

(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
2022 年 7 月 14 日 (14.07.2022)



(10) 国际公布号
WO 2022/148467 A1

(51) 国际专利分类号:
G06F 40/289 (2020.01)

(21) 国际申请号: PCT/CN2022/071144

(22) 国际申请日: 2022 年 1 月 10 日 (10.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110034450.2 2021 年 1 月 11 日 (11.01.2021) CN

(71) 申请人: 百果园技术 (新加坡) 有限公司
(BIGO TECHNOLOGY PTE. LTD.) [SG/SG]; 新加坡
巴西班让路枫树商业城 30 号楼 15 层
31A, Singapore 117440 (SG)。

(72) 发明人: 及

(71) 申请人 (仅对 US): 张建宁 (ZHANG, Jianning) [CN/
SG]; 新加坡巴西班让路枫树商业城 30 号楼
15 层 31A, Singapore 117440 (SG)。

(74) 代理人: 北京泽方誉航专利代理事务所 (普通
合伙) (BEIJING ZFANG PATENT AGENCY); 中
国北京市丰台区南四环西路 186 号四区 2 号楼
7 层 23 室陈照辉, Beijing 100071 (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,
CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB,
GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT,
JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,
LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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,
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, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, 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,

(54) Title: CROSS-LANGUAGE DATA ENHANCEMENT-BASED WORD SEGMENTATION METHOD AND APPARATUS

(54) 发明名称: 一种基于跨语言数据增强的分词方法及装置

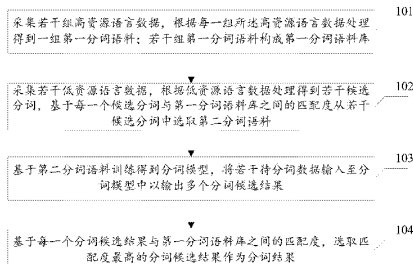


图 1

- 101 Acquire multiple groups of high-resource language data, and process each group of the high-resource language data to obtain a group of first word segment materials, multiple groups of first word segment materials constituting a first word segment corpus
- 102 Acquire several low-resource language data and process the low-resource language data to obtain several candidate word segments, and select second word segment materials from among the several candidate word segments on the basis of the degree of matching between each candidate word segment and the first word segment corpus
- 103 Train a word segmentation model on the basis of the second word segment materials and input, into the word segmentation model, several data to be subjected to word segmentation to output multiple candidate word segmentation results
- 104 Select, on the basis of the degree of matching between each candidate word segmentation result and the first word segment corpus, a candidate word segmentation result having the highest degree of matching as a word segmentation result

(57) Abstract: Embodiments of the present application provide a cross-language data enhancement-based word segmentation method and apparatus. The technical solution provided in the embodiments of the present application is: acquiring high-resource language data and processing same to obtain word segment materials, acquiring low-resource language data to obtain candidate word segments, and screening the candidate word segments according to the word segment materials obtained from the high-resource language data to select candidate word segments having a high degree of matching with the word segment materials as word segment materials of the low-resource language data, training a word segmentation model on the basis of the low-resource word segment materials to automatically output candidate word segmentation results for the low-resource language data on the basis of the model, and selecting a word segmentation result according to the degree of matching between the candidate word segmentation results and the word segment materials of the high-resource language data. The use of high-resource language materials to automatically augment and validate model training data in low-resource languages solves the problem of imbalance between data resources and annotation resources in different languages, thereby providing an easier and more efficient solution for an iterative word segmentation model.

[见续页]

IE, IS, IT, LT, LU, LV, MC, 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)。

根据细则4.17的声明：

- 关于申请人有权要求在先申请的优先权(细则4.17(iii))

本国际公布：

- 包括国际检索报告(条约第21条(3))。
-

(57) 摘要：本申请实施例提供了一种基于跨语言数据增强的分词方法及装置。本申请实施例提供的技术方案通过采集高资源语言数据处理得到分词语料，采集低资源语言数据获取候选分词，并根据从高资源语言数据处获得的分词语料对候选分词进行甄选，选择与分词语料匹配度高的作为低资源语言数据的分词语料，并根据低资源分词语料进行分词模型的训练，可以实现对低资源语言数据基于模型自动输出分词候选结果，结合分词候选结果与高资源语言数据的分词语料的匹配度进行选取分词结果，通过使用高资源语言的语料对低资源语言的模型训练数据进行自动扩充和验证，解决了数据资源和标注资源在不同语言之间不平衡的问题，为迭代分词模型提供了一种更轻量而高效的解决方式。