

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
29 September 2005 (29.09.2005)

PCT

(10) International Publication Number
WO 2005/091167 A2

(51) International Patent Classification⁷: **G06F 17/28**

(21) International Application Number:
PCT/US2005/008863

(22) International Filing Date: 16 March 2005 (16.03.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/802,479 16 March 2004 (16.03.2004) US

(71) Applicant (for all designated States except US):
GOOGLE INC. [US/US]; 1600 Amphitheatre Park-
way, Mountain View, CA 94043 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **WU, Jun** [CN/US];
20398 Franklin Ave, Saratoga, CA 95070 (US). **ZHU,**
Huican [CN/US]; 1560 San Thomas Aquino Road #32,
San Jose, CA 95130 (US). **ZHU, Hongjun** [CN/US]; 557
Carrick Court, Sunnyvale, CA 94087 (US).

(74) Agent: **KUO, Jung-hua**; P.O. Box 3275, Los Altos, CA
94024 (US).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN,
CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI,
GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE,
KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD,
MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG,
PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ,
TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA,
ZM, ZW.

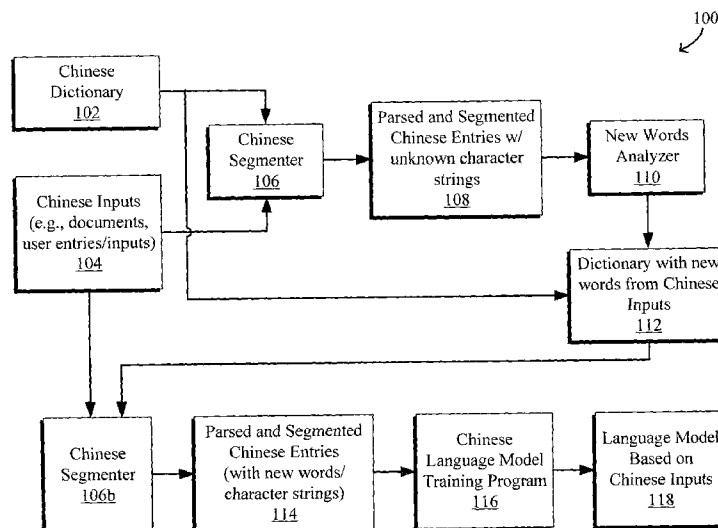
(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,
ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO,
SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN,
GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— without international search report and to be republished
upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.

(54) Title: SYSTEMS AND METHODS FOR TRANSLATING CHINESE PINYIN TO CHINESE CHARACTERS



(57) Abstract: Systems and methods to process and translate pinyin to Chinese characters and words are disclosed. A Chinese language model is trained by extracting unknown character strings from Chinese inputs, e.g., documents and/or user inputs/queries, determining valid words from the unknown character strings, and generating a transition matrix based on the Chinese inputs for predicting a word string given the context. A method for translating a pinyin input generally includes generating a set of Chinese character strings from the pinyin input using a Chinese dictionary including words derived from the Chinese inputs and a language model trained based on the Chinese inputs, each character string having a weight indicating the likelihood that the character string corresponds to the pinyin input. Ambiguous user input may be classified as non-pinyin or pinyin by identifying an ambiguous pinyin/non-pinyin ASCII word in the user input and analyzing the context to classify the user input.

WO 2005/091167 A2