拓尔思副董事长兼总经理施水才：

为“漂亮的皮囊”注入有趣的NLP灵魂

【董事长面对面】

漂亮的皮囊千篇一律，有趣的灵魂万里挑一。我们的任务就是要做有趣的灵魂，做美丽皮囊背后的大脑。

本报记者 赵学毅 李乔宇

ChatGPT站上风口，人工智能迈向新纪元。在名单越写越长的ChatGPT概念股中，拓尔思信息技术股份有限公司（以下简称“拓尔思”）是少有的在早期就从事相关业务，并有项目落地的企业之一。在这场风起云涌的技术追逐战中，拓尔思站到了风口之前，且其股价在月初创下新高。

面对我国类ChatGPT产品与ChatGPT的差距，国内企业该如何在算力、算法和数据层面突围、赶超？国内最早从事自然语言处理（NLP）技术研发的拓尔思正在做什么？

迎着春日和煦的暖阳，《证券日报》记者来到位于北京市海淀区中关村西三旗金隅科技园，走进这家既年轻又“老练”的公司，对拓尔思副董事长兼总经理施水才进行专访。

今年以来，ChatGPT成为资本市场上最受关注的话题之一，引资金竞相追逐。拓尔思所擅长的NLP技术，则是ChatGPT背后的主要技术基础。作为中文全文检索技术的始创者，领先的大数据、人工智能和数据安全产品及服务提供商，拓尔思有着得天独厚的创新优势，同时也肩负着快速缩小与ChatGPT差距的重任。

紧跟时代浪潮 “209工程”火速启动

在拓尔思的办公楼内，记者看到多个会议室均以知名科学家命名，爱因斯坦、玛丽·居里、巴贝奇、伽利略、达尔文、达·芬奇……

“所有研发人员的会议室都以科学家的名字命名，作为踩在巨人肩膀上的他们，希望通过这种方式表示敬意。”施水才解释说。此时，这些会议室大多处于忙碌状态。

听着会议室里不时传出热烈的讨论声，记者仿佛穿行在时空隧道里，那是一代一代科学探索者们的真切对话。

每个会议室的门上都贴着不同项目的研发时间表，覆盖来自拓尔思的中国版ChatGPT产品在内的“209工程”赫然在列。显然，ChatGPT正在激起我国技术创新的千层浪，包括拓尔思在内的科技企业正在积极探索类ChatGPT产品赛道。

早在2015年，拓尔思就参与了一项“高考机器人”的项目，通过对十亿量级数据库的搭建，帮助该项目实现对于试题答案进行分析和科学评价。除了机器人自动撰稿，在北京冬奥会和卡塔尔世界杯期间，拓尔思的NLP技术也承担了虚拟人播报脚本生成的重任；在直播电商领域，拓尔思正在为新农人电商提供直播文案辅助生成技术……

“从分词、句法、语义等信息的信息抽取、自动聚类、自动分类，到自然语言的理解和生成，我们做了全方位的研究，但这仍然不够。”谈及ChatGPT，施水才表示，“我们虽然有一些相关的技术和应用案例，但对比ChatGPT，明显感觉到我们的产品还有不小差距，这种差距既有算力上的，也有算法和数据层面的。”因其海量数据以及大模型工程化的特征，施水才将ChatGPT的成功形容为“大力出奇迹”。“接下来，拓尔思面临的挑战就是如何缩小这些差距，做出与国际水平比肩的产品出来。”

为了紧跟时代变化的浪潮，同时也为了使自己的技术不落后于人，拓尔思开启了“209工程”。

“209工程”的名字来自项目正式启动的日期（2023年2月9日），其计划用3个月至6个月的时间，基于通用AIGC大模型，以“专业大模型+领域知识数据”为核心，以NLP技术突破来推动更多AIGC商业落地。“顺利的话，年内大家就能够看到来自拓尔思的专注于垂直行业的类ChatGPT产品。”施水才表示，为了支持新技术的研发，拓尔思单独成立了数字经济研究院，设立了多个新部门，全力推动垂直行业的专业大模型的研发。

施水才透露，未来要进一步支持“209工程”，拓尔思有意加大研发投入，尤其会加大在AI技术领域的研发投入，研发费用占比会提升至20%左右。据拓尔思披露的财报数据显示，2022年前三季度，公司研发费用为9592.28万元，占总营业收入15.06%。

锚定B端场景 “大力出奇迹”有望复制

志存高远，拓尔思要做的“专业大模型+领域知识数据”，希望能在ChatGPT的基础上更进一步。

ChatGPT看上去效果惊人，但在准确度上有待商榷。施水才告诉记者，这是因为目前GPT大模型本质作为概率模型以及提示训练的机制所导致的。在很多ToC场景中，人们能够接受这种误差。但在ToB场景中，信息的提取和检索要求准确全面，任何分析和预测都需要有理有据和正确的分析框架。

AI ToB的关键在于领域知识的建立和领域模型的再训练。在部分ToB场景中，可用的数据集是有限的，需要更多工程化和特定方法的干预；部分ToB场景中，用户对安全性、一致性、规范性、意识形态敏感性的要求更高，需要更多的定制和额外的相关工作等等。而这些都是拓尔思不断在思考、研究、开发和应用实践去解决的问题。

帮助ChatGPT“大力出奇迹”的海量数据和大模型技术未必遥远。在NLP技术和算法层面，拓尔思有着长期的技术投入和不断推陈出新的产品；在数据层面，作为A股第一家上市的大数据技术企业，拓尔思从2010年起就投资建设自有的大数据中心，常年持续采集的海量网络数据资源；拓尔思还拥有海量大数据的规模化治理加工能力优势，包括NLP自动化技术平台、组织流程、质量控制等，具有支撑AI技术开发和赋能的完备的数据能力。

未来，拓尔思预计新业务会成为其增量收入的主要来源。施水才告诉记者，拓尔思打造的垂直行业的专业版ChatGPT产品有望能够满足B端用户对于内容生成的需求，亦有望打开其背后SaaS产业的市场空间。“SaaS服务平台所带来的收入有望在三年内收入占比超过50%，这意味着这项业务还有5亿元至10亿元的增量空间。”

赋能数字化转型 勇做有趣的灵魂

类ChatGPT产品可能会带来生产工具的革新，有望加速各行业的数字化转型。比如在电商客服领域，人工智能可以利用多人对话的能力达到提质增效的效果；同时能够通过交流和沟通提供情感呵护的陪伴机器人也有望实现。虚实结合的应用场景亦有望在类ChatGPT产品的推动下加速落地。

无论是元宇宙还是虚拟数字人，都曾经获得市场关注。在类ChatGPT产品落地的过程中，拓尔思等NLP领域的从业者的任务就是“为漂亮的皮囊注入有趣的灵魂”。施水才表示：“漂亮的皮囊千篇一律，有趣的灵魂万里挑一。我们的任务就是要做有趣的灵魂，做美丽皮囊背后的大脑。”

作为最早涉足NLP技术的从业者之一，新技术引发了施水才对行业格局的思考。ChatGPT的出现意味着不限制场景的通用人工智能发展到了一个新阶段，可能引发互联网产业的变革。在这个过程中，自然语言处理技术也取得了颠覆性的突破，为拓尔思带来了新启示。

NLP技术在人工智能领域具有较高的地位，有人将之形容为“人工智能皇冠上的一颗明珠”。在深度学习算法被发明后，图像的识别已经能够实现较高的准确率，但自然语言处理的难度远高于图像处理。

新技术的出现拉近了人们与NLP之间的距离，同时也对传统NLP业务造成冲击。因此，在AIGC领域的赶超成为当下我国科技企业的重要任务。拓尔思责无旁贷，成为这场赶超赛的领队人。

在结束了近两个小时的专访后，施水才面对镜头微笑着表示：我们都赶上了大有可为的好时代，应该怀揣梦想，拥抱变化，构筑新愿景，并矢志不渝，追光而行，为中国式现代化做出新的贡献。

（证券日报官网及两微一端已同步推出视频报道《拓尔思施水才：ChatGPT兴起预示着自然语言处理时代到来》，敬请关注）

拓尔思：自然语言处理的春天来了

本报记者 李乔宇

数据、算法、算力是保证AIGC产出质量的三大核心要素，构建起个性化、专业性的内容自动生成壁垒将成为人工智能厂商比肩前沿技术及保持领先地位的良方。

作为国内最早从事NLP技术研发的企业之一，拓尔思致力于在商业落地过程中将技术、产品、场景完美融合，打造类人助手全新模式的上市公司。在2023年开年ChatGPT春风到来之前，拓尔思已将NLP技术相关产品服务真正运用到商业实践中并产生效益。

通用大模型 开启通用人工智能一扇门

NLP被誉为“人工智能皇冠上的一颗明珠”，是人工智能认知能力的核心，对于人工智能发展至关重要。NLP主要做三件事：让机器听懂人话（理解）、讲人话（生成）、干人事（应用）。在深度学习算法被发明后，图像和语音识别已经能够实现超过人类的准确率。但由于语言规律是错综复杂的，不仅要有一定的知识依赖，而且语言的使用还基于环境和上下文，因此自然语言处理的难度远高于图像处理。

ChatGPT的火爆在于它具备非常强的语言理解能力和生成能力，使机器能够理解人类的意图，实现类人对话。2023年3月15日，OpenAI发布的ChatGPT4.0在识图能力、文字输入限制、回答准确等方面较前款产品有了飞跃式提升，并且其在各种专业测试和学术基准上表现优良。

拓尔思副董事长兼总经理施水才曾表示：“ChatGPT是人工智能生成内容（以下简称AIGC）技术进展的里程碑，该模型使利用人工智能进行内容创作的技术成熟度大幅提升，有望成为新的全行业生产力工具，提升内容生产效率与丰富度。”施水才评价ChatGPT是NLP技术革新的一大步，并开启了通用人工智能的一扇大门。

通用大模型 在各行业垂直应用落地

目前，ChatGPT正在引领一场深刻的变革，其技术可以用于智能客服、智能对话、智能搜索、智能推荐、机器写稿、辅助办公、虚拟员工等应用场景，在各行各业的商业落地中加速成熟。自2000年起，拓尔思就自主研发NLP技术，见证了人工智能技术发展的整个过程。大模型要落地，离不开数据、算法、算力、应用等综合能力的聚合，拓尔思在这方面拥有深厚的积累。

数据层面，拓尔思从2010年自建数据中心以来，已采集了超过10年的互联网公开数据，拥有规模及质量均位列业界前茅的另类数据资产。目前，拓尔思拥有来自境内外、各行各业的公开数据规模超1300亿条，数据类型涵盖新闻、资讯、政策、视频、图片、百科、社交等多模态，数据总量达100TB以上。

算法层面，拓尔思长年深耕政务、媒体、金融、专利等行业的信息化建设，已积累了30+专业知识库，涵盖通用语义分析、人物/机构、行业分类、专利、媒体、金融、科技情报、乡村振兴等领域；30000+标签模型，覆盖媒体、舆情、金融风控、产业投研、智能消保、开源情报、政务应用等场景；350+深度学习算法模型，包括NLP、金融监管、风控征信、公共安全、产业服务、传播分析、事件研判、舆情态势等通用模型、指数模型、领域模型等。

应用层面，拓尔思具备数据标注、模型设计、训练、优化、评估、部署等一站式AI工程化落地服务能力，同时在政务、媒体、金融、舆情、安全、专利等行业有着丰富的应用场景实践，有助于专业大模型贴合用户场景进行快速落地，产生业务价值。

NLP的春天已至，拓尔思正在积极拥抱大模型的发展，并充分挖掘多年积累的行业知识和专业模型优势，持续保持在AI领域的技术竞争力。在大有可为的好时代，拓尔思将继续坚定理想，拥抱变化，构筑新愿景，满怀新期待，力争为国家数字经济的创新发展贡献力量。

【记者手记】

探寻人工智能皇冠明珠上的光辉

李乔宇

ChatGPT的浪潮将其背后的NLP技术推至风口。作为国内最早从事NLP技术研发的企业之一，拓尔思股价仅在今年一季度就实现了170.05%的上涨。

满怀新奇，笔者时隔两年，走进拓尔思的新办公地，与上次不同的是，位于北京市海淀区的新办公楼更加宽敞明亮，原本在北京分隔三地的拓尔思办公地，终于汇聚到了一起。

站在公司前台大厅的展示屏前，笔者看到一帧帧繁忙景象的图片：数以千计台装载着拓尔思大数据智能系统的服务器24小时不间断地运转着，上亿条信息被归集、加工、存储和分析……展示屏画面流转，拓尔思副董事长兼总经理施水才两年前曾向笔者描绘的业务版图就这样生动地展示出来。

早在2000年，我国第一批NLP从业者就已经开始涌现，拓尔思正是国内NLP领域的先行者。在这段发展史中，我国NLP技术一直保持着技术和应用层面的突破和创新。在近十年的时间里，我国NLP技术一度实现了与国际先进水平并肩。

但成绩已成过往，新技术引发新变革，以施水才为代表的一代科研人员又开始了新思考。伴随着“209工程”火速启动，“拓尔思版”ChatGPT提上日程。

ChatGPT的热度让大家开始关注谁会成为中国的OpenAI。但或许我们并不需要中国的OpenAI，我们要做的是中国的拓尔思，是世界领先的国产AIGC和国产NLP。

施水才对NLP的描述颇有几分夸父逐日的意味，他说NLP之所以被称为人工智能皇冠上的明珠，是由于其只能不断靠近，却终究难以摘取。皇冠上的明珠难摘取，太阳难靠近。但谁说夸父的子孙没有在一代一代更接近太阳呢？


施水才接受本报记者采访 郑凯鹏/摄


拓尔思员工在进行产品研发 郑凯鹏/摄

本版主编 沈明 责编 石柳 美编 王琳 制作 李波 E-mail:zmzx@zqrb.net 电话 010-83251785