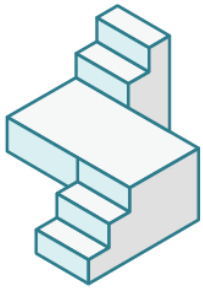
BRIKO WHITEPAPER



Bríko

Briko Realistic International
Knowledge Organization

Version:0.6

草案 @ March 04, 2019

Briko项目概述

什么是Briko？

Briko是一个基于区块链技术的开放语料库，以及基于该语料库的人工智能开发平台。

Briko还是一个语料贡献者、人工智能模型设计者、应用软件开发者和使用者多方共赢的合作社区。

Briko的价值是什么？

Briko平台的意义在于：

- 为语料库的建设提供全新的驱动模式，显著降低语料库获取的难度和维护成本
- 使人工智能模型和算法能更快的为实际应用直接服务
- 降低广大开发者使用机器学习模型及进行自然语言处理（NLP）应用开发的门槛

Briko平台构建了一个完整的NLP产业生态结构，并且所有参与方都能从中获得收益：

- 大众贡献语料数据
- 研究人员提供人工智能模型和算法
- 软件工程师提供软件产品

谁是Briko的受益者？

- Briko为中小企业、开发者提供价格合理、可定制的语料库及NLP机器学习模型的API接口，使他们能够快速低成本的开发应用。
- Briko向语料贡献者支付看得见的劳动回报。
- Briko为NLP及其他人工智能科研领域的模型和算法提供开放的使用接口。
- Briko是区块链社区中为数不多的有劳动支撑价值的落地项目。

背景与问题

为什么需要语料库？ -- 基于深度学习的自然语言处理

自然语言处理（Natural Language Processing，NLP）正在成为人工智能领域的一个重要分支，它关注如何处理与应用人类交流表达使用的语言。它的应用给诸如客服系统，电子邮件回复，电话交谈等人与人的交流场景，以及智能应用，IoT设备，甚至网络搜索等人机交互应用，都带来了技术上的革新和进步。

过去由于技术的限制，自然语言处理远远不能达到人与人自然交流的水平，因此应用场景非常受限，而深度学习将自然语言处理带入新纪元。新出现的预训练方法，例如词向量（word2vec）[1,2]，GloVe[3]，FastText[8]，以及seq2seq[6,7]，Transformer[4,5]等基于深度学习的模型，乃至迁移学习（Transfering Learning）[9]，对偶学习（Dual Learning）[10]等方法，都在不断提升机器NLP的表现。

基于深度学习的NLP模型，其性能主要有三方面决定：一是模型架构及训练方法，二是训练集，也就是语料，三是部署的软硬件环境。模型架构及训练方法有大量的学术文章详细介绍，内容或由作者开源，或有开发者自己根据论文实现后开源供社区参考，使用者有足够的资源选择合适自己应用场景的模型及训练方法。训练出一个优秀模型，真正的壁垒之一在于语料库，而在生产环境中部署，还有软硬件上的壁垒。

基于深度学习的NLP的表现虽然超越了传统方法， 但在工业界中的部署与应用依然有限。除了上边提到的壁垒，业界主要面临的挑战是，基于深度学习的方法在开发成本，时间和难度上并不清晰，且对于设备算力也有极高要求，因此很多公司并没有快速的转向深度学习的方法。

需要什么样的语料库？ -- 语料库的现状和技术难点

语言模型依赖语料作为“原材料”来进行训练，语料库的质量直接影响到模型的表现。为了达到最贴合语境的效果，除了选择最合适的模型与训练方法，我们还需要海量内容的，符合具体语境的，经精细标记、筛选、校验的，实时更新的语料库对模型进行训练。如此才可以让模型给出当前语境下更合理的预测，并能够应对最新的词语及表达方式。

优秀的语料库，尤其是细分学科和具体场景下的语料库，通常被大企业掌握，中小企业和开发者即便可以使用，也需要付出高额的代价。传统的开源语料库由于难以有效连接使用者，缺乏对语料贡献者和检验者的激励，很难保证高质量，高精度和及时的更新。而一旦高质量的语料库数据垄断在大企业手中，势必造成技术上的垄断，扼杀创新。

建设和维护一个庞大的语料库非常困难，需要大量人工来添加、筛选、修改、审核语料。现有的开源语料库不足以满足各种需求，且很难有足够的资源帮助维护，我们亟需一个完善的激励体系让更多人加入到语料库建设的工作中来。

为什么语料库要使用区块链技术？

区块链技术的出现将在全球范围内彻底改变工业和商业并推动经济变革，通过互联网“实现经济价值的低成本转移”[11]和灵活分配，让前文中的这些问题有了全新的解决思路。

- 区块链上存储的数据是开放且不可篡改的，接受所有人的见证，不被单一公司或机构垄断。这使语料、模型和应用的归属权有据可查；语料的劳动贡献及使用历史可追溯；侵权和恶意提交语料行为也很容易被发现和取证。
- 智能合约用代码清晰地明确了劳动价值如何分配；所有的合约执行的记录，都保存在链上，无法被篡改或撤回。一份语料素材的经济价值转移过程对所有利益相关者都是完全透明的。这保护了语料贡献者的利益，使得来自使用者支付的价值能公平地转移给贡献者。
- 借助开放共生互利的生态和新型收益分配模式，可在语料库的贡献者和使用者之间建立前所未有的连接，推动语料贡献/审核模式上的革新。保证语料库能根据使用者的需求，得到高质量，精准和及时的更新。语料及模型的贡献者和使用者因收益分配模式的变化，会更倾向于协作互利，去实现价值的共享和放大。
- 基于区块链技术的架构有利于第三方开发者，吸引更多开发者的投入，构建基于Briko平台上各种资源的第三方应用、工具、插件，和Briko团队一起完善平台，扩展语料和模型在产品化上的多样性，从长远上会为整个行业带来种种积极的可能。

机遇与方案

区块链技术的出现有望帮助我们建立一个自由开放的语料市场和人工智能模型的平台，并将语料和模型的贡献者和使用者都纳入到语料库建设的闭环生态中。通过区块链技术，Briko平台可以重新定义价值分配，整合NLP相关产业上下游各环节的利益关系，实现几方之间低成本的价值转移，并通过公平透明的智能合约最大化各方利益，激励之前无法获得足够驱动的贡献者参与进来，改变语料库建设的驱动力。

对于语料和模型的贡献者，被确认有效的劳动将与相关贡献者建立确权关系，同时他们会直接得到BGE (Briko Granted Endorsement) 奖励作为付出劳动的凭证。当Briko平台的语料与服务被购买使用时，Briko将通过二级市场公开回购的方式将Briko的收入转移给BGE持有者。Briko作为一家注册在加拿大的非营利组织，加拿大税务局（CRA）的监管保证Briko的运作是非营利性质的，我们也会公开Briko的账目供社区监督，这样保证了平台收入向贡献者的最大化转移，实现了贡献者和使用者之间最公平的价值转移。而贡献越多，质量越高，可期待的收入便越丰厚。权益明确的语料贡献机制将不断吸引贡献者参与其中，提供最新最优质的语料数据。

对于语料的使用者，他们可以像在自选超市一样，根据标签和分类选择自己需要的细分语料库和模型库，购买、定制语料和服务。量身定制的语料库可以训练模型达到特定场景下最优的表现，使用

者也不再需要从数据垄断的大企业高价购买语料。对于希望将人工智能直接应用在自己产品中的开发者，可以使用Briko平台上根据各种语料训练出的不同模型。这为中小规模的开发团队节省了大量的研发成本。

在应用层面，我们希望整合优秀的开源自然语言处理模型及框架，并利用Briko平台高质量、有时效性的语料库，为使用者搭建一个人工智能平台，提供易用、可靠、灵活的模型训练框架以及API接口，供使用者定制训练自己的模型，或直接调用Briko提供的API接口，快速地开发及部署自然语言处理相关产品。

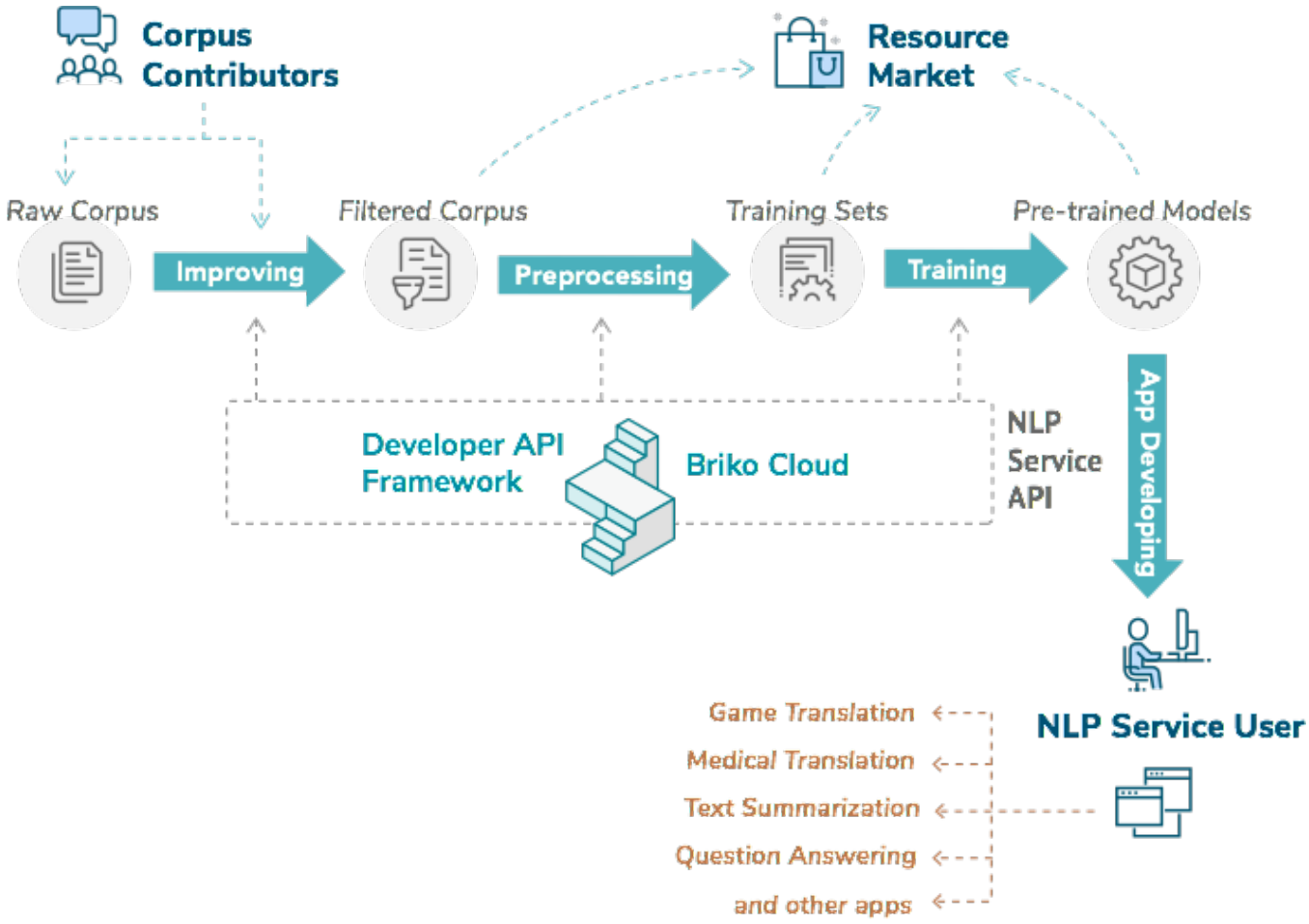
Briko平台的潜在使用者包括商业机构和学术机构。

对于商业机构，他们可以将自己的专业语料库纳入Briko开发平台，在获取BGE奖励的同时，也会有更多人加入到专业语料库的建设中。同时Briko的开发平台为他们提供了低成本试错的机会，使他们以最小的代价快速的搭建、测试模型，争取占领市场的先机。而对于中小企业，乃至个人开发者，他们可以低成本的使用优质的语料资源，同时Briko开发平台的自有模型和API接口也极大地降低了技术门槛，也使得NLP领域的创新不再限于大公司和少数核心人才。

对于学术机构，Briko开发平台将提供标准的语料库与模型，为学者提供各模型表现的基准线（baseline），方便研究成果间的比较。同时，科研学术界可以在平台上看到自己设计的人工智能模型和算法被哪些领域应用，在应用时存在哪些问题，为未来的科研方向提供指引。

高质量低成本的语料库，以及开放的开发平台将极大降低开发门槛，促进自然语言处理相关应用的创新与繁荣。

技术与架构



系统架构

Briko系统由基于区块链技术社区协作维护的Briko语料库，以及可以使用该语料库的Briko人工智能开发平台构成。

语料库

Briko将基于EOS开源项目搭建一套区块链系统运行语料库软件，这套区块链系统将作为EOS主网的侧链存在，以丰富EOS主网生态为目的。语料库的软件主体将基于EOS智能合约技术搭建，实现一套具有高度安全性、高效、并有很高自动化程度的软件架构。大部分主要功能，例如语料库相关任务发放，任务领取，任务提交，任务结算，BGE回购等，将由智能自动合约完成，所有簿记工作和资产管理，也将全部记录到链上，供查询和公示。所有Briko项目涉及的智能合约源代码将全部开源。

为了实现如下目的（不限于）：

- 劳动量计酬（语料翻译，整理，打分等）

- 惩罚恶意使用者，保护经济系统
- 语料库使用者付费（API使用，翻译模型使用，语料下载等）

Briko将实现一个基于代币（BGE）的经济系统（后详）。

在提供网页和移动端应用的同时， Briko将提供基于rest的语料库访问API，通过访问Briko语料API， 开发者可以将语料上传、审核等任务嵌入在自己的应用之中，在实现人机识别等任务的同时，获取BGE奖励。

Briko语料库系统将以EOS智能合约形式提供对EOS侧链上语料库的访问、贡献和使用及授权记录。

Briko人工智能开发平台

Briko的另一部分是基于链上语料库的开源人工智能开发平台。平台框架内融合主流的自然语言处理模型，并将它们模块化、通用化，提供一个开发平台供使用者搭建和训练自己的定制模型。同时Briko系统会开源由语料库训练得到的自有模型，还会提供基于自有模型的常用NLP任务的API接口，使用者可以直接调用API实现翻译、问答、语义分析等任务。Briko的人工智能平台将大幅降低相关应用场景的技术门槛，以促进NLP相关应用的创新与发展。

- **模型搭建与训练**

随着深度学习在NLP领域的深入研究，越来越多准确高效的模型与方法不断涌现（例如Transformers[4,5]，Seq2eq[6,7]等模型，以及BERT，GPT-2等预训练模型）。Briko人工智能平台会纳入最新最优秀的模型供开发者使用，通过模块化、通用化等手段，尽量简化模型搭建、训练的过程，最大程度上为使用者节省时间成本。

Briko会兼容当下主要的深度学习框架（framework），例如TensorFlow，PyTorch，Keras，Caffe等。同时Briko会开源自己的平台，鼓励广大开发者参与到平台的建设中来。

- **Briko开源模型**

训练模型通常需要耗费大量算力，尤其在模型复杂，数据库庞大的NLP应用中，常常需要多块GPU训练几天甚至数周之久。除了为使用者提供便捷的模型搭建框架，Briko系统会训练自有的通用模型，并将这些训练好的模型开源。这些通用模型可以直接拿来使用，帮助使用者节省训练所需的大量时间及运算资源。使用者亦可以通过迁移学习（Transfer Learning）快速高效的训练针对自己应用场景的自有模型。Briko系统的开源模型会随着语料库的不断更新时时保持进化。

- Briko人工智能API

深度学习模型的复杂艰涩无疑对大多数开发者并不友好，而对于一些要求并不苛刻的通用场景，开发者并不需要搭建和训练自己的模型。基于Briko自有的开源模型，Briko系统提供的API可以完全屏蔽模型的部分，使开发者可以更方便的实现一些通用化的功能，例如翻译，问答，情感识别分析等。我们会随着技术的发展以及应用场景的需要，不断推出新的API。

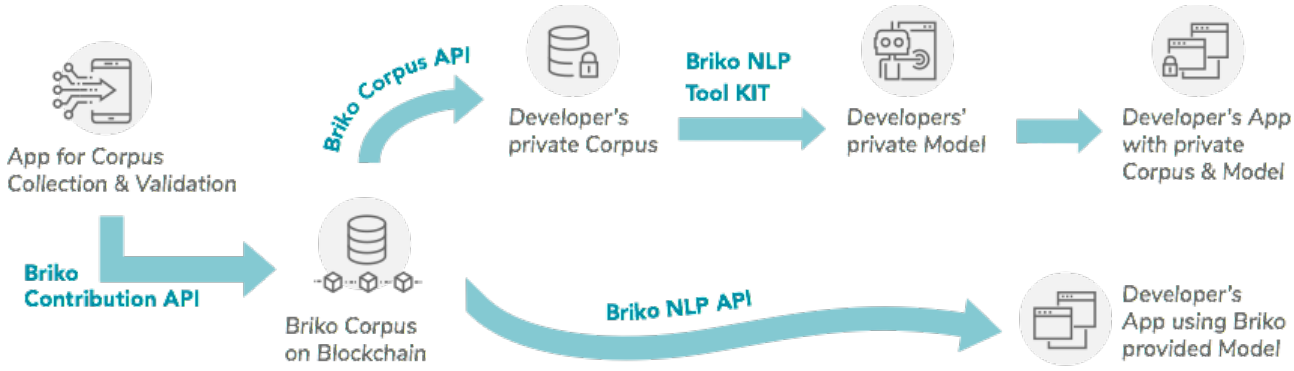
Briko 产品框架

设计理念

Briko框架是一个基于区块链的开放平台，我们希望它能够：

- 打破大公司的数据垄断，低成本地提供高质量的语料库
- 大幅降低NLP应用开发的门槛，缩减开发成本，促进NLP相关应用的繁荣发展

Briko是一个语料贡献者、模型设计者和应用开发者多方共赢的合作社区。通过基于区块链技术的智能合约，透明高效地最大化各个群体的利益，推动整个生态的开放和发展。



Briko 产品框架

Briko 平台的自有应用

- Briko网站

Briko.org 网站是Briko项目的官方网站，会提供Briko语料库、算法和模型的主要管理功能，包括模型的评测，语料库的分类和管理，授权管理和追踪，语料上传，任务分发、处理、提交，访问区块链存储，钱包管理等基本功能。第一步我们会将所有任务放在Briko网站上分发。

- **Briko移动端应用软件**

Briko移动端应用软件可以方便普通用户和Briko语料库交互。Briko将提供覆盖2大主流移动平台的客户端（Android/IOS），Briko移动端应用软件将提供完整的语料上传，审核任务分发、提交，访问区块链存储，钱包管理等基本功能。移动端app会在网站完成之后开发。

第三方Dapp

Briko是一个非盈利组织，我们致力于降低技术的门槛，让更多人使用最前沿的技术，创造更多价值。因此Briko鼓励开发者使用我们提供的模型与接口开发第三方应用，以下几个“抛砖引玉”的Dapp案例是我们举例用来说明基于Briko语料库和模型库产品化的一些可能，希望能启发广大开发者和创业者创造出更多的应用。

- **小说翻译机**

根据内容的类型、题材、领域做针对性优化和模型训练的机器翻译工具。让用外语写成的文本通过机器翻译之后，只通过极少量或无需对成品做人工校准或干预，就能成为可流畅阅读的作品。此产品的C端用户的主要需求集中在娱乐休闲类内容领域，目标是信息获取而非外语习得。应用形式可能为给B端用户的API接口或工具应用或直接开放给C端用户使用的网页服务。

- **应用文档写作助手**

文书类内容的翻译+应用文写作辅助系统。借助丰富精准的语料资源和训练模型，用户可选择语言、文书类型、应用场景，获得符合该语言文化背景的格式建议、文字翻译、语法修正、拼写检查，可切换书面/日常的表达方式，推荐常用句型、高频用词，建议符合礼仪的寒暄等。不光能满足个人和企业跨语言沟通时的常见写作需求，如邮件、报告、规划、通知、论文等，提高沟通效率及有效性，还可以帮助各国游客或移民解决许多日常生活中会遇到的问题，如投诉、询问、致歉、致贺等。

- **语言学习机器人**

满足语言学习的常见需求，通过产品体验优化可面向不同年龄段的用户群。如机器人可以回答：“苹果用法语怎么说？”；“英语怎么说‘可以借用洗手间吗？’”；“翻译下这段话（用户将机器人交给说外语的服务人员）”。 这个应用案例结合了问答、跨语言翻译、语音识别等应用领域，体现了Briko平台更多发展可能性——用于语音识别类模型训练的语料库采集及优化。

Briko 生态

区块链世界的颠覆性之一在于改变了个人的驱动力，改变了组织合作的形式和效率，进而发展出可以重新定义经济与生产活动的力量。

而这些改变是从何而来？

- 通过持有项目代币 (BGE)，参与者在整个生态里既是贡献者，也是价值红利的分享者。
- 整个生态的价值增长，会通过BGE价值的上涨，公平地分配给生态里的所有参与者、贡献者、投资者。
- 这成为生态里各环节的投入参与的驱动力之一。驱动人们自发地合作，为生态贡献力量，创造价值。
- 而区块链的分布式、平等、透明、共识，不光改变人与人之间的信任关系，提高了合作效率，也提供了更多参与方式，赋予个人更多能量和机会。

Briko生态的愿景是构建一个普通人提供数据，研究人员提供模型，开发者提供产品的价值网络，具有自组织的特征，并通过自组织不断进化。成员所创造的价值会在整个生态中共享。Briko生态的协作方式鼓励多样性，鼓励创新，建立系统和有序的共生关系。

价值相关者

- Briko 项目方 ：Briko项目的创建者和建设者。
- 开发者：包括第三方DApp、工具、插件的研发者、协议层的代码贡献者、合约开发者、为Briko网络及生态提供其它各种技术服务、应用研发的技术从业者。
- 资源贡献者：语料的贡献、编辑、审核者，模型提供者等相关劳动者
- 资源使用者：语料、模型、应用和服务的使用者。包括中小企业、个人开发者、学术机构等等。

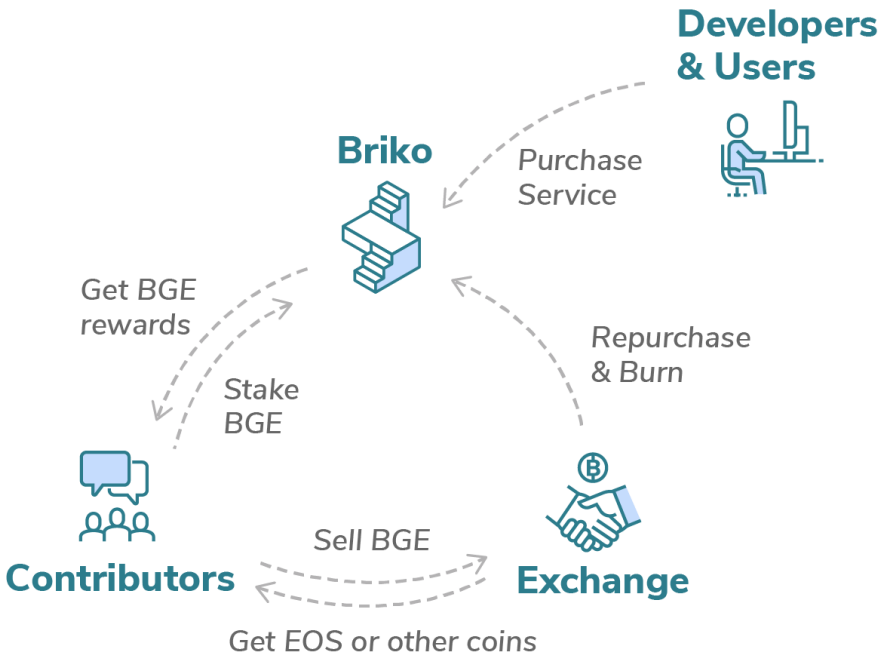
Briko 经济系统

Briko Granted Endorsement (BGE) -- Briko平台流通的虚拟代币

BGE是一种在EOS侧链上发行的代币。它的本质与市面上常见的Bitcoin等通用虚拟货币略有区别：BGE的存在主要并非为了作为一般等价物进行广泛的交换，而是作为Briko平台中衡量对AI模型和语

料库的贡献，例如新模型提供，语料素材提供、语料审核工作，管理维护工作等劳动的一个等价物。

作为整个Briko生态系统中的通用货币，在此生态系统之内，获得BGE需要触发智能合约来完成：用户可以通过提交有效的语料，提交有用的模型，参与审核其他用户提交的语料等方式来获得BGE，作为对用户贡献的认可和奖励。平台的服务，包括语料的使用权将全部通过EOS交易，Briko在得到收入后会通过二级市场公开回购的方式回收用户手中的BGE，完成价值转移。回收的BGE会直接销毁，不再流通。抵押机制是指在贡献者领取语料任务时，需要首先抵押一笔BGE，任务顺利完成后返还。这旨在通过经济手段保证语料和模型库的质量，减少恶意提交降低整个社区审核压力和工作量。若Briko系统最终判断用户恶意提交任务，则之前抵押的BGE将被Briko系统收回，销毁或作为鼓励之后用户贡献的基金。



用户贡献奖励机制 -- 如何激励用户做出贡献？

- 鼓励社群用户参与，奖励方式和内容多样化

用户在成功提交有效的语料、模型和审核意见后，会得到以BGE为主要形式的奖励。此外，Briko平台会定期设立多样化的目标，以鼓励社群成员贡献和协作：我们会引入经验等级和能力评价体系，奖励工作量大，工作质量优秀的贡献者。同时Briko社区也会根据需求的变化和发展，持续进化和迭代的Briko的激励机制。

随着NLP应用的发展，对语料和模型的需求也会不断变化，Briko平台的生态和经济系统也需要保持灵活性和成长性以适应这样发展。Briko会同社群一起共同演进具体的激励机制，去引导用户的贡献行为，使之产出能，帮助Briko上的语料库和模型库在最新和最有需要的方向上持续得到完善。

Briko平台的使用者亦可提出悬赏，提高针对此类贡献的奖励，以征集其所需要的特定用途和格式语料或模型，通过经济杠杆有效解决语料和模型库缺失的难题。

- **应用驱动的奖励模型——贡献越多，收益越大**

Briko平台希望建立可持续的经济系统还体现在它对做出特殊贡献的用户给予持续的奖励。如果用户贡献的模型或语料被大量应用，为社区创造了大量价值，根据智能合约的记录历史，Briko平台会奖励这些资源的杰出贡献者，让他们获得额外的收益。Briko平台所采用的应用驱动的奖励模型，结合区块链技术的优势，将改变用户贡献价值衡量的难题，通过经济手段鼓励用户贡献最优质的语料和模型。

用户贡献审核机制 -- 如何保障用户贡献的质量？

用户贡献的语料和模型需要经过审核通过。审核者是社群中的一个角色，它的存在旨在保障用户提交语料的质量，防止语料库受到恶意提交的攻击。审核者可由语料贡献者兼任，他们是社群的重要组成部分，在Briko平台上的有效工作在会得到相应的经济回报。

Briko平台也提供了对审核者的审核质量进行把关的机制。和提交语料时一样，审核者在提交审核意见之前也会自动抵押一定BGE，待审核意见得到确认，系统会将先前抵押的BGE返还用户。同理，未通过审核的贡献会受到惩罚，以保证审核质量。

为了防止提交者和审核者的共谋攻击，Briko平台将会随机对审核者进行测试，在大量审核任务之中加入随机的有已知有效答案的审核问题，识别恶意提交语料和审核的用户加倍惩罚。

用户信誉值系统 -- 如何鼓励用户长期参与？

Briko的经济系统鼓励用户长期参与，并对信任的贡献突出的用户给予长期支持。用户信誉值系统即是根据这个目的而设立的。持续贡献高质量语料和模型，提供高质量审核判断的用户将获得高信誉值。而信誉值高的用户获得更高的以BGE为主要形式的收益，并且可以参与社区规则制定和管理工作。

用户信誉系统旨在提供时间维度上的用户奖励依据，鼓励社群建设和高质量用户的长期参与。同时，对加入社区的新用户而言，用户信誉系统可以起到鼓励新用户为积累用户信誉值而学习社区规则和运作方式，贡献高质量资源的作用。

核心团队

崔淼

长期从事软件开发工作，10余年移动软件开发经验。北京邮电大学硕士。曾任职多家公司里项目主管，高级经理。熟练使用各种语言，开发工具。具有各类大型项目开发，管理经验。具有独立开发软件的经验和能力。

黄威

多年系统软件开发及虚拟货币系统管理经验。2008起研究点对点网络系统和基于点对点网络的虚拟货币系统。早期Bitcoin网络的参与者。前香港大学和多伦多大学研究员。

赵舒泽

多伦多大学电子工程博士。研究领域包括数据中心、AI芯片的高效节能应用以及基于深度学习的自然语言处理。

冯宇晖

多伦多大学机电工程硕士。研究领域包括智能机器人人机交互功能的开发与应用以及智能机械在工业，医疗等多领域内的集成与应用。多年应用软件开发经验。

霍炬

科技和互联网领域多年开发、架构和团队管理经验。经历包括创立国内最早的企业搜索云计算供应商Ginkgotek；盛大创新院高级研究员；以独立顾问身份为纽约时报中文网、FTchinese 等国际化媒体集团提供技术咨询等。同时，他也是中文著名的科技领域写作者，其早期的blog和目前的微信订阅号《歪理邪说》都有巨大影响力。

Reference

[1] Mikolov, Tomas et al. “Efficient Estimation of Word Representations in Vector Space.” CoRR abs/1301.3781 (2013).

[2] Mikolov, Tomas et al. “Distributed Representations of Words and Phrases and their Compositionality.” NIPS (2013).

[3] Pennington, Jeffrey et al. “Glove: Global Vectors for Word Representation.” EMNLP (2014).

[4] Vaswani, Ashish et al. “Attention Is All You Need.” NIPS (2017).

[5] Dehghani, Mostafa et al. “Universal Transformers.” CoRR abs/1807.03819 (2018).

[6] Sutskever, Ilya et al. “Sequence to Sequence Learning with Neural Networks.” NIPS (2014).

[7] Cho, Kyunghyun et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.” EMNLP (2014).

[8] Joulin, Armand, et al. "Bag of tricks for efficient text classification." arXiv preprint arXiv:1607.01759(2016).

[9] Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." ACL (2018).

[10] He, Di, et al. "Dual learning for machine translation." Advances in Neural Information Processing Systems. 2016.

[11]野口悠紀雄 .(2016). 区块链革命：分布式自律型社会出现. 东方出版社, 2018.