PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [7] : **G10L 13/00**	**A2**	**(11) International Publication Number:** **WO 00/30070** **(43) International Publication Date:** 25 May 2000 (25.05.00)

(21) International Application Number: PCT/US99/27210

(22) International Filing Date: 17 November 1999 (17.11.99)

(30) Priority Data:
60/108,778 17 November 1998 (17.11.98) US

(71) Applicant: LERNOUT & HAUSPIE SPEECH PRODUCTS N.V. [BE/BE]; Flanders Language Valley 50, B-8900 Ieper (BE).

(71)(72) Applicant and Inventor: CARUS, Alwin, B. [US/US]; 20 East Quinobequin Road, Burlington, MA 02468 (US).

(74) Agents: SUNSTEIN, Bruce, D. et al.; Bromberg & Sunstein LLP, 125 Summer Street, Boston, MA 02110-1618 (US).

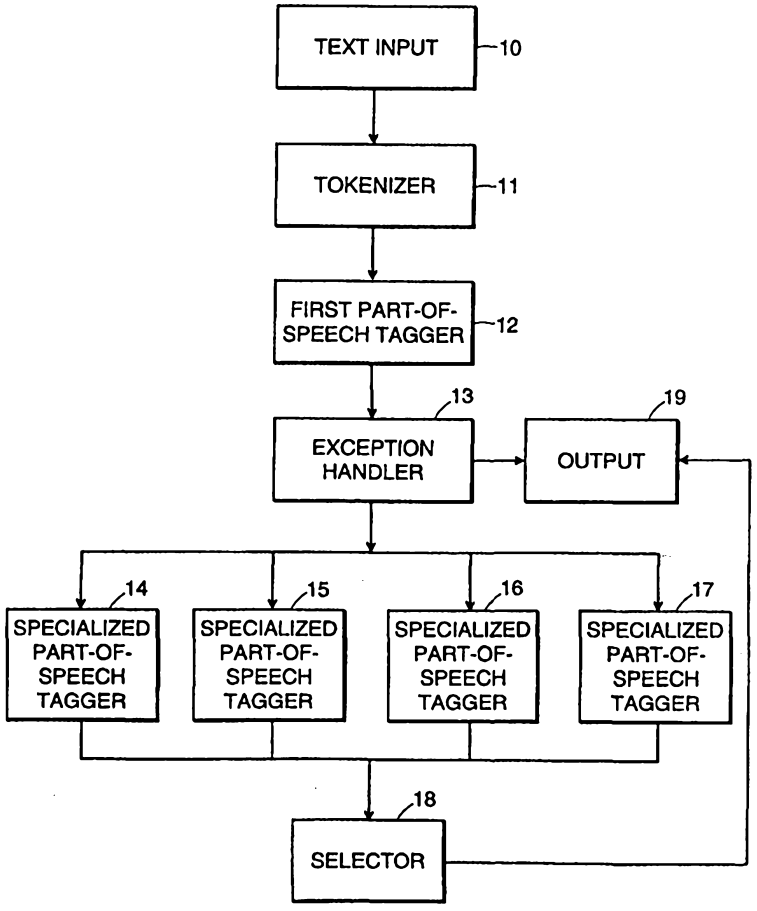
(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: METHOD AND APPARATUS FOR IMPROVED PART-OF-SPEECH TAGGING

(57) Abstract

A tagging device for identifying parts-of-speech of text includes a first part-of-speech tagger that provides, at a first output, a part-of-speech tag for each term in the text and a set of specialized part-of-speech taggers, having an output coupled to a device output and also having an input. The set of specialized part-of-speech taggers provide a set of candidate part-of-speech tags for each term provided at the input to the set of specialized part-of-speech taggers. An exception handler, coupled to the first output provides, in response to each term in the text, a part-of-speech tag from the first output to the device output, unless the term in the text is included in an exception list, in which case the term is provided to the input of the set of specialized part-of-speech taggers. A voting procedure may be used to select a part-of-speech tag from the set of candidate part-of-speech tags produced by the specialised part-of-speech-taggers for terms on the exception list.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Method and Apparatus for Improved Part-of-Speech Tagging

Technical Field

The present invention relates generally to part-of-speech tagging of text and more particularly to the contextual part-of-speech disambiguation of words and phrases in text.

Background Art

The identification of the part-of-speech of words and phrases in text is useful in many different areas such as word and text processing (e.g. proofreading), information retrieval and natural-language database query, information and fact extractions, natural language understanding and machine translation. Many different methods exist by which the part-of-speech may be identified and tagged such as the Markov model, decision tree, connectionist, transformational, nearest neighbor, on-line learning, and maximum entropy. These methods are well described in the art. See, for example Weischedel, R., Meteer, M., Scwartz, R., Ramshaw, L., and Palmucci, J., "Coping with Ambiguity and Unknown Words Through Probabilistic Models", *Computational Linguistics* (1993); Black, E., Jelinek, F., Lafferty, J., Mercer, R., and Roukos, S., "Decision Tree Models Applied to the Labeling of Text with Parts-of-Speech", *Darpa Workshop on Speech and Natural Language* (Harriman, N.Y., 1992); Schmid, H., "Part of Speech Tagging with Neutral Networks," *Proceedings of 15th International Conference on Computational Linguistics* (COLING) (Yokohama, Japan 1994); Brill, E, "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", *Computational Linguistics* 21(4) , pp. 543-565, Dec. 1995; Daelemans, W., Zavrel, P., Berck, P., Gillis, S., "MBT: A Memory-Based Part of Speech Tagger-Generator," *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, pp.14-27, (1996); Ratnaparkhi, A., "A Maximum Entropy Part-of-Speech Tagger," *Proceedings of the*

First Empirical Methods in Natural Language Processing Conference, May 17-18 (University of Pennsylvania 1996). The foregoing references are herein incorporated by reference.

Even the most accurate part-of-speech taggers in the prior art result in some residual error. Improved performance and accuracy may be obtained by producing larger and slower part-of-speech taggers. Another method developed for improving the accuracy of part-of-speech taggers is described in Brill, E., Wu, J., "Classifier Combination for Improved Lexical Disambiguation," *Proceedings of the 19th International Conference on Computational Linguistics and Association for Computational Linguistics* (COLING-ACL)(Montreal, Canada, 1998) and van Halteren, H., Zavrel, J., Daelemans, W., "Improving Data Driven Wordclass Tagging by System Combination," *Proceedings of 19th International Conference on Computational Linguistics and Association for Computational Linguistics*(COLING-ACL)(Montreal, Canada 1998), pp491-497. The foregoing references are herein incorporated by reference. The method described in the foregoing references involves processing the entire text with four different part-of-speech taggers. A part-of-speech tag is then selected from the results of the four part-of-speech taggers using a selection procedure. Although such a method improves accuracy, the improvement comes at a cost of computational speed and complexity.

Summary of the Invention

In accordance with one aspect of the invention, a tagging device for identifying parts-of-speech of terms in a text comprises a first part-of-speech tagger, a set of specialized part-of-speech-taggers and an exception handler. As used in this description and the following claims, the word "set" refers to a set that includes at least one member. The first part-of-speech tagger provides, at a first output, a part-of-speech tag for each term in the text. As used in this description and the following claims, the word "term" refers to a word and optionally to a word or a phrase. In other words, the tagging device is operative on each word in the text, and optionally the tagger may be operative as well on phrases in the text. The set of specialized part-of-speech taggers has an output

coupled to a device output and also has an input and provides a set of candidate part-of-speech tags for each term provided at the input to the set of specialized part-of-speech taggers. The exception handler, coupled to the first output, provides, in response to each term in the text, a part-of-speech tag from the first output to the device output, unless the term in the text is included in an exception list. If the term in the text is included in the exception list, the term is provided to the input of the set of specialized part-of-speech taggers.

In a further embodiment, the set of specialized part of speech taggers includes a plurality of specialized part-of-speech taggers and the tagging device further includes a selector, coupled to the output of the set of specialized part-of-speech taggers. The selector also has an output coupled to the device output. The selector selects a part-of-speech tag from the set of candidate part-of-speech tags using a voting procedure and provides the selected part-of-speech tag at the device output.

In another further embodiment, at least one member of the set of specialized part-of-speech taggers is optimized for processing terms on the exception list. The exception list may include terms which account for a predetermined percentage of errors produced by the first part-of-speech tagger.

In yet another embodiment, the voting procedure generates a score for each unique candidate part-of-speech tag from the set of candidate part-of-speech tags based on predetermined characteristics of each specialized part-of-speech tagger in the set of specialized part-of-speech taggers. The voting procedure may select the part-of-speech tag with the highest score. The tagging device may further include a tokenizer, coupled to the first part-of-speech tagger, for parsing the text into a set of word tokens.

In an alternative embodiment, a method for identifying parts-of-speech of terms in a text comprises: (a) using a first part-of-speech tagger to determine the part-of-speech of each term in the text; (b) identifying each term in the text which is included in an exception list; (c) providing the part-of-speech tag from step (a) as a device output for each term not included in the exception list; and (d) using a set of specialized part-of-speech taggers to determine a set of candidate part-of-speech tags for each term included in the exception list. In a further

embodiment, the method, wherein the set of specialized part-of-speech taggers includes a plurality of taggers, further includes: (e) selecting a part-of-speech tag from the set of candidate part-of-speech tags using a voting procedure and (f) providing the part of speech tag selected in step (e) as the device output for each term included in the exception list.

In a further embodiment, at least one member of the set of specialized part-of-speech taggers is optimized for processing terms on the exception list. The exception list may include terms which account for a predetermined percentage of errors produced by step (a). In the above embodiments, the voting procedure generates a score for each unique candidate part-of-speech tag from the set of candidate part-of-speech tags, the score being based upon predetermined characteristics of each specialized part-of-speech tagger in the set of specialized part-of-speech taggers. The voting procedure may select the candidate part-of-speech tag with the highest score. The method may further include, before step (a), parsing the text into word tokens.

In another alternative embodiment, a digital storage medium encoded with instructions which, when loaded into a computer, may establish any of the devices previously discussed.

Brief Description of the Drawings

The present invention will be more readily understood by reference to the following detailed description taken with the accompanying drawings, in which:

Fig. 1 is a block diagram of a tagging device in accordance with an embodiment of the invention.

Fig. 2 is a block diagram showing the voting procedure utilized by the tagging device of Fig. 1 in accordance with a preferred embodiment of the invention.

Fig. 3 is a block diagram showing the flow of control for a method of part-of-speech tagging in accordance with an embodiment of the invention

Detailed Description of Specific Embodiments

Figure 1 shows a block diagram of a tagging device in accordance with an embodiment of the invention. Text is input at a text input **10** and then the text is parsed into word tokens using a tokenizer **11**. The tokenizer **11** may be one of general use in the art (for example, U.S. Patent No. 5,721,939, "Method and Apparatus for Tokenizing Text", or U.S. Patent No. 4,991,094, "Method for Language-Independent Text Tokenization using a Character Categorization", herein incorporated by reference). The tokenized text is then placed into a text buffer in order to be processed by the tagging device. A first part of speech tagger **12** processes the tokenized text. The first part-of-speech tagger **12** may be one of general use in the art such as Markov model, decision tree, connectionist, transformational, nearest neighbor, on-line learning, or maximum entropy. Preferably, the first part-of-speech tagger **12** is a fast and accurate part-of-speech tagger. In one embodiment of the invention, the first part-of-speech tagger **12** is a Brill transformational tagger implemented by an Abney-like finite-state automaton (FSA) (See Brill, E., "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging," *Computational Linguistics* 21 (4), Dec. 1995, pp. 543-565, herein incorporated by reference). Accordingly, the first part of speech tagger **12** is produced by generating Brill part-of-speech tagging transformational rules against a first part-of-speech tagged corpus of text.

An exception handler **13** is coupled to the first part-of-speech tagger **12**. If the term being processed is not found on an exception list, the part-of-speech tag identified by the first part of speech tagger will be the output **19** of the tagging device. When the exception handler **13** encounters a term found on the exception list, the term is routed to a set of specialized part-of- speech taggers **14-17** coupled to the exception handler **13** for further processing. The set of specialized part-of-speech taggers may include n members, where n can be a number greater than or equal to one. In the embodiment shown in Figure 1, the set of specialized part-of-speech taggers includes four specialized part-of-speech taggers **14-17**.

Preferably, the exception list includes terms which are known to have inaccurate tagging results using the first part-of-speech tagger **12**. The terms included in the exception list are identified by running the first part-of-speech tagger **12** against a second part-of-speech tagged corpus of text to identify the residual error of the first part-of-speech tagger **12**. The frequency distribution of the part-of-speech tagging errors by the term associated with the errors is generated in order to identify the terms which account for the most frequently occurring errors produced by the first part-of-speech tagger **12**. Terms which account for a predetermined percentage of the errors produced by the first part-of-speech tagger **12** are included in the exception list. In one embodiment, the predetermined percentage is 90%.

Each specialized part-of-speech tagger is generated using the exception list described above. Each of the specialized part-of-speech taggers **14-17** may be one generally known in the art. As discussed above, some examples of part-of-speech taggers are the Markov model, decision tree, connectionist, transformational, nearest neighbor, on-line learning, and maximum entropy. The specialized part-of-speech taggers are produced by the methods appropriate to each style of tagger, however, each specialized part-of-speech tagger is trained specifically on the terms included in the exception list. Preferably, each specialized part-of-speech tagger is of a different type. In one embodiment, the specialized part-of-speech taggers **14-17** are trigram, Brill transformational, memory-based learning, and maximum entropy part-of-speech taggers.

If there is only one specialized part-of-speech tagger in the set of specialized part-of-speech taggers, the output of the specialized part-of-speech tagger is the device output **19** for each term in the text that is included in the exception list. As discussed above, for each term not found on the exception list, the device output **19** will be the output of the first part-of-speech tagger **12**.

If the set of part-of-speech taggers consists of a plurality of specialized part-of-speech taggers, as shown in Figure 1, each specialized part-of-speech tagger **14-17** will produce a candidate part-of-speech tag for the term being processed by the set of specialized part-of-speech taggers. Each candidate part-

of-speech tag produced by the set of specialized part of speech taggers is provided to a selector 18. The selector 18 uses a voting procedure to select one of the candidate part-of-speech tags. Figure 2 is a block diagram showing the voting procedure according to an embodiment of the invention. At block 20, each specialized part-of-speech tagger processes the term and identifies a candidate part-of-speech tag. At block 21, the voting procedure creates a list of unique candidate part-of-speech tags identified by the specialized part-of-speech taggers. A score is then calculated for each unique candidate part-of-speech tag at block 22.

In one embodiment, the voting procedure uses pre-computed values of precision and recall for each specialized part-of-speech tagger to calculate a score (block 22) for each unique candidate part-of-speech tag produced by the set of specialized part-of-speech taggers. Precision is defined as the percentage of tokens tagged X by the part-of-speech tagger that are also tagged X in the training corpus. Recall is defined as the percentage of tokens tagged X in a training corpus that are also tagged X by the part-of-speech tagger. For example, the word "that" has several possible parts of speech, such as coordinating conjunction (CS), determiner (DT), qualifier (QL) or WH-pronoun (WPR). If a specialized part-of-speech tagger produced the tag DT in fifty instances of the word "that" of which forty-five as identified by the training corpus are correct, then the precision is .90 (=45/50). If there were fifty instances of "that" tagged DT in the training corpus and the specialized tagger tagged forty-eight of them as DT, then the recall is .96 (=48/50).

As mentioned above, the values of precision and recall may be used to determine the score for each unique candidate part-of-speech tag at block 22. The score for a candidate part-of-speech tag may be calculated by adding the precision of each specialized part-of-speech tagger which produced a particular candidate part-of-speech tag to an amount equal to (1-recall) of each specialized part-of-speech tagger which produced the particular candidate part-of-speech tag. The candidate part-of-speech tag with the highest accumulated score is

selected at block 23 as the part-of-speech tag for the term being processed by the set of specialized part-of-speech taggers.

Table 1 shows example results for the word "that" using a set of specialized part-of-speech taggers consisting of trigram, Brill transformational, memory-based learning, and maximum entropy part-of-speech taggers. The candidate part-of-speech tags are defined as determiner (DT) and coordinating conjunction(CS).

Specialized Tagger	Candidate Tag	Precision	Recall
Trigram	DT	.83	.93
Transformational	CS	.87	.87
Memory-Based	DT	.88	.89
Maximum Entropy	CS	.91	.93

Table 1. Example results from specialized taggers for the term "that" in a given instance

The calculation of the scores for the candidate part-of-speech tags "DT" and "CS" would be as follows:

$$Score_{DT} = .83 + .88 + (1-.93) + (1-.89) = 1.89$$

$$Score_{CS} = .87 + .91 + (1-.87) + (1-.93) = 1.98$$

In this example, the candidate part-of-speech tag CS has the higher score and would be selected as the part-of-speech tag for the word "that".

Returning to Figure 1, the output **19** of the tagging device will be the output of selector **18** for each term in the text that is found on the exception list. Otherwise, the output **19** of the tagging device will be the output of the first part-of-speech tagger **12**. The use of the specialized part-of-speech taggers **14-17** in combination with the first part-of-speech tagger **12** improves the performance and accuracy of the first part-of-speech tagger **12**. This is accomplished by

training each specialized part of speech tagger **14-17** to improve the accuracy of those terms which produce the largest error rates for the first part-of-speech tagger **12**.

Figure 3 illustrates the flow of control for a method of identifying the parts-of-speech of terms in a text in accordance with an embodiment of the invention. The text input at block **30** is parsed into word tokens at block **31**. The tokenized text is then placed in a text buffer at block **32** and processed at block **33** by a first part-of-speech tagger. At block **34**, if the term being processed is not found on an exception list, the output at block **37** will be the part-of-speech tag produced by the first part-of-speech tagger. The exception list is described above with respect to Figure 1. If the term being processed is found on the exception list, the term will be processed by a set specialized part-of-speech taggers at block **35**. The set of specialized part-of-speech taggers produce a set of candidate part-of-speech tags. If the set of specialized part of speech taggers includes only one specialized part-of-speech tagger, the output at block **37** will be the output of the specialized part-of-speech tagger as determined at block **35**. If the set of specialized part-of-speech taggers includes a plurality of specialized part-of-speech taggers, a voting procedure is used at block **36** to select a part-of-speech tag from the set of candidate part-of-speech tags. The voting procedure for an embodiment of the invention is described above with respect to Figure 2. At block **37** the output for a term in the text that is found on the exception list will be the part-of-speech tag selected in step **36**.

Although various exemplary embodiments of the invention have been disclosed, it should be apparent to those skilled in the art that various changes and modifications can be made which will achieve some of the advantages of the invention without departing from the true scope of the invention. These and other obvious modifications are intended to be covered by the appended claims.

What is claimed is:

1. A tagging device for identifying parts-of-speech of text, the device comprising:

a first part-of-speech tagger that provides, at a first output, a part-of-speech tag for each term in the text;

a set of specialized part-of-speech-taggers, having an output coupled to a device output and also having an input, the set of specialized part-of-speech taggers providing a set of candidate part-of-speech tags for each term provided at the input to the set of specialized part-of-speech taggers; and

an exception handler, coupled to the first output, to provide, in response to each term in the text, a part-of-speech tag from the first output to the device output, unless the term in the text is included in an exception list, in which case the term is provided to the input of the set of specialized part-of-speech taggers.

2. A tagging device according to claim 1, wherein the set of specialized part of speech taggers includes a plurality of specialized part-of-speech taggers, the device further comprising:

a selector, coupled to the output of the set of specialized part-of-speech taggers, the selector having an output coupled to the device output, for selecting a part-of-speech tag from the set of candidate part-of-speech tags using a voting procedure and providing the selected part-of-speech tag at the device output.

3. A tagging device according to claim 1, wherein at least one member of the set of specialized part-of-speech taggers is optimized for processing terms on the exception list.

4. A tagging device according to claim 1, wherein the exception list includes terms which account for a predetermined percentage of errors produced by the first part-of-speech tagger.

5. A tagging device according to claim 2, wherein the voting procedure generates a score for each unique part-of-speech tag from the set of candidate part-of-speech tags based on predetermined characteristics of each specialized part-of-speech tagger in the set of specialized part-of-speech taggers.

6. A tagging device according to claim 5, wherein the voting procedure selects the candidate part-of-speech tag with the highest score.

7. A tagging device according to claim 1, further including a tokenizer, coupled to the first part-of-speech tagger, for parsing the text into a set of word tokens.

8. A method for identifying parts-of-speech of text, the method comprising;

(a) using a first part-of-speech tagger, to determine the part-of-speech of each term in the text;

(b) identifying each term in the text which is included in an exception list;

(c)providing the part-of-speech tag from step (a) as a device output for each term not included in the exception list; and

(d) using a set of specialized part-of-speech taggers to determine a set of candidate part-of-speech tags for each term in the text that is included in the exception list.

9. A method according to claim 8, wherein the set of specialized part-of-speech taggers includes a plurality of taggers, the method further including:

(e) using a voting procedure to select a part-of-speech tag from the set of candidate part-of-speech tags; and

(f) providing the part of speech tag selected in step (e) as the device output for each term in the text that is included in the exception list.

10. A method according to claim 8, wherein at least one member of the set of specialized part-of-speech taggers is optimized for processing terms on the exception list.

11. A method according to claim 8, wherein the exception list includes terms which account for a predetermined percentage of errors produced by step (a).

12. A method according to claim 9, wherein the voting procedure generates a score for each unique candidate part-of-speech tag from the set of candidate part-of-speech tags, the score being based upon predetermined characteristics of each specialized part-of-speech tagger in the set of specialized part-of-speech taggers.

13. A method according to claim 12, wherein the voting procedure selects the part-of-speech tag with the highest score.

14. A method according to claim 8, further including, before step (a), parsing the text into word tokens.

15. A digital storage medium encoded with instructions which, when loaded into a computer, establishes a device for identifying the parts-of-speech of text, the device including:

a first part-of-speech tagger that provides, at a first output, a part-of-speech tag for each term in the text;

a set of specialized part-of-speech-taggers, having an output coupled to a device output and also having an input, the set of specialized part-of-speech taggers providing a set of candidate part-of-speech tags for each term provided at the input to the set of specialized part-of-speech taggers; and

an exception handler, coupled to the first output, to provide, in response to each term in the text, a part-of-speech tag from the first output to the device output, unless the term in the text is included in an exception list, in

which case the term is provided to the input of the set of specialized part-of-speech taggers.

16. A storage medium according to claim 15, wherein the set of specialized part of speech taggers includes a plurality of specialized part-of-speech taggers, the device further comprising:

a selector, coupled to the output of the set of specialized part-of-speech taggers, the selector having an output coupled to the device output, for selecting a part-of-speech tag from the set of candidate part-of-speech tags using a voting procedure and providing the selected part-of-speech tag at the device output.

17. A storage medium according to claim 15, wherein at least one of the set of specialized part-of-speech taggers is optimized for processing terms on the exception list.

18. A storage medium according to claim 15, wherein the exception list includes terms which account for a predetermined percentage of errors produced by the first part-of-speech tagger.

19. A storage medium according to claim 16, wherein the voting procedure generates a score for each unique part-of-speech tag from the set of candidate part-of-speech tags, the score being based upon predetermined characteristics of each specialized part-of-speech tagger in the set of specialized part-of-speech taggers.

20. A storage medium according to claim 19 wherein the voting procedure selects the part-of-speech tag with the highest score.

21. A storage medium according to claim 15 , the device further including a tokenizer, coupled to the first part-of-speech tagger, for parsing the text into a set of word tokens.

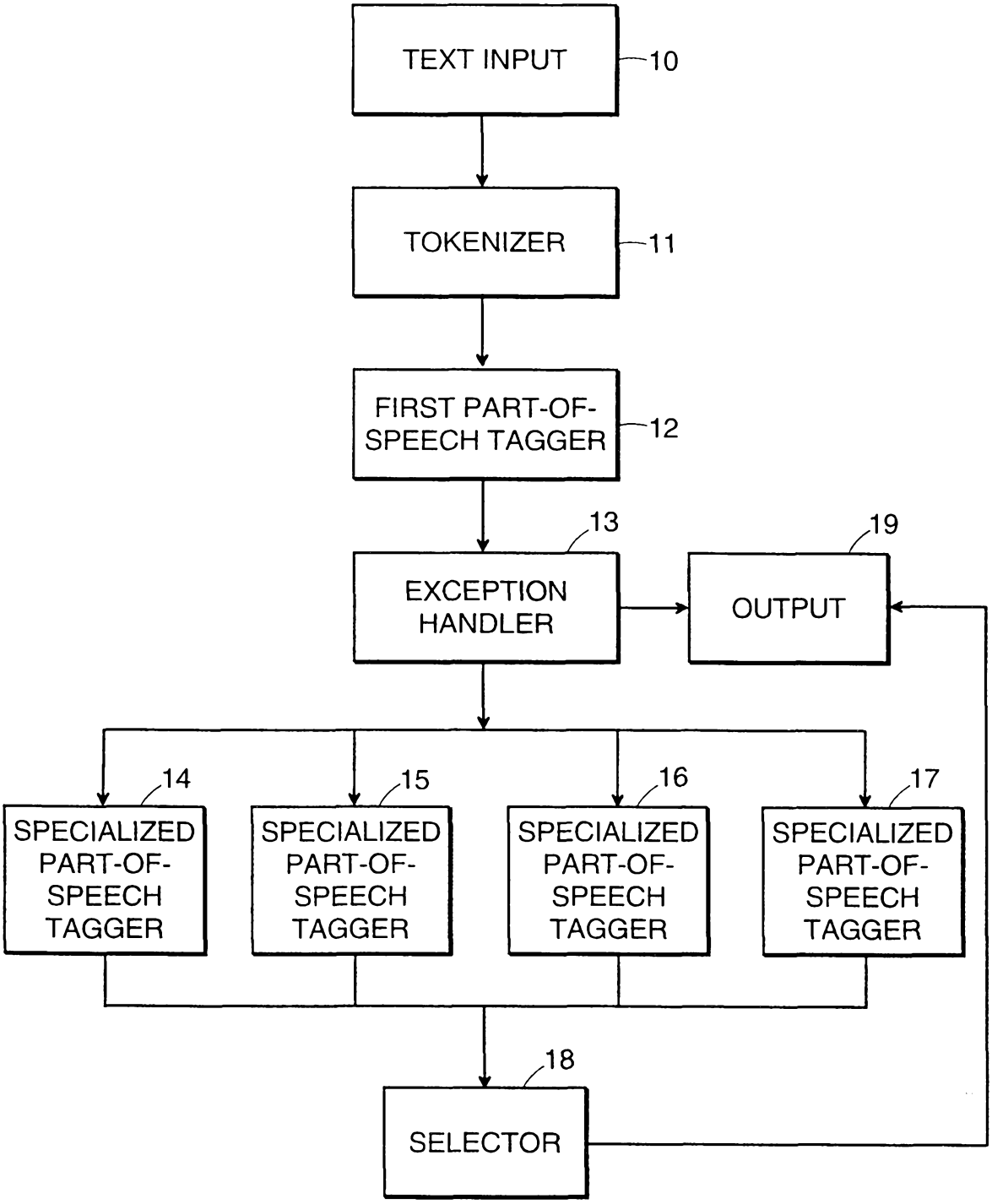

TEXT INPUT
10
TOKENIZER
11
FIRST PART-OF-SPEECH TAGGER
12
13
EXCEPTION HANDLER
19
OUTPUT
14
SPECIALIZED PART-OF-SPEECH TAGGER
15
SPECIALIZED PART-OF-SPEECH TAGGER
16
SPECIALIZED PART-OF-SPEECH TAGGER
17
SPECIALIZED PART-OF-SPEECH TAGGER
18
SELECTOR

FIG. 1

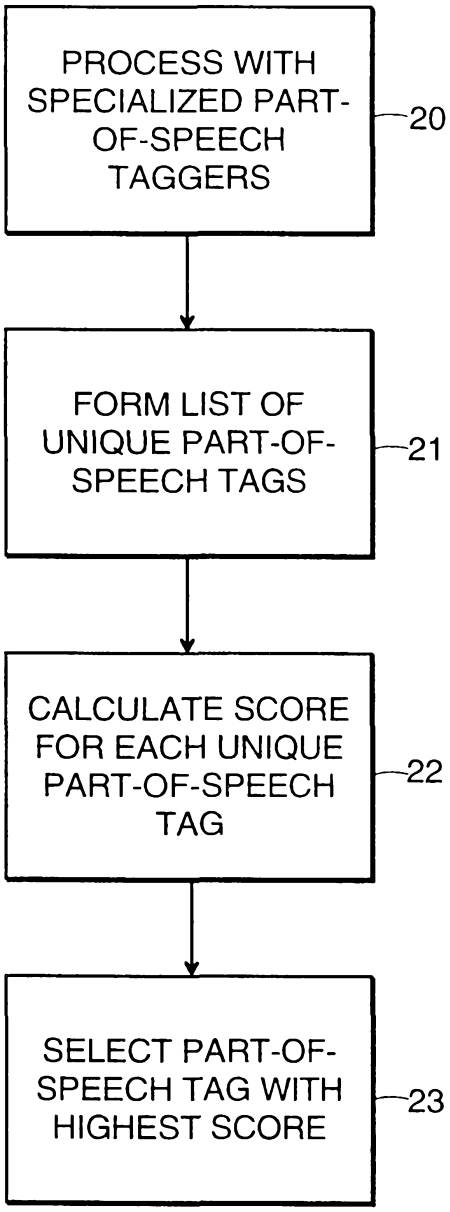

PROCESS WITH SPECIALIZED PART-OF-SPEECH TAGGERS
20
FORM LIST OF UNIQUE PART-OF-SPEECH TAGS
21
CALCULATE SCORE FOR EACH UNIQUE PART-OF-SPEECH TAG
22
SELECT PART-OF-SPEECH TAG WITH HIGHEST SCORE
23

FIG. 2

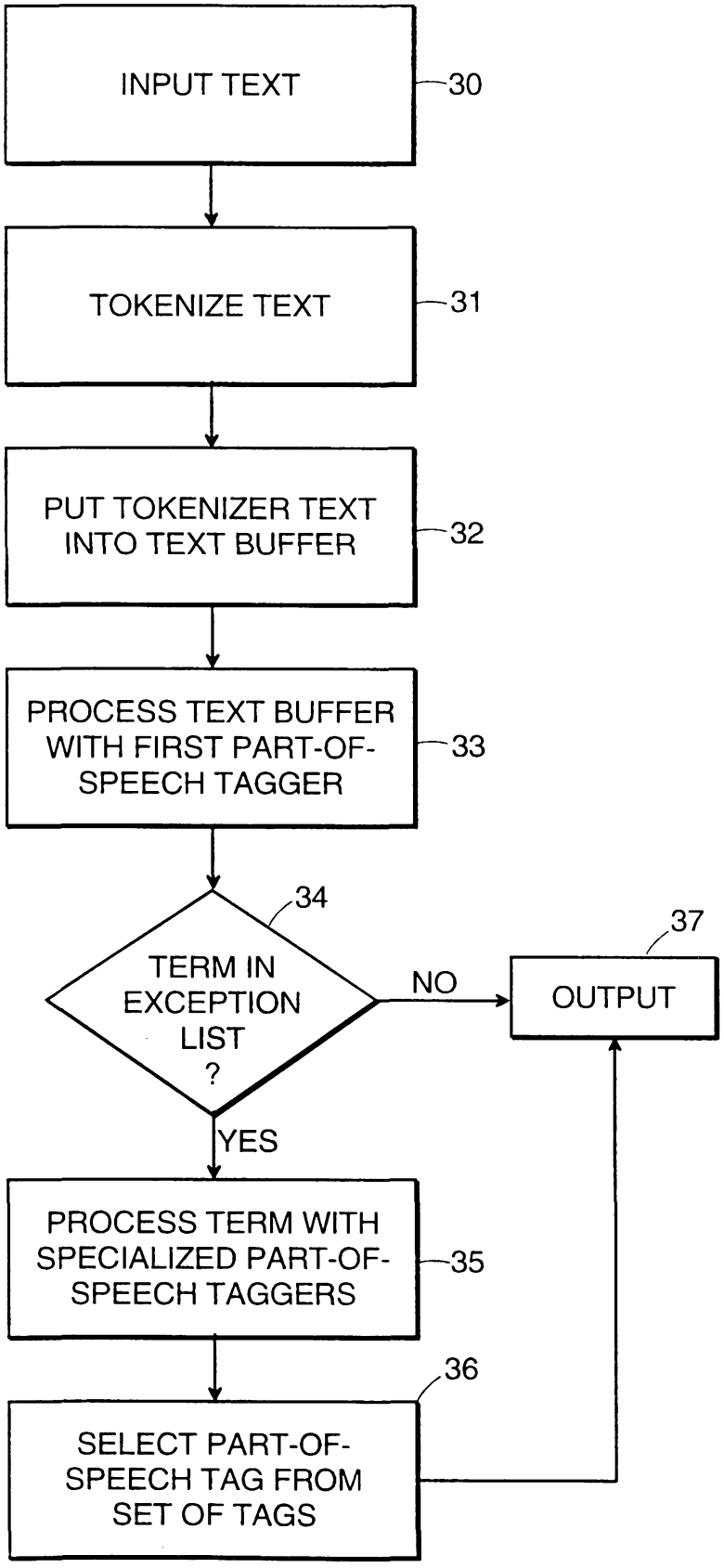

INPUT TEXT
30
TOKENIZE TEXT
31
PUT TOKENIZER TEXT INTO TEXT BUFFER
32
PROCESS TEXT BUFFER WITH FIRST PART-OF-SPEECH TAGGER
33
34
TERM IN EXCEPTION LIST ?
NO
37
OUTPUT
YES
PROCESS TERM WITH SPECIALIZED PART-OF-SPEECH TAGGERS
35
36
SELECT PART-OF-SPEECH TAG FROM SET OF TAGS

FIG.3