* C L M S *

CLAIMS:

1. A method of tokenization and named entity recognition of ideographic language, said method including the steps of:

generating a word lattice for a string of ideographic characters using finite state grammars and a system lexicon,

said finite state grammars are a dynamic and complementary extension of said lexicon for creating named entity hypotheses and said lexicon includes single ideographic characters, words, and predetermined features of said characters and words;

generating segmented text by determining word boundaries in said string of ideographic characters using said word lattice dependent upon a contextual language model and one or more entity language models; and

recognizing one or more named entities in said string of ideographic characters using said word lattice dependent upon said contextual language model and said one or more entity language models.

2. The method according to claim 1, further including the step of combining said contextual language model and said one or more entity language models.

3. The method according to claim 1, wherein said contextual language model and said one or more entity language models are each class-based language models.

4. The method according to claim 1, wherein said contextual language model and said one or more entity language models incorporate local and contextual linguistic information, respectively, for producing prioritized word and corresponding category sequences.

REPLACEMENT SHEET
19 OCT 2001

* C L M S *

5. The method according to claim 4, wherein said contextual language model and said one or more entity language models are dependent upon an n-gram paradigm.

6. The method according to claim 1, wherein said lattice generating step includes the step of generating one or more elements of said lattice using said lexicon.

7. The method according to claim 1, wherein said finite state grammars run on said predetermined features contained in said lexicon to suggest possible entities, entity boundaries and entity categories.

8. An apparatus for tokenization and named entity recognition of ideographic language, said apparatus including:

means for generating a word lattice for a string of ideographic characters using finite state grammars and a system lexicon,

said finite state grammars are a dynamic and complementary extension of said lexicon for creating named entity hypotheses and said lexicon includes single ideographic characters, words, and predetermined features of said characters and words;

means for generating segmented text by determining word boundaries in said string of ideographic characters using said word lattice dependent upon a contextual language model and one or more entity language models; and

means for recognizing one or more named entities in said string of ideographic characters using said word lattice dependent upon said contextual language model and said one or more entity language models.

9. The apparatus according to claim 8, further including means for combining said contextual language model and said one or more entity language models.

REPLACEMENT SHEET
19 OCT 2001

10. The apparatus according to claim 8, wherein said contextual language model and said one or more entity language models are each class-based language models.

11. The apparatus according claim 8, wherein said contextual language model and said one or more entity language models incorporate local and contextual linguistic information, respectively, for producing prioritized word and corresponding category sequences.

12. The apparatus according to claim 11, wherein said contextual language model and said one or more entity language models are dependent upon an n-gram paradigm.

13. The apparatus according to claim 8, wherein said lattice-generating means includes means for generating one or more elements of said lattice using said lexicon.

14. The apparatus according to claim 8, wherein said finite state grammars run on said predetermined features contained in said lexicon to suggest possible entities, entity boundaries and entity categories.

15. A computer program product having a computer readable medium having a computer program recorded therein for tokenization and named entity recognition of ideographic language, said computer program product including:

computer program means for generating a word lattice for a string of ideographic characters using finite state grammars and a system lexicon,

said finite state grammars are a dynamic and complementary extension of said lexicon for creating named entity hypotheses and said lexicon includes single ideographic characters, words, and predetermined features of said characters and words;

REPLACEMENT SHEET
19 OCT 2001

* C L M S *

computer program means for generating segmented text by determining word boundaries in said string of ideographic characters using said word lattice dependent upon a contextual language model and one or more entity language models; and

computer program means for recognizing one or more named entities in said string of ideographic characters using said word lattice dependent upon said contextual language model and said one or more entity language models.

16. The computer program product according to claim 15, further including computer program means for combining said contextual language model and said one or more entity language models.

17. The computer program product according to claim 15, wherein said contextual language model and said one or more entity language models are each class-based language models.

18. The computer program product according to claim 15, wherein said contextual language model and said one or more entity language models incorporate local and contextual linguistic information, respectively, for producing prioritized word and corresponding category sequences.

19. The computer program product according to claim 18, wherein said contextual language model and said one or more entity language models are dependent upon an n-gram paradigm.

20. The computer program product according to claim 15, wherein said computer program means for lattice-generating includes computer program code means for generating one or more elements of said lattice using said lexicon.

21. The computer program product according to claim 15, wherein said finite state grammars run on said predetermined features contained in said lexicon to suggest possible entities, entity boundaries and entity categories

REPLACEMENT SHEET
19 OCT 2001