(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2026 年 9 月 3 日 (03.09.2026)

WIPO | PCT

(10) 国际公布号
WO 2026/178988 A1

(51) 国际专利分类号:
G06F 40/30 (2020.01)

(21) 国际申请号: PCT/CN2025/095833

(22) 国际申请日: 2025 年 5 月 19 日 (19.05.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202510210103.9 2025年2月25日 (25.02.2025) CN

(71) 申请人: 无限星途科技私人有限公司 (X STAR TECHNOLOGY PTE. LTD.) [SG/SG]; 新加坡新达城1号楼, 37-03, 淡马锡林荫道7号 038987 (SG)。

(72) 发明人: 杜新峰 (DU, Xinfeng); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。 贾货贺 (JIA, Zihe); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。 刘兵 (LIU, Bing); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。 张磊 (ZHANG, Lei); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。 吕舒瑶 (LV, Shuyao); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。 刘涛 (LIU, Tao); 中国上海市浦东新区中国（上海）自由贸易试验区浦东大道2123号三层 200135 (CN)。

(74) 代理人: 北京超凡宏宇知识产权代理有限公司 (CHOFN INTELLECTUAL PROPERTY); 中国

(54) **Title:** TEXT SEGMENTATION METHOD AND APPARATUS, ELECTRONIC DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) **发明名称:** 文本切分方法、装置、电子设备及计算机可读存储介质

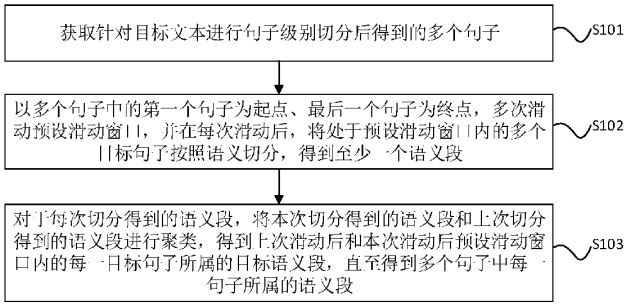

图 2

S101 Acquire a plurality of sentences obtained by performing sentence-level segmentation on target text

S102 By using the first sentence among the plurality of sentences as a starting point and the last sentence as an end point, slide a preset sliding window multiple times, and upon each sliding, segment a plurality of target sentences within the preset sliding window on the basis of semantics to obtain at least one semantic segment

S103 For the semantic segment obtained by each segmentation, cluster the semantic segment obtained by the current segmentation and the semantic segment obtained by the previous segmentation to obtain, for each target sentence within the preset sliding window after the previous sliding and the current sliding, a target semantic segment to which the target sentence belongs, until, for each sentence among the plurality of sentences, a corresponding semantic segment is obtained

(57) **Abstract:** The present disclosure relates to computer technology, and provides a text segmentation method and apparatus, an electronic device, and a computer-readable storage medium. The method comprises: acquiring a plurality of sentences obtained by performing sentence-level segmentation on target text; by using the first sentence among the plurality of sentences as a starting point and the last sentence as an end point, sliding a preset sliding window multiple times, and upon each sliding, segmenting a plurality of target sentences within the preset sliding window on the basis of semantics to obtain at least one semantic segment; and for the semantic

WO 2026/178988 A1

[见续页]

北京市海淀区北四环西路68号左岸工社1215-1218室 100080 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UY, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

segment obtained by each segmentation, clustering the semantic segment obtained by the current segmentation and the semantic segment obtained by the previous segmentation to obtain, for each target sentence within the preset sliding window after the previous sliding and the current sliding, a target semantic segment to which the target sentence belongs, until, for each sentence among the plurality of sentences, a corresponding semantic segment is obtained. The present disclosure can ensure semantic coherence and improve the accuracy and reliability of question answering in RAG systems.

(57) 摘要: 本公开涉及计算机技术，提供一种文本切分方法、装置、电子设备及计算机可读存储介质，包括：获取针对目标文本进行句子级别切分后得到的多个句子；以多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，直至得到多个句子中每一句子所属的语义段。本公开能够保证语义连贯，提升RAG系统问答的准确性和可靠性。