PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/20	A1	(11) International Publication Number: WO 00/62193
		(43) International Publication Date: 19 October 2000 (19.10.00)

(21) International Application Number: PCT/SG99/00029

(22) International Filing Date: 8 April 1999 (08.04.99)

(71) Applicant *(for all designated States except US)*: KENT RIDGE DIGITAL LABS [SG/SG]; 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

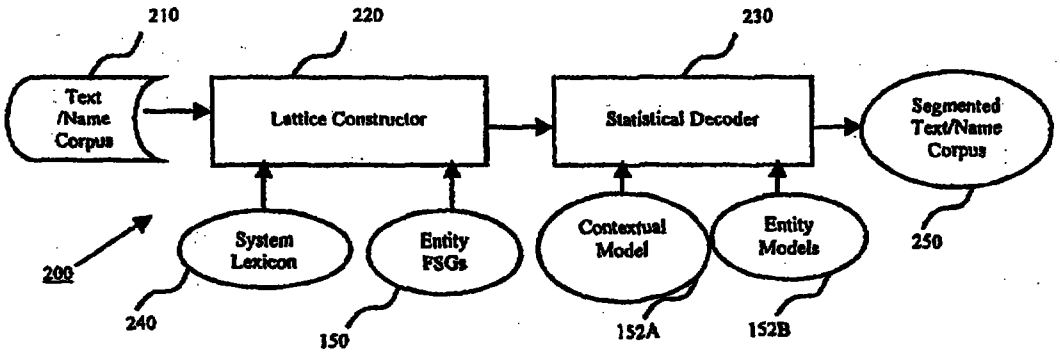
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: BAI, Shuanhu [CN/SG]; Blk 403, Clementi Avenue 1 #03–192, Singapore 120403 (SG). WU, Horng, Jyh, Paul [CN/SG]; Blk 122, Jurong East Street 13, #04–35, Singapore 600122 (SG). LI, Haizhou [CN/SG]; Blk 413, Pandan Gardens #11–132, Singapore 600413 (SG). LOUDON, Gareth [GB/SG]; 61 Jalan Puteh Jerneh, Singapore 278077 (SG).

(74) Agent: SPRUSON & FERGUSON PTE LTD.; 51 Bras Basah Road, #02–03 Plaza By The Park, Singapore 189554 (SG).

(81) Designated States: CN, JP, SG, US.

Published
With international search report.

(54) Title: SYSTEM FOR CHINESE TOKENIZATION AND NAMED ENTITY RECOGNITION



(57) Abstract

A system (100, 200) for tokenization and named entity recognition of ideographic language is disclosed. In the system, a word lattice is generated for a string of ideographic characters using finite state grammars (150) and a system lexicon (240). Segmented text is generated by determining word boundaries in the string of ideographic characters using the word lattice dependent upon a contextual language model (152A) and one or more entity language models (152B). One or more named entities is recognized in the string of ideographic characters using the word lattice dependent upon the contextual language model (152A) and the one or more entity language models (152B). The contextual language model (152A) and the one or more entity language models (152B) are each class–based language models. The lexicon (240) includes single ideographic characters, words , and predetermined features of the characters and words.