，第 99–107 页，中国苏州。计算语言学协会。

Yinhan Liu、Jiatao Gu、Naman Goyal、Xian Li、Sergey Edunov、Marjan Ghazvininejad、Mike Lewis 和 Luke Zettlemoyer。2020年。 面向神经机器翻译的多语言去噪预训练。 计算语言学协会学报，8：726–742。

Jordan J. Louviere、Terry N. Flynn 和 A. A. J. Marley。2015年。 最佳-最差量表法：理论、方法与应用。剑桥大学出版社。Louis Martin、Benjamin Muller、Pedro Javier Ortiz Suárez、Yoann Dupont、Laurent Romary、Éric de la Clergerie、Djamé Seddah 和 Benoît Sagot。2020年。CamemBERT：一种美味的法语语言模型。载于计算语言学协会第*58*届年会论文集。计算语言学协会。

Paulius Micikevicius、Sharan Narang、Jonah Alben、Gregory F. Diamos、Erich Elsen、David García、BorisGinsburg、MichaelHouston、OleksiiKuchaiev、Ganesh Venkatesh 和 Hao Wu。2017年。 混合精度训练。*CoRR*，abs/1710.03740。

Shashi Narayan、Shay B. Cohen 和 Mirella Lapata。2018年。 别告诉我细节，只给我摘要！用于极端摘要生成的主题感知卷积神经网络。载于*2018*年自然语言处理实证方法会议论文集，第 1797–1807 页，比利时布鲁塞尔。计算语言学协会。

Myle Ott、Sergey Edunov、Alexei Baevski、Angela Fan、Sam Gross、Nathan Ng、David Grangier 和 Michael Auli。2019年。Fairseq：用于序列建模的快速、可扩展工具包。载于《*2019*年北美计算语言学协会分会会议论文集（演示）》，第48–53页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Stamatis Outsios、Konstantinos Skianis、Polykarpos Meladianos、ChristosXypolopoulos 和 Michalis Vazirgiannis。2018年。来自大规模希腊语网络内容的词嵌入。*arXiv* 预印本 *arXiv:1810.06694*。

Katerina Papantoniou 和 Yannis Tzitzikas。2020年。希腊语自然语言处理：简要综述。载于第十一届希腊人工智能会议，SETN 2020，第101–109页，纽约，纽约州，美国。计算机协会。

AlecRadford、Karthik Narasimhan、TimSalimans、Ilya Sutskever 等。2018年。通过生成式预训练提升语言理解能力。

AlecRadford、JeffreyWu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019年。语言模型是无监督的多任务学习者。*OpenAI* 博客，1(8):9。

Rico Sennrich、Barry Haddow 和 Alexandra Birch。2016年。 使用子词单元进行稀有词的神经机器翻译。载于第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文），第1715–1725页，德国柏林。计算语言学协会。

Emma Strubell、Ananya Ganesh 和 Andrew McCallum。2019年。自然语言处理中的深度学习能源与政策考量。载于 第*57*届计算语言学协会年会论文集，第3645–3650页，意大利佛罗伦萨。计算语言学协会。

Yuqing Tang、Chau Tran、Xian Li、Peng-Jen Chen、Na- man Goyal、Vishrav Chaudhary、Jiatao Gu 和 Angela Fan。2020年。采用可扩展的多语言预训练和微调进行多语言翻译。

Nguyen Luong Tran、DuongMinh Le 和 Dat Quoc Nguyen。2021年。BARTpho：面向越南语的预训练序列到序列模型。*CoRR*，abs/2109.09701。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。注意力机制就是一切。神经信息处理系统进展，30。

GuillaumeWenzek、Marie-Anne Lachaux、Alexis Con- neau、Vishrav Chaudhary、Francisco Guzmán、Ar- mand Joulin 和 Edouard Grave。2020年。CCNet：从网络爬取数据中提取高质量单语数据集。载于 第*12*届语言资源与评估会议论文集，第4003–4012页，法国马赛。欧洲语言资源协会。

Adina Williams、Nikita Nangia 和 Samuel Bowman。2018年。一个通过推理理解句子的广覆盖挑战语料库。载于*2018*年北美计算语言学协会分会年会：人类语言技术论文集，第*1*卷（长篇论文），第1112–1122页，新奥尔良，路易斯安那州。计算语言学协会。

Tianyi Zhang、Varsha Kishore、FelixWu、Kilian Q. Weinberger 和 Yoav Artzi。2019年。Bertscore：使用 BERT 评估文本生成。*CoRR*，abs/1904.09675。

# 附录

## A 附录——GreekSUM摘要

在本附录部分，我们展示了从 GreekSUM 摘要测试集中随机选取的 5 个文档的参考文本和模型摘要。

| | | |
|---|---|---|
| | 文档 | «Ο κύβος ερρίφθη. Ο Άρμιν Λάσετ θα είναι ο υποψήφιος Καγκελάριος για την Χριστιανική Ένωση», δήλωσε πριν από λίγο ο Αρχηγός της Χριστιανοκοινωνικής Ένωσης (基社盟) και Πρωθυπουργός της Βαυαρίας Μάρκους Ζέντερ, αναγνωρίζοντας το αποτέλεσμα της ψηφοφορίας του προεδρείου του Χριστιανοδημοκρατικού Κόμματος (基民盟), το οποίο σε ποσοστό 77,5% τάχθηκε υπέρ της υποψηφιότητας του κ. Λάσετ. Πριν από την συνεδρίαση του προεδρείου του 基民盟, ο κ. Ζέντερ είχε δηλώσει ότι παραχωρεί στο 基民盟 το προβάδισμα στην επιλογή του υποψήφιου Καγκελάριου της Χριστιανικής Ένωσης (基民盟/基社盟) και σήμερα επανέλαβε ότι δέχεται το αποτέλεσμα «χωρίς μηνισικακία» και ότι τάσσεται υπέρ της ενότητας της Χριστιανικής Ένωσης. |
| ABSTRACT | Gold | Ο Άρμιν Λάσετ θα είναι ο υποψήφιος των 基民盟 και 基社盟 για την καγκελαρία της Γερμανίας στις εκλογές του Σεπτεμβρίου. |
| | BART-random | Ο Αρμιν Λάσετ θα είναι ο υποψήφιος πρωθυπουργός της Χριστιανικής Ένωσης, μετά από σχετική συνεδρίαση. |
| | mBART25 | Ο πρωθυπουργός της Βαυαρίας δέχθηκε το αποτέλεσμα της ψηφοφορίας του προεδρείου του 基民盟, το οποίο σε ποσοστό 77, 5% τάχθηκε υπέρ της υποψηφιότητας του Αρμιν Λάσετ. |
| | mBART50 | Σε ποσοστό 77, 5% τάχθηκε υπέρ της υποψηφιότητας του Αρμιν Λάσετ στο προεδρείο του 基民盟, ο Πρωθυπουργός της Βαυαρίας Μάρκους Ζέντερ. |
| | GreekBART | Υπέρ του Αρμιν Λάσετ τάσσεται ο Μάρκους Ζέντερ, αναγνωρίζοντας το αποτέλεσμα της ψηφοφορίας του προεδρείου του基民盟. |

表9：示例1-GreekSUM 摘要

| | | |
|---|---|---|
| | 文档 | Κλειστή είναι η λεωφόρος Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί κυκλοφοριακό πρόβλημα, καθώς έχει χυθεί μεγάλη ποσότητα λαδιού από φορτηγό, στην συμβολή με την λεωφόρο Βασ. Όλγας και είναι επικίνδυνη η διέλευση των οχημάτων. Η Τροχαία έχει διακόψει την κυκλοφορία στο κατερχόμενο ρεύμα στο ύψος της οδού Ριζάρη και στο ανερχόμενο από την αρχή της Αρδητού και κάνει εκτροπή, αλλά έχει δημιουργηθεί μποτιλιάρισμα. Στο σήμειο που έχουν χυθεί τα λάδια βρίσκονται συνεργεία του Δήμου, που ρίχνουν πριονίδι και άλλα υλικά για να αντιμετωπίσουν την ολοσθηρότητα του οδοστρώματος και να αποκατασταθεί η κυκλοφορία。 |
| ABSTRACT | Gold | Η Τροχαία έχει διακόψει την κυκλοφορία στο κατερχόμενο ρεύμα στο ύψος της οδού Ριζάρη και στο ανερχόμενο από την αρχή της Αρδητού και κάνει εκτροπή- Μποτιλιάρισμα στο σημείο。 |
| | BART-random | Η Τροχαία, που έχει δημιουργηθεί στο ύψος του Παναθηναϊκού, έχει διακοπεί την κυκλοφορία των οχημάτων στην λεωφόρο Βασ.Α. |
| | mBART25 | Κλειστή είναι η λεωφόρος Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί κυκλοφοριακό πρόβλημα. |
| | mBART50 | Κυκλοφοριακό πρόβλημα στο ύψος του Παναθηναϊκού Σταδίου, καθώς έχει χυθεί μεγάλη ποσότητα λαδιού από φορτηγό σε λεωφόρο Βασ. Κωνσταντίνου. |
| | GreekBART | Κυκλοφοριακό πρόβλημα έχει δημιουργηθεί στην Λεωφόρο Βασ. Κωνσταντίνου στο ύψος του Παναθηναϊκού Σταδίου, με αποτέλεσμα να έχει δημιουργηθεί μποτιλιάρισμα. |

表 10：示例 2-GreekSUM 摘要

| | | |
|---|---|---|
| | 文档 | Η Καγκελάριος Άνγκελα Μέρκελ δεν θα παραστεί στην επίσημη δεξίωση που θα παραθέσει την Παρασκευή ο Ομοσπονδιακός Πρόεδρος Φρανκ-Βάλτερ Σταϊνμάιερ προς τιμήν του Προέδρου της Τουρκίας Ρετζέπ Ταγίπ Ερντογάν, σύμφωνα με κυβερνητικές πηγές τις οποίες επικαλείται το περιοδικό 《明镜周刊》。Η δεξίωση αλλά και οι στρατιωτικές τιμές με τις οποίες θα υποδεχθεί τον προσκεκλημένο του ο Γερμανός Πρόεδρος προκαλούν σοβαρές αντιδράσεις στον πολιτικό κόσμο της χώρας。 Η Μέρκελ είναι πάντα προσκεκλημένη του Ομοσπονδιακού Προέδρου σε δεξιώσεις ή επίσημα δείπνα που παρατίθενται προς τιμήν υψηλών προσκεκλημένων。 Η ίδια ωστόσο συνηθίζει να παρευρίσκεται μόνο σε εξαιρετικές περιπτώσεις。 Η τελευταία φορά που παρέστη σε κάτι ανάλογο ήταν το επίσημο δείπνο που είχε παρατεθεί το 2015 προς τιμήν της Βασίλισσας Ελισάβετ, ενώ την προηγούμενη χρονιά είχε παρευρεθεί στο δείπνο με τον Εμίρη του Κατάρ。 Αντιθέτως, δεν είχε παρευρεθεί στην δεξίωση προς τιμή του Κινέζου Προέδρου Σι Τζινπίνγκ το 2017。 Η Καγκελάριος όμως δεν θα είναι η μόνη που θα απορρίψει την πρόσκληση του Σταϊνμάιερ。 Ο Πρόεδρος των Φιλελευθέρων (FDP) Κρίστιαν Λίντνερ ανακοίνωσε ότι δεν σκοπεύει να παραστεί, καθώς δεν επιθυμεί«να συμμετάσχει στην προπαγάνδα του Ερντογάν». Την ίδια στάση θα τηρήσει και η εκπρόσωπος του κόμματος για την εξωτερική πολιτική, Μπιτζάν Ντζιρ-Σαράι, ενώ σύσσωμη η ηγετική ομάδα των Πρασίνων, οι συμπρόεδροι Αναλένα Μπέρμποκ και Ρόμπερτ Χάμπεκ και οι επικεφαλής της Κοινοβουλευτικής Ομάδας Κάτριν Γκέρινγκ-Έκαρτ και Άντον Χοφράιτερ, δήλωσαν ότι θα απέχουν από την δεξίωση. Το ίδιο ισχύει και για τους επικεφαλής της Εναλλακτικής για την Γερμανία (德国选择党) Άλεξάντερ Γκάουλαντ και Αλίς Βαϊντέλ και για την επικεφαλής της Κ。 Ο。 της Αριστεράς Σεβίμ Νταγκντελέν。 Αντιθέτως, την πρόθεσή του να παραστεί στην δεξίωση στο Προεδρικό Ανάκτορο 贝尔维尤 εξέφρασε ο πρώην Πρόεδρος των Πρασίνων Τζεμ Έζντεμιρ, διευκρινίζοντας ταυτόχρονα ότι ο Τούρκος Πρόεδρος “δεν είναι κανονικός Πρόεδρος και δεν αξίζει” να παρατεθεί δεξίωση προς τιμήν του。 Με την παρουσία του, δήλωσε ο κ。 Έζντεμιρ στην 《每日镜报》, ελπίζει να στείλει ένα μήνυμα τόσο προς την Τουρκία όσο και προς την τουρκογερμανική κοινότητα: 《Η αντιπολίτευση στην Γερμανία είναι μέρος της πολιτικής αυτής της χώρας, είμαστε ένα σταθερό και απαραίτητο συστατικό στοιχείο της δημοκρατίας μας。 Ο κ. Ερντογάν θα πρέπει να με ανεχθεί》。 |
| ABSTRACT | Gold | Η καγκελάριος είναι πάντα προσκεκλημένη του ομοσπονδιακού προέδρου σε δεξιώσεις ή δείπνα προς τιμήν υψηλών προσκεκλημένων, ωστόσο δίνει το παρών μόνο σε εξαιρετικές περιπτώσεις。 |
| | BART-random | Δεν θα παραστεί στην επίσημη δεξίωση που θα παραθέσει την Τουρκία προς τιμήν του Ρετζέπ Ταγίπ Ερντογάν ο εκπρόσωπος της Γερμανίας Άνγκελα Μέρκελ。 |
| | mBART25 | Αντιδράσεις από τον πολιτικό κόσμο της χώρας προκαλούν η δεξίωση που θα παραθέσει ο Φρανκ-Βάλτερ Σταϊνμάιερ προς τιμήν του Προέδρου της Τουρκίας - Δεν θα είναι η μόνη που θα απορρίψει την πρόσκληση του Σταϊνμάιερ。 |
| | mBART50 | Η Μέρκελ είναι πάντα προσκεκλημένη του Ομοσπονδιακού Προέδρου σε δεξιώσεις ή επίσημα δείπνα που παρατίθενται προς τιμήν υψηλών προσκεκλημένων。 Η ίδια ωστόσο συνηθίζει να παρευρίσκεται μόνο σε εξαιρετικές περιπτώσεις。 |
| | GreekBART | Από τον πολιτικό κόσμο της Γερμανίας。 Η Άνγκελα Μέρκελ δεν θα παραστεί στην επίσημη δεξίωση προς τιμήν του Γερμανού Προέδρου Φρανκ-Βάλτερ Σταϊνμάιερ。 |

表11：示例 3-GreekSUM 摘要

| | | |
|---|---|---|
| | 文档 | Από το 2011 και μετά αρκετοί εκατοντάδες άνθρωποι έχουν πεθάνει στην προσπάθειά τους να βγάλουν την τέλεια 自拍。 Οι περισσότεροι θάνατοι έχουν λάβει χώρα στην Ινδία。 Ακολουθεί η Ρωσία, οι Ηνωμένες Πολιτείες και ύστερα το Πακιστάν με τους νεκρούς συνολικά να φτάνουν τους 259。 Βέβαια υπάρχουν κάποια σημεία, τα οποία σύμφωνα με έρευνες, παρουσιάζουν μεγαλύτερη επικινδυνότητα, όπως το νερό και οι ψηλές κυλιόμενες σκάλες。 Οι πιο «συνηθισμένες» αιτίες θανάτου από 自拍 συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές。 Όσον αφορά τα στατιστικά στοιχεία τα 3/4 των θυμάτων είναι άνδρες και κάτω από την ηλικία των 30。 Αν και οι γυναίκες βγάζουν περισσότερες 自拍 σύμφωνα με τις μελέτες, οι άνδρες είναι πιο επιρρεπείς στον κίνδυνο. Ακόμα, οι τουρίστες είναι αυτοί που πλήττονται πιο συχνά στην προσπάθεια να βγάλουν μια φωτογραφία που θα εντυπωσιάσει τους ακολούθους τους. Οι αρχές ψάχνουν τρόπους προκειμένου να αποτρέψουν τους θανάτους. Για παράδειγμα η ρωσική αστυνομία μοίρασε φυλλάδια, τα οποία εμπεριείχαν προειδοποιήσεις σχετικά με τους κινδύνους που«καραδοκούν» πίσω από μια 自拍. Στις Ηνωμένες Πολιτείες, τα εθνικά πάρκα έχουν εκδώσει οδηγούς για το πώς να βγάζεις «ασφαλείς» 自拍, ενώ στην Ινδία υπάρχουν επίσημα σχεδιασμένες πινακίδες που προειδοποιούν για υψηλού κινδύνου περιοχές ή αλλιώς “禁止自拍区”。 Αν και η εμμονή με τις 自拍 δεν φαίνεται να περνάει οι αρχές κάνουν ότι μπορούν για να περιορίσουν την επικινδυνότητα και τους θανάτους。 |
| ABSTRACT | Gold | Οι πιο“συνηθισμένες” αιτίες θανάτου από σελφιε συμπεριλαμβάνουν πνιγμό, πτώση, και τη σύγκρουση με κινούμενο όχημα - Άνδρες κάτω των 30 τα περισσότερα θύματα。 |
| | BART-random | Οι Ηνωμένες Πολιτείες, Ινδία, Αν。 και Πακιστάν και Αν。 Ινδία αναζητούν αναζητούν στοιχεία για να βγάλουν την τέλεια 自拍 τους στην προσπάθειά τους。 |
| | mBART25 | Η Ινδία μετράει τους 259 θανάτους από 自拍,τα οποία συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές。 Οι αρχές ψάχνουν τρόπους προκειμένου να αποτρέψουν τους θανάτους。 |
| | mBART50 | Στην Ινδία, τα εθνικά πάρκα έχουν εκδώσει οδηγούς για το πώς να βγάζεις«ασφαλείς» 自拍, ενώ στην Ινδία υπάρχουν επίσημα σχεδιασμένες πινακίδες που προειδοποιούν για υψηλού κινδύνου περιοχές. |
| | GreekBART | Πολλοί άνθρωποι έχουν πεθάνει στην προσπάθειά τους να βγάλουν μια 自拍, με τις “συνηθισμένες” αιτίες να συμπεριλαμβάνουν τον πνιγμό, την πτώση, τη σύγκρουση με κινούμενο όχημα και τις φωτιές. |

表12：示例4-GreekSUM 摘要

| | | |
|---|---|---|
| 文档 | | Στην απώλεια του Μίκη Θεοδωράκη αναφέρθηκε ο πρωθυπουργός Κυριάκος Μητσοτάκης στην έναρξη της συνεδρίασης του Υπουργικού Συμβουλίου, κηρύσσοντας τριήμερο εθνικό πένθος. Ο πρωθυπουργός ειδικότερα δήλωσε: "Τη σημερινή μας συνεδρίαση σκιάζει δυστυχώς μία πολύ θλιβερή είδηση: Ο Μίκης Θεοδωράκης περνά πια στην αιωνιότητα. Η φωνή του σίγησε και μαζί του σίγησε και ολόκληρος ο Ελληνισμός. Όπως είχε γραφτεί και για τον Παλαμά, 《όλοι είχαμε ξεχάσει πως είναι θνητός》。 Όμως, μας αφήνει παρακαταθήκη τα τραγούδια του, την πολιτική του δράση, αλλά και την εθνική του προσφορά σε κρίσιμες στιγμές。 Η Ρωμιοσύνη σήμερα κλαίει。 Και γι' αυτό και με απόφαση της κυβέρνησης από σήμερα κηρύσσεται τριήμερο εθνικό πένθος。 Όπως ξέρετε, είχα την τιμή να τον γνωρίζω για πολλά χρόνια και σχετικά πρόσφατα μάλιστα τον είχα επισκεφτεί。 Οι συμβουλές του ήταν πάντα πολύτιμες για μένα, κυρίως αυτές που αφορούσαν στην ενότητα του λαού μας και στην υπέρβαση των διαχωριστικών γραμμών。 Πιστεύω πως η καλύτερη τιμή προς αυτόν τον παγκόσμιο Έλληνα θα είναι εμείς, με το καθημερινό μας έργο, να κάνουμε πράξη αυτό ακριβώς το μήνυμά του。 Ο Μίκης είναι η Ιστορία μας και πρέπει να τη συνεχίσουμε όπως θα ήθελε και εκείνος。" Πέθανε ο Μίκης Θεοδωράκης - Ορφάνεψε η Ρωμιοσύνη Έλενα Ακρίτα - Ο 'Ηλιος (που κρύφτηκε) και ο Χρόνος(που χάθηκε), Μίκη Μίκης Θεοδωράκης: Τα 5 τραγούδια του σπουδαίου μουσικού που 《μιλούν》 στην ψυχή της Ελλάδας Ο πολιτικός Μίκης Θεοδωράκης: Πάντα στο πλευρό των απλών ανθρώπων。 |
| ABSTRACT | Gold | Η Ρωμιοσύνη σήμερα κλαίει δήλωσε ο πρωθυπουργός στην έναρξη της συνεδρίασης του υπουργικού συμβουλίου αναφερόμενος στο θάνατο του Μίκη Θεοδωράκη。 |
| | BART-random | Ο πρωθυπουργός κατά την έναρξη της συνεδρίασης του Υπουργικού Συμβουλίου κηρύσσοντας την απώλεια του Μίκη Θεοδωράκη。 |
| | mBART25 | Ο πρωθυπουργός Κυριάκος Μητσοτάκης απο το υπουργικό συμβούλιο για τον θάνατο του Μίκη Θεοδωράκη。 |
| | mBART50 | Τριήμερο εθνικό πένθος κηρύχθηκε στη συνεδρίαση του υπουργικού συμβουλίου, με τον πρωθυπουργό να σημειώνει ότι ο Μίκης Θεοδωράκης περνά πια στην αιωνιότητα。 |
| | GreekBART | Το δικό του μήνυμα για την απώλεια του Μίκη Θεοδωράκη έστειλε ο πρωθυπουργός Κυριάκος Μητσοτάκης κατά τη συνεδρίαση του Υπουργικού Συμβουλίου。 |

表 13：Example 5-GreekSUM 希腊语摘要

## B 附录-GreekSUM 标题

在附录的第二部分，我们展示了从 GreekSUM 标题测试集中随机选取的 5 篇文档的参考文本标题和模型生成标题。

| | |
|---|---|
| 文档 | Ένας 33χρονος έχασε τη ζωή του, ύστερα από σύγκρουση δύο αυτοκίνητων, έξω από τη Θεσσαλονίκη. Όπως έγινε γνωστό, το θανατηφόρο τροχαίο συνέβη στις 2.15 μετά τα μεσάνυχτα σε παράδρομο της Εγνατίας Οδού, στο ύψος του Ωραιοκάστρου. Σύμφωνα με την Αστυνομία, ο 33χρονος, οδηγός του ενός οχήματος, διακομίστηκε στο νοσοκομείο Παπαγεωργίου, όπου όμως λίγη αργότερα υπέκυψε στα τραύματά του, ενώ η οδηγός του άλλου οχήματος υπέστη ελαφρά τραύματα. Οι ακριβείς συνθήκες υπό τις οποίες προκλήθηκε η σύγκρουση ερευνώνται από το αρμόδιο τμήμα τροχαίας. |
| TITLE Gold | Τροχαίο δυστύχημα στη Θεσσαλονίκη με έναν νεκρό |
| TITLE BART-random | Τροχαίο έξω από τη Θεσσαλονίκη - Δύο τραυματίες |
| TITLE mBART25 | Θεσσαλονίκη: Νεκρός 33χρονος ύστερα από σύγκρουση δύο αυτοκίνητων |
| TITLE mBART50 | Θεσσαλονίκη: Νεκρός 33χρονος ύστερα από σύγκρουση δύο αυτοκίνητων |
| TITLE GreekBART | Τροχαίο στη Θεσσαλονίκη: Νεκρός 33χρονος σε παράδρομο |

表 14：示例 1-GreekSUM 标题

| | |
|---|---|
| 文档 | Ολες οι χώρες της Ευρωπαϊκής Ενωσης συμφωνούν ότι δεν θα πληρώσουν τη Ρωσία απευθείας σε ρούβλια για τις εισαγωγές ρωσικού φυσικού αερίου, δήλωσαν υψηλόβαθμοι ευρωπαίοι αξιωματούχοι, σημειώνοντας ότι οι επόμενες πληρωμές είναι προγραμματισμένες για τις 20 Μαΐου. «Αυτό που γνωρίζουμε, και υπάρχει συναίνεση επάυτού μεταξύ όλων των κρατών μελών, είναι ότι κανείς δεν είναι πρόθυμος να πληρώσει σε ρούβλια», δήλωσε ο ένας αξιωματούχος κατά την διάρκεια ενημέρωσης των δημοσιογράφων και προσθέτοντας ότι η Ευρωπαϊκή Επιτροπή δεν γνωρίζει πόσοι αγοραστές έχουν ανοίξει λογαριασμούς για πληρωμές προμήθειας φυσικού αερίου μέσω της俄罗斯天然气工业银行. Στο μεταξύ, ανώτερος αξιωματούχος της Ευρωπαϊκής Ένωσης δήλωσε πως και μόνο το άνοιγμα τραπεζικού λογαριασμού σε ρούβλια στην 俄罗斯天然气工业银行 ενδέχεται να αποτελεί παραβίαση των κυρώσεων που έχει επιβάλει η ΕΕ σε βάρος της Ρωσίας, όμως η ΕΕ δεν έχει ένδειξη πως κάποια εταιρεία φυσικού αερίου της ΕΕ έχει κάνει κάτι τέτοιο。 Ο αξιωματούχος δήλωσε πως 《ex πρώτης όψεως》 το άνοιγμα τραπεζικών λογαριασμών σε ρούβλια από εισαγωγείς φυσικού αερίου φαίνεται ότι παραβιάζει τις κυρώσεις。 Ο αξιωματούχος πρόσθεσε πως η Ευρωπαϊκή Επιτροπή δεν έχει κάποια επίσημη ένδειξη ότι εταιρείες της ΕΕ έχουν δημιουργήσει στην俄罗斯天然气工业银行 λογαριασμούς σε ρούβλια για την πληρωμή του φυσικού αερίου. Επίσης διευκρίνισε πως η Πολωνία και η Βουλγαρία χρησιμοποίησαν τις υφιστάμενες μεθόδους πληρωμής για το ρωσικό αέριο, πριν η Μόσχα αναστείλει χθες, Τετάρτη, τις προμήθειες των χωρών αυτών με αέριο, και πως δεν χρησιμοποίησαν τον μηχανισμό που προτείνει η Μόσχα για να πληρώσουν σε ρούβλια。 “Σύμφωνα με τις πληροφορίες μας, αμφότερες οι χώρες επέμειναν στην αρχική μορφή πληρωμής», δήλωσε ο αξιωματούχος σε δημοσιογράφους. Ωστόσο δύο πηγές είπαν σήμερα στο Ρόιτερς ότι λίγες ευρωπαϊκές εταιρείες έχουν αρχίσει να πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο, αν και μεγάλοι πελάτες της δεν το έχουν κάνει ακόμη。 “Μερικές εμπορικές εταιρείες, ίσως περισσότερες από πέντε, έχουν αρχίσει τις πληρωμές», είπε μία πηγή, ζητώντας να μην κατονομαστεί, επειδή δεν είχε εξουσιοδοτηθεί να μιλήσει στα μέσα ενημέρωσης. |
| TITLE Gold | Φυσικό αέριο: Όλες οι χώρες της ΕΕ συμφωνούν ότι δεν θα πληρώσουν τη Ρωσία σε ρούβλια |
| BART-random | Ε.Ε.: «Δεν θα πληρώσουν» οι χώρες της ΕΕ για το φυσικό αέριο σε ρούβλια |
| mBART25 | ΕΕ: Οι χώρες δεν πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο |
| mBART50 | ΕΕ: Οι χώρες δεν πληρώνουν σε ρούβλια τη Ρωσία για το φυσικό αέριο |
| GreekBART | ΕΕ: Δεν θα πληρώσουμε τη Ρωσία σε ρούβλια για το φυσικό αέριο |

表 15：示例 2-GreekSUM 标题

| | | |
|---|---|---|
| 文档 | | Στις ημέρες του Πάσχα έχει προσαρμοστεί το πρόγραμμα λειτουργίας λεωφορείων, τρόλεϊ, ηλεκτρικού και μετρό. Ειδικότερα, τα λεωφορεία και τα τρόλεϊ σήμερα, Μεγάλη Παρασκευή, θα κινούνται με πρόγραμμα Σαββάτου. Οι συρμοί στο μετρό θα διέρχονται από τους σταθμούς ανά 7 λεπτά από τις 09.00 έως τις 17.00 και ανά 10 λεπτά τις υπόλοιπες ώρες. Υπενθυμίζεται πως δεν θα ισχύσει η δίωρη παράταση λειτουργίας που εφαρμόζεται τις Παρασκευές. Στον ηλεκτρικό οι συρμοί θα διέρχονται από τους σταθμούς ανά 10, 5 λεπτά. Τα λεωφορεία και τα τρόλεϊ θα κινηθούν με πρόγραμμα Κυριακής, ενώ θα αποσυρθούν νωρίτερα, ώστε να βρίσκονται στα αμαξοστάσια στις 23.00. Τα λεωφορεία θα κινηθούν με πρόγραμμα Κυριακής και τα τρόλεϊ με ειδικό πρόγραμμα Κυριακής. Τόσο στα δρομολόγια των λεωφορείων όσο και σ΄ αυτά των τρόλεϊ θα εφαρμοστεί ειδικό πρόγραμμα Σαββάτου. Ακινητοποιημένοι θα μείνουν την Τετάρτη 1η Μαΐου οι συρμοί του ηλεκτρικού (πρώην ΗΣΑΠ), τα λεωφορεία, τα τρόλεϊ, αλλά και ο σιδηρόδρομος, λόγω 24ωρης απεργίας των εργαζομένων, που θα συμμετάσχουν στις απεργιακές συγκεντρώσεις για την Πρωτομαγιά. Όπως αναφέρουν σε ανακοίνωσή τους οι εργαζόμενοι στον πρώην ΗΣΑΠ, 「είναι μέρα αγώνα, τιμής και μνήμης. Θυμόμαστε και τιμάμε τους πρωτοπόρους αγωνιστές και τα θύματα των εργατικών αγώνων για βελτίωση των συνθηκών δουλειάς για αξιοπρεπείς αμοιβές και την κατοχύρωση των δικαιωμάτων μας. Ανασυγκροτούμαστε, θέτουμε τους στόχους μας και προχωράμε σε νέους αγώνες. Διεκδικούμε και παλεύουμε για την αναπλήρωση απωλειών από τις μνημονιακές πολιτικές λιτότητας, για πραγματικές αυξήσεις στους μισθούς και στις κοινωνικές παροχές». Και προσθέτουν «υπερασπιζόμαστε τον δημόσιο χαρακτήρα των συγκοινωνιών. Διεκδικούμε την υπογραφή νέας Συλλογικής Σύμβασης Εργασίας. Αγωνιζόμαστε για ασφαλείς, φθηνές συγκοινωνίες. Με αγώνες κατακτάμε τα δικαιώματά μας». |
| TITLE | Gold | Πάσχα 2019: Πώς θα κινηθούν λεωφορεία, τρόλεϊ, ηλεκτρικός και μετρό |
| | BART-random | Μέσα Μαζικής Μεταφοράς: Πώς θα κινηθούν σήμερα τα Μέσα Μεταφοράς |
| | mBART25 | Μέσα Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικό και μετρό |
| | mBART50 | Μέσα Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικό και μετρό |
| | GreekBART | Πάσχα: Πώς θα κινηθούν σήμερα λεωφορεία, τρόλεϊ, ηλεκτρικού και μετρό |

表16：示例3-GreekSUM 标题

| | |
|---|---|
| 文档 | Συνάντηση με οικονομικούς παράγοντες από το Σίτι του Λονδίνου έχει αυτή την ώρα ο Αλέξης Τσίπρας στο κέντρο της βρετανικής πρωτεύουσας. Τον Έλληνα πρωθυπουργό υποδέχθηκε ο αντιπρόεδρος της Επιτροπής Πολιτικής του Σίτι, Τομ Σλέι (汤姆·斯莱) Επισημαίνεται ότι η Επιτροπή υπέχει θέση Διοίκησης του Σίτι του Λονδίνου. Από την αίθουσα της 《Παλιάς Βιβλιοθήκης》, ο πρωθυπουργός θα απευθυνθεί σε έναν κύκλο περισσότερων από εκατό σημαίνοντων στελεχών της επενδυτικής/χρηματοπιστωτικής κοινότητας του Σίτι και, σύμφωνα με πληροφορίες στη συνέχεια θα ακολουθήσει συνάντηση σε πιο στενό κύκλο συμμετεχόντων. Στον απόηχο της απόφασης του 欧元集团 για την ελάφρυνση του χρέους, οι επαφές του Αλέξη Τσίπρα με σημαντικούς εκπροσώπους της επενδυτικής/χρηματοπιστωτικής της κοινότητας του οικονομικού κέντρου της Ευρώπης, σηματοδοτούν ένα ευκρινές διεθνές μήνυμα για τις προοπτικές της ελληνικής οικονομίας και της«επόμενης μέρας», στην περίοδο μετά την ολοκλήρωση των μνημονίων. Όπως ανέφερε κυβερνητικός αξιωματούχος, οι σημερινές συναντήσεις είναι ένας σημαντικός σταθμός σε μια«αλυσίδα» επαφών και συνομιλιών που θα συνεχιστούν στο αμέσως επόμενο διάστημα των καλοκαιρινών μηνών και το φθινόπωρο. Ενδεικτική της ευνοϊκής συγκυρίας για την ελληνική οικονομία και το στοίχημα της ανάκαμψης, η χθεσινοβραδινή αναβάθμιση, από τον αμερικανικό οίκο αξιολόγησης 标准普尔 της μακροπρόθεσμης πιστοληπτικής ικανότητας της χώρας σε B+, χαιρετίζοντας την απόφαση του 欧元集团。 Στις 18:00 το απόγευμα ώρα Ελλάδας, ο πρωθυπουργός θα περάσει το κατώφλι της 唐宁街 10 προκειμένου να συναντηθεί με την πρωθυπουργό της Βρετανίας, Τερέζα Μέι. Στη συνέχεια θα έχει συνάντηση με τον αρχηγό του Εργατικού Κόμματος, Τζέρεμι Κόρμπιν. |
| TITLE<br>Gold<br>BART-random<br>mBART25<br>mBART50<br>GreekBART | Συνάντηση με οικονομικούς παράγοντες από το Σίτι του Λονδίνου έχει ο Αλέξης Τσίπρας<br>Μήνυμα Τσίπρα στο Λονδίνο για το χρέος<br>Συνάντηση Τσίπρα με οικονομικούς παράγοντες στο Σίτι<br>Συνάντηση Τσίπρα με οικονομικούς παράγοντες στο Σίτι<br>Βλέμματα στο Λονδίνο για την ελληνική οικονομία |

表17：示例4-GreekSUM 标题

| | |
|---|---|
| 文档 | Επιβατικό τρένο εκτροχιάστηκε σήμερα περίπου 20 χλμ. βόρεια της Ραμπάτ, με αποτέλεσμα να σκοτωθούν έξι άνθρωποι και άλλοι 86 να τραυματιστούν, σύμφωνα με επίσημο απολογισμό που ανακοινώθηκε στον τόπο του δυστυχήματος. «Ο εκτροχιασμός προκάλεσε έξι θανάτους, σύμφωνα με τον τρέχοντα απολογισμό, και 86 τραυματίες σε σοβαρή κατάσταση», δήλωσε ο Μοχάμεντ Ραμπί Ραχίλ, γενικός διευθυντής της εταιρίας σιδηροδρόμων ONCF, ο οποίος μετέβη επί τόπου。《Ξεκίνησε έρευνα για τον προσδιορισμό των αιτιών του δυστυχήματος», πρόσθεσε, σε βίντεο που αναρτήθηκε στα μέσα κοινωνικής δικτύωσης。 Θεαματικές εικόνες του δυστυχήματος、 που σημειώθηκε γύρω στις 13:00 ώρα Ελλάδας, περίπου 20 χλμ. βόρεια της πρωτεύουσας Ραμπάτ, στο ύψος της κοινότητας Σιντί Μπουκναντέλ, κάνουν τον γύρο των μέσων κοινωνικής δικτύωσης, που είναι πολύ επικριτικά εναντίον της ONCF。 Οι εικόνες δείχνουν πολλά βαγόνια εκτροχιασμένα κοντά σε μια γέφυρα στους αγρούς、 ενώ η μηχανή είναι πλήρως κατεστραμμένη。 Ο οδηγός της αμαξοστοιχίας είναι νεκρός、 σύμφωνα με πολλά τοπικά ΜΜΕ。 Ο βασιλιάς αποφάσισε να αναλάβει τα έξοδα της κηδείας των θυμάτων και οι τραυματίες θα διακομιστούν στο στρατιωτικό νοσοκομείο της Ραμπάτ με βασιλικές οδηγίες、 αναφέρεται σε ανακοίνωση του γραφείου του βασιλιά。 |
| TITLE Gold | Εκτροχιασμός τρένου στο Μαρόκο: Στους 6 οι νεκροί- 86 τραυματίες |
| BART-random | Ραμπάτ: 20 νεκροί από εκτροχιασμό τρένου |
| mBART25 | ΗΠΑ: Επιβατικό τρένο εκτροχιάστηκε - Έξι νεκροί και 86 τραυματίες |
| mBART50 | ΗΠΑ: Επιβατικό τρένο εκτροχιάστηκε - Έξι νεκροί και 86 τραυματίες |
| GreekBART | Εκτροχιασμός τρένου στη Ραμπάτ: Έξι νεκροί και 86 τραυματίες |

表18：Example 5-GreekSUM 标题