



- (51) **International Patent Classification:**
G06F 17/27 (2006.01)
- (21) **International Application Number:**
PCT/MY2014/000153
- (22) **International Filing Date:**
29 May 2014 (29.05.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
PI 2013004281 27 November 2013 (27.11.2013) MY
- (71) **Applicant:** MIMOS Berhad [MY/MY]; Technology Park Malaysia, 57000 Kuala Lumpur (MY).
- (72) **Inventors:** CHU MIN XIAN, Benjamin; c/o MIMOS Berhad, Technology Park Malaysia, 57000, Kuala Lumpur (MY). BAHLS, Daniel; c/o MIMOS Berhad, Technology Park Malaysia, 57000, Kuala Lumpur (MY). LUKOSE, Dickson; c/o MIMOS Berhad, Technology Park Malaysia, 57000, Kuala Lumpur (MY).
- (74) **Agent:** YAP, Kah Hon; Pyprus Sdn Bhd, Suite 8-02, 8th Floor, Plaza First Nationwide 161, Jalan Tun H.S. Lee, 50000, Kuala Lumpur (MY).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): 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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, 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).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— with international search report (Art. 21(3))

(54) **Title:** A METHOD AND SYSTEM FOR AUTOMATED ENTITY RECOGNITION

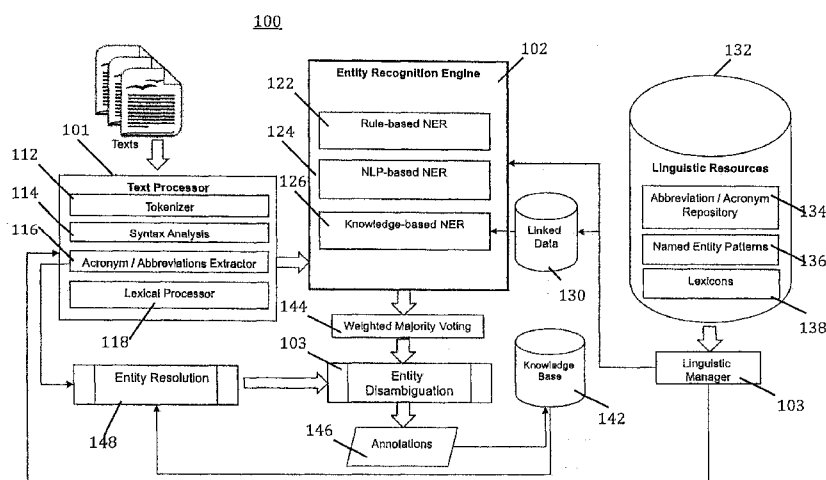


FIG. 1

(57) **Abstract:** The present invention provides a system for extracting concept and named-entities from a text-containing document. An entity recognition engine (102) is provided to process an entity with a Rule-based Named-Entity Recognition (NER) (122), a Natural-Language-Processing (NLP) based NER (124), and a knowledge-based NER (126). The NERs are further scored and weighted, wherein the highest weighted score will be taken. A method thereof is also provided.

A Method and System for Automated Entity Recognition

Field of the Invention

[0001] The present invention relates to information extraction. More
5 specifically, the present invention relates to a system and method for automated entity recognition.

Background

[0002] Entity Recognition, a subset of information extraction, is adapted for recognizing and extracting entities that include personal name, location, and
10 organization, and etc. from a document. In general, some features, such as string label, indicator function for each word are used for the entity recognition. The labeled training data and extracted features are then used to train a classifier to obtain optimized parameters through a numerical optimization algorithm. Given the features of a word in testing data, the classifier is able to recognize the label of which it belongs to.

15 [0003] State-of-the-art Named-Entity Recognition (NER) techniques identify entities based on rules or statistical training methods. Such methods are at least limited in their classification abilities, because the language structure of a sentence itself does not give enough clues for determining the category of an entity at a fine-grained level. Misclassification is a main setback of the conventional methods.

[0004] The problems existed in the state-of-the art system and method for NER are known mainly with the following challenges: 1) Training datasets dependency (limited gazetteers); and 2) Inflexibility for fine-grained entity type classification.

[0005] In relation to training datasets dependency, it is known that NER systems developed for one domain do not typically perform well on another domains. Considerable effort is required in tuning an NER system intended for one domain into another for a new domain; this is true for both rule-based and trainable statistical systems. It remains a question whether the altered system is capable of recognizing generic and domain-specific entities after the effort has been spent.

10 [0006] Further, state-of-the-art NER systems is adapted for a fixed entity type, for example, PERSON, LOCATION, ORGANIZATION, etc, is usually inflexible for configuration.

Summary

[0007] In accordance with one aspect of the present invention, there is a system for extracting concept and named-entities from a text-containing document having sentences therein. The system comprises a text processor adapted for tokenizing the sentences to extract possible entities from the document; an entity recognition module for identifying concepts and named-entities; an entity resolution module for resolving acronyms or abbreviations identified from the sentence into recognized entities based on a knowledge base; an entity disambiguation module adapted for disambiguating entities based on a sentence context. The entity recognition module operationally carries out a Rule-based Named-Entity Recognition (NER), a Natural-Language-Processing (NLP) based NER, and a knowledge-based NER on each entity identified

through the text processor, wherein a score is assigned to each NER, and a result of the NER having a highest score will be stored into a knowledge base.

[0008] In one embodiment, the system further comprises a weighted majority voting module, wherein each NER is pre-assigned with a weighted vote, and the score
5 assigned to each NER on the entity is further weighted with the respective weighted votes to obtaining ranking scores, a result of the NER with the highest ranking score of the entity will be stored into the knowledge base.

[0009] In another embodiment, the text processor comprises a tokenizer for tokenizing the sentences, a syntax analyzer for parsing the tokenizing sentences to
10 identify noun phrases, an acronym/abbreviation extractor adapted for identifying acronym and abbreviations through the entity resolution module, and a lexical processor for identifying noun phrases and compound words.

[0010] In a further embodiment, the system may further comprise linguistic resources having an abbreviation/acronym repository, a named-entity and a lexicon
15 database, wherein the linguistic resources is operationally accessed by the text processor and the entity recognition engine; and Linked Data that is operationally accessed by the knowledge-based NER for processing entities. There is further provided a pattern matcher adapted for searching entity patters from the linguistic resources and a term finder adapted for finding and matching a target string from the Linked Data to
20 processes the target string to identify entities for processing the knowledge-based NER.

[0011] In another aspect of the present invention, there is also provided a method of extracting concept and named-entities from a text-containing document having sentences therein. The method comprises tokenizing the sentences for extracting

possible entities from the document; performing a Rule-based Named-Entity Recognition (NER) on the entities; performing a Natural-Language-Processing (NLP) based NER on the entities; performing a knowledge-based NER on the On each entity identified through the text processor; scoring each of the NERs processed on the
5 entities; resolving acronyms and abbreviations identified from the sentence into recognised entities based on a knowledge base; disambiguating entities based on a sentence context; and storing the entity that is processed by the NER with a highest score.

[0012] In another embodiment, the method further comprises defining a
10 weighted vote to each NER; weighting each NER's score for each entity to obtain a ranking score for each NER; storing the results of the NER having the highest ranking score on the knowledge database.

[0013] The step of tokenizing the sentence may further comprise tokenizing the sentences; parsing the tokenized sentences to identify noun phrases; identifying
15 acronym and abbreviations through an entity resolution module; and identifying noun phrases and compound words through lexical processor.

[0014] In yet another embodiment, the method may further comprise accessing the linguistic resources by the text processor and the entity recognition engine to extract and recognize entities; and accessing the Linked Data by the knowledge-based NER to
20 process the entity. The linguistic resources have an abbreviation/acronym repository, a named-entity and a lexicon database. Further, it may also comprise searching entity patterns through a pattern matcher from the linguistic resources; and finding and

matching a target string a term finder from the Linked Data to processes the target string to identify entities for processing the knowledge-based NER.

Brief Description of the Drawings

[0015] Preferred embodiments according to the present invention will now be described with reference to the figures accompanied herein, in which like reference numerals denote like elements;

[0016] FIG. 1 illustrates a block diagram of a Named-Entity Recognition (NER) system in accordance with one embodiment of the present invention;

[0017] FIG. 2 illustrates a process of recognizing entity in accordance with one embodiment of the present invention;

[0018] FIGs. 3A-C exemplify an example of a sentence that is being processed to extract its entities therefrom; and

[0019] FIG. 4 exemplifies a scoring and weighted voting process of an entity E1.

15

Detailed Description

[0020] Embodiments of the present invention shall now be described in detail, with reference to the attached drawings. It is to be understood that no limitation of the scope of the invention is thereby intended, such alterations and further modifications in the illustrated device, and such further applications of the principles of the invention as illustrated therein being contemplated as would normally occur to one skilled in the art to which the invention relates.

20

[0021] FIG. 1 illustrates a block diagram of a Named-Entity Recognition (NER) system 100 in accordance with one embodiment of the present invention. The NER system 100 is capable of automatically identifying concepts and named-entities from a document having natural language texts. The NER system 100 comprises a text processor 101, an entity-recognition engine 102, linguistic resources manager 103, a
5 disambiguation module 104.

[0022] A document 190 having machine-readable text is inputted to the system 100 for processing. The document 190 can be originated in various formats available. The text is first processed by the text processor 101 to extract and tokenize the
10 sentences for all possible entities. The text processor 101 includes a tokenizer 112, a syntax analyser 114, an acronym/abbreviation extractor 116 and a Lexical processor 118 for processing the text. The tokenizer 112 is provided for splitting the document into tokenized sentences and eventually into tokenized words. The main role of the tokenizer is to produce eventually a series of word tokens as the basic units for syntax
15 analyser to produce the corresponding syntax or part-of-speech classes for these units. Tokenization techniques are well known in the art, and it is possible to adapt any of those in the present invention. The syntax analyser 114 is adapted for analysing syntax to produce syntax structure and entity patterns. It parses the tokenized sentences to identify all possible noun phrases to be matched by a pattern matcher for entity
20 recognition. The acronym / abbreviations extractor 116 extracts all acronyms / abbreviations for resolving by the entity resolution module 148. The lexical processor 118 is adapted to identify noun phrase or compound words from the target string. It supplements the text preprocessing in generating noun phrases and compound words. It plays a role to identify whether a token exists as a single compound word or noun

phrase, or there are several tokens that made up the compound word or noun phrase. It helps the entity recognition module to identify entities.

[0023] The linguistic manager 103 is connected to the linguistic resource knowledge base 132. The linguistic manager 103 is adapted to manage the interaction
5 between the pattern matcher and linguistic resources, and the interaction between the term finder and the Linked Data 130. The linguistic resources knowledge base 132 has an abbreviation/acronym repository 134, a named-entity pattern repository 136 and a lexicon 138. The lexicon may include generic and specific lexicons for various domains. The linguistic manager 103 feeds the required information from the linguistic
10 resources database 132 to the entity recognition engine 102 for processing the processed text at the entity recognition engine 102.

[0024] The Linked Data 130 is used with referenced to the linguistic resources in identifying textual representations (such as names, labels, etc) for each of the knowledge base concepts.

15 [0025] After the text is being processed by the text processor 101 to obtain a potential entity, the potential entity is further processed by the entity recognition engine 102. It is adapted for identifying concept and named-entities. The entity recognition engine 102 has a rule-based NER 122, a natural-language-processing (NLP) based NER 124 and a knowledge-based NER 126 for processing the text. In this entity recognition
20 engine, the text are process under the Rule-based, the NLP-based and the knowledge-based NER respectively, and a score is assigned under each NER process for the text. The scores will be used to compare with each other to determine which NER results shall be taken.

[0026] To process the entities, the entity recognition module 102 comprises a pattern matcher and a term finder. The preprocessed texts are being process by the pattern matcher to search for entity patterns from the linguistic resources 132, and subsequently the term finder for finding and matching a target string from Linked Data 5 130 for identifying entities for processing the knowledge-based NER 126.

[0027] The scores from the entity recognition engine 102 are further weighted through the weighted majority voting module 144. The weighted majority-voting module 144 computes a final ranking score for results filtering between the rule-based, NLP-based, and knowledge-based NERs. The results of the NER that obtained the 10 highest ranking score will be annotated through the annotator 146 and stored into the knowledge base 142.

[0028] The Linked Data 130 provides the background knowledge for the knowledge-based NER 126. The Link Data 130 may be taken or extracted from any opened knowledge bases, such as DbPedia, Agrovoc, Freebase, etc. When it is 15 combined with the linguistic resources 132 that include thesaurus, dictionary, etc, they can be configured dynamically to produce a knowledge pattern structure, i.e. dynamic configuration regardless of the domain of the knowledge base. With this enriched knowledge pattern structure, it provides sufficient and rich background knowledge to the knowledge based NER 126 for entity recognition.

20 [0029] The entity resolution module 148 is configured to resolve acronyms or abbreviations of the target sentence into the recognized entities from the knowledge base 142.

[0030] The entity disambiguation module 104 is configured to disambiguate entities based on the sentence context. Many systems and methods for disambiguate entities are well known in the art and they can be adapted for the present invention. In one embodiment of the present invention, the disambiguation system and method
5 disclose in the Malaysia patent application entitled "A METHOD AND SYSTEM FOR AUTOMATED WORD SENSE DISAMBIGUATION" filed on the same day as the present application, can also be adapted.

[0031] FIG. 2 illustrates a process of recognizing entity in accordance with one embodiment of the present invention. The process starts with searching for entity
10 pattern through a pattern matcher at step 202. The method is done by referencing the tested text, which is a potential entity determined through the text processor 101, with the linguistic resources from the linguistic resources database 132. It follows with a step 204 of finding and matching the potential entity from the linked data 130 through a term finder. When the potential entity can be found or matched with the Linked Data
15 130 at step 206, a property type for the potential entity is being retrieved at step 209 for knowledge-based NER 126. When no match can be found with the Linked Data 130, the potential entity is processed under the rule-based NER 122 and the NLP-based NER at steps 207 and 208 respectively. The potential entity is being processed by the Role-based NER 122 or the NLP-based NER 124 or the knowledge-based NER 126, the
20 potential entity processed by the entity recognition engine 102 will be assigned with a matching score of similarity measure at step 210. The system 100 weighted for majority voting at step 212. The potential entity with the best score is selected at step 214. The entity is then stored on the knowledge base 412. The entity is stored together with the related information resulting from the NERs.

[0032] FIGs. 3A-C exemplify an example of a sentence that is being processed to extract its entities therefrom. The exemplified sentence is *"The iPad Mini is a mini tablet computer designed, developed, and marketed by Apple Inc. which was announced on October 23, 2012."*. The exemplified sentence can also herewith referred as a target string. As shown in FIG. 3A, the target string is parsed by the text processor 101 to retrieve its phrase syntax structure through the syntax analyzer 114. It uses the pattern matcher to search for the entity patterns from the linguistic resources. Among the targeted string, the words "Apple" and "Inc." being a pattern number 1 will be identified as an entity, or named-entity. And the series of string "October", "23", ",", and "2012" being identified as pattern number 2 will be identified as a date.

[0033] As shown in FIG. 3B, the targeted string is further processed to extract possible named entities through the pattern matcher to identify all named-entities from the target string above. In this case, "iPad Mini" and "Apple Inc" will be identified as the possible entities and the "October 23, 2012" will be identified as date. The system will go on to search and match for possible acronyms/abbreviations. The entity resolution module 148 is also used to resolve the existing entities when required. In addition to the syntax structure, the text processor also perform noun phrase or compound words identification through lexical processor 118.

[0034] Returning to the entity resolution module 148, its main role is for resolving the entities to identify acronyms/abbreviations. Systems and methods for resolving entities to identify acronyms/abbreviations therefrom are well known in the art and they can be adapted for the present invention. In one proposed embodiment, the entity resolution may include searching and querying the Linked Data 130 and linguistic resources 132 to find all related information for each entity, storing these information as

context vectors, generating the number of salient words from the document or text, $N \sim \text{Poisson}(x)$, where x is a fixed prior parameter. Salient words refer to important and meaningful words (e.g. nouns), exclusive stopwords (e.g. preposition words). Based on the identified salient words, a document vector can be formed. A similarity function
 5 can be used to compute similarity scores by iterating through each of the elements between the context and document vectors and the final score is averaged out to resolve entities.

[0035] As shown in FIG. 3C, with the potential named-entities identified, the term finder finds and matches the target string from the Linked Data. For example, the
 10 label extracted from the Linked Data 130 is being matched with the target string. The label is defined in the Linked Data 130. Through the property type available in the Linked Data 130, such as `rdfs:label`, the string label can be obtained. Each of the potential entities is being assigned with a match score with similarity measure.

[0036] FIG. 4 exemplifies derivations of a ranking score for an entity for one
 15 embodiment of the present invention. As provided in table 402, an entity E1 that has been processed based on NLP-based, Rule-based and knowledge-based NERs are assigned with similarity scores of 0.67, 0.51 and 0.70 respectively. The weighted votes for NLP-based, Rule-based and knowledge-based NERs are preset as 1, 3 and 5 respectively as shown in table 404. The scores in table 402 is then multiply directly
 20 with the weighted votes in table 404 to obtain the ranking scores of the entity E1 that are processed under NLP-based, Rule-based and knowledge-based NERs. The highest ranking score that falls above a certain specific threshold 408 is taken, for which, the entity that processed under the corresponding NER is stored on the knowledge base 142. The specific threshold can be defined by the user.

[0037] While specific embodiments have been described and illustrated, it is understood that many changes, modifications, variations, and combinations thereof could be made to the present invention without departing from the scope of the invention.

Claims

1. A system (100) for extracting concept and named-entities from a text-containing document having sentences therein, the system comprising:
 - a text processor (101) adapted for tokenizing the sentences to extract possible
 - 5 entities from the document;
 - an entity recognition module (102) for identifying concepts and named-entities wherein the entity recognition module (102) operationally carries out a Rule-based Named-Entity Recognition (NER) (122), a Natural-Language-Processing (NLP) based NER (124), and a knowledge-based NER (126) on each entity identified through the
 - 10 text processor (101);
 - an entity resolution module (148) for resolving acronyms or abbreviations identified from the sentence into recognized entities based on a knowledge base (142);
 - and
 - an entity disambiguation module (103) adapted for disambiguating entities based
 - 15 on a sentence context,
 - wherein a score is assigned to each NER, and a result of the NER having a highest score will be recorded on a knowledge base (142).
2. The system (100) according to claim 1, further comprising a weighted majority
- 20 voting module (144), wherein each NER is pre-assigned with a weighted vote, and the score assigned to each NER on the entity is further weighted with the respective weighted votes to obtaining ranking scores, a result of the NER with the highest ranking score of the entity will be recorded on the knowledge base (142).
- 25 3. The system (100) according to claim 1, the system (100) further comprising:
 - linguistic resources (132) having an abbreviation/acronym repository (134), a
 - named-entity (136) and a lexicon database 138, wherein the linguistic resources (132) is
 - operationally accessed by the text processor (101) and the entity recognition engine
 - (102); and
 - 30 Linked Data (130) that is operationally accessed by the knowledge-based NER
 - (126) for processing entities.

4. The system (100) according to claim 4, the system (100) further comprising:
a pattern matcher adapted for searching entity patterns from the linguistic
resources and a term finder adapted for finding and matching a target string from the
5 Linked Data (130) to processes the target string to identify entities for processing the
knowledge based NER.
5. A method of extracting concept and named-entities from a text-containing
document (190) having sentences therein, the method comprising:
- 10 tokenizing the sentences for extracting possible entities from the document
(190);
- performing a Rule-based Named-Entity Recognition (NER) (122) on the
entities;
- performing a Natural-Language-Processing (NLP) based NER (124) on the
15 entities;
- performing a knowledge-based NER (126) on the on each entity identified
through the text processor (101);
- scoring each of the NERs processed on the entities;
- resolving acronyms and abbreviations identified from the sentence into
20 recognised entities based on a knowledge base (142);
- disambiguating entities based on a sentence context; and
- storing the entity that is processed by the NER with a highest score.
6. The method according to claim 5, the method further comprising:
- 25 defining a weighted vote to each NER;

weighting each NER's score for each entity to obtain a ranking score for each NER;

storing the results of the NER having the highest ranking score on the knowledge database.

5

7. The method according to claim 5, wherein tokenizing the sentence further comprising:

tokenizing the sentences;

parsing the tokenized sentences to identify noun phrases;

10 identifying acronym and abbreviations through an entity resolution module (148); and

identifying noun phrases and compound words through lexical processor (118).

8. The method according to claim 5, the method further comprising:

15 accessing the linguistic resources (132) by the text processor (101) and the entity recognition engine (102) to extract and recognize entities; and

accessing the Linked Data (130) by the knowledge-based NER (126) to process the entity,

20 wherein the linguistic resources have an abbreviation/acronym repository (134), a named-entity (136) and a lexicon database (138).

9. The method according to claim 8, the method further comprising:

searching entity patterns through a pattern matcher from the linguistic resources (132); and

25 finding and matching a target string through a term finder from the Linked Data (130) to process the target string to identify entities for processing the knowledge-based NER (126).

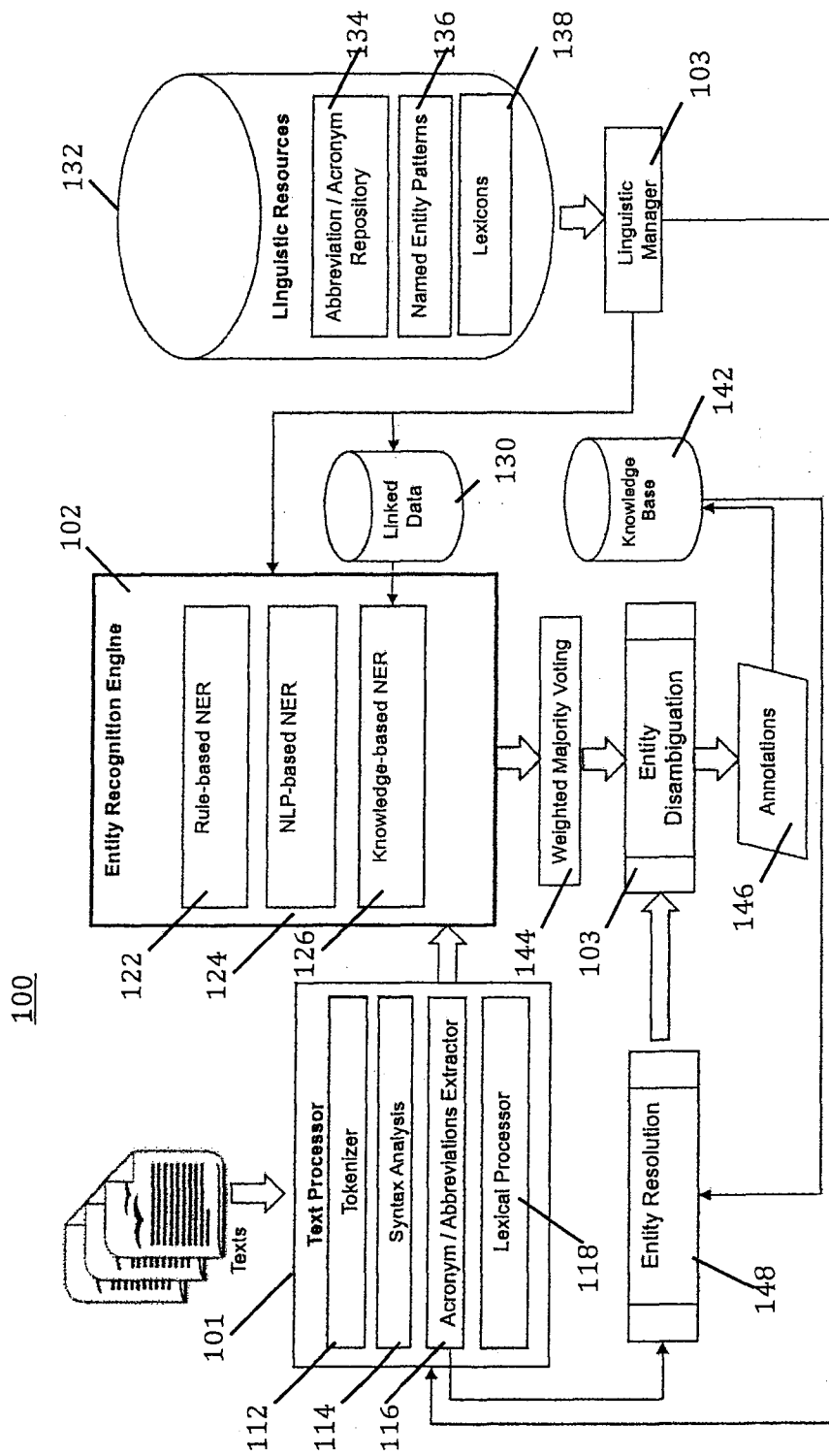


FIG. 1

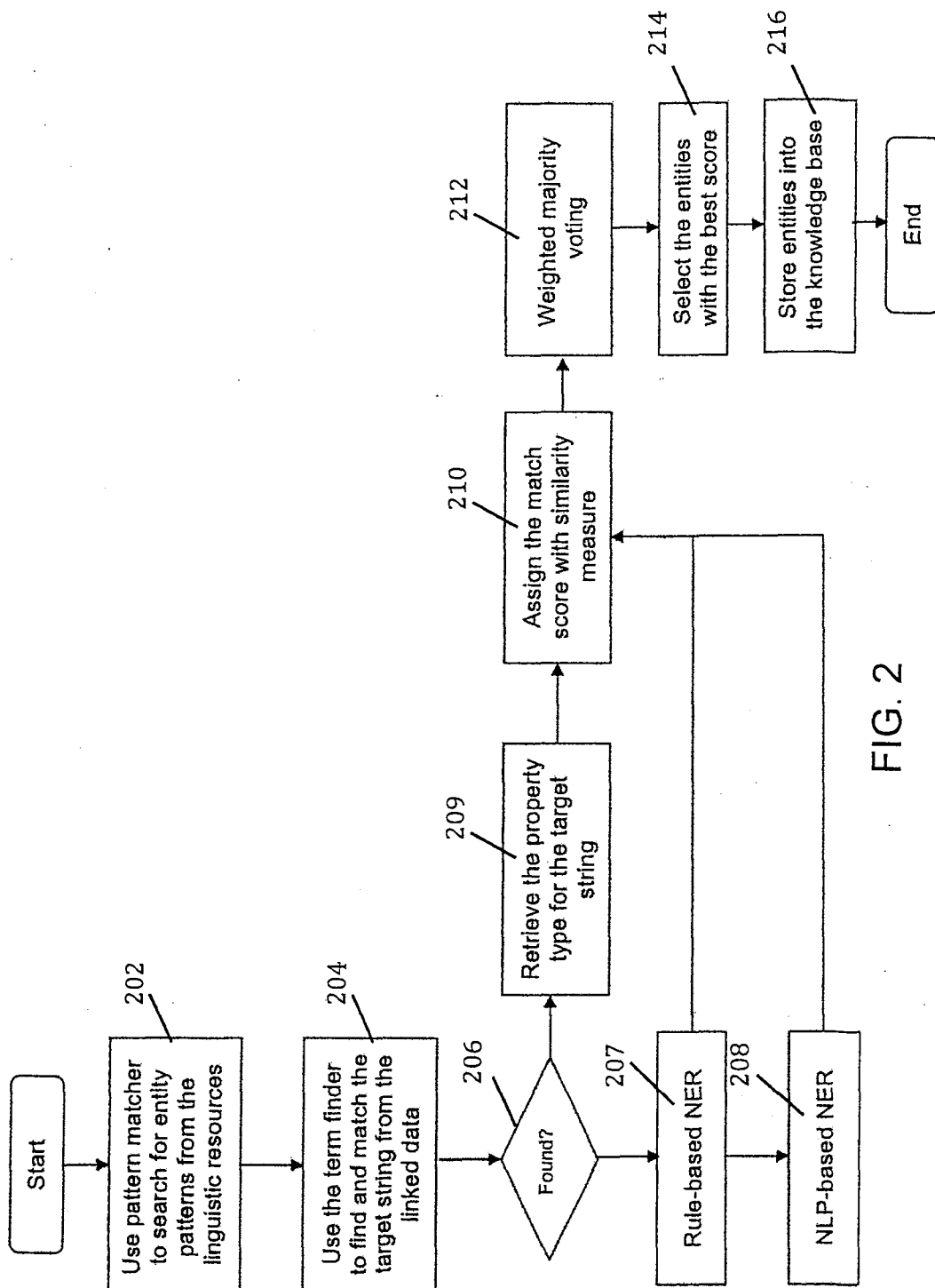


FIG. 2

Sentence:
The iPad Mini is a mini tablet computer designed, developed, and marketed by Apple Inc. which was announced on October 23, 2012.

Step (1): Parse the sentence to retrieve the phrase syntax structure

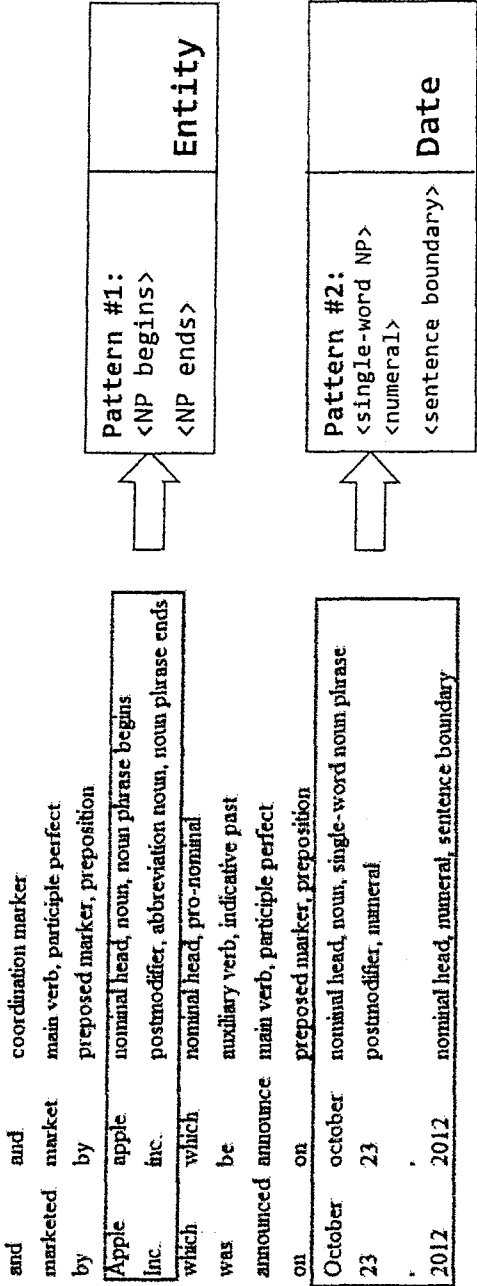
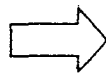


FIG. 3A

Sentence:

The iPad Mini is a mini tablet computer designed, developed, and marketed by Apple Inc. which was announced on October 23, 2012.

Step (2): Extract possible named entities (use of the pattern matcher) by matching all defined patterns from linguistic resources



iPad Mini	Entity
Apple Inc	Entity
October 23, 2012	Date

Step (3): Extract all possible acronyms / abbreviations and use entity resolution to resolve to the existing entities

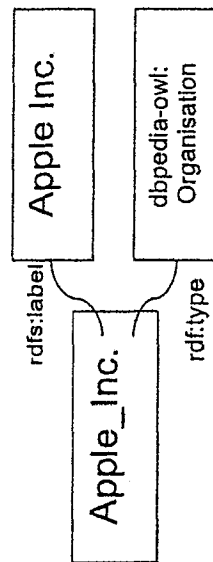
(*) In addition to the syntax structure, perform noun phrase (compound words) identification with the use of lexical processor to extract possible named entities

FIG. 3B

Sentence:

The iPad Mini is a mini tablet computer designed, developed, and marketed by Apple Inc. which was announced on October 23, 2012.

Step (4): From all the potential entities identified, use the term finder to find and match the target string from the linked data (i.e. match the label to the target string)



Step (5): Assign the match score with similarity measure

Step (6): Retrieve the property type for the target string
(For this case, the entity type will be identified as "Organisation")

FIG. 3C

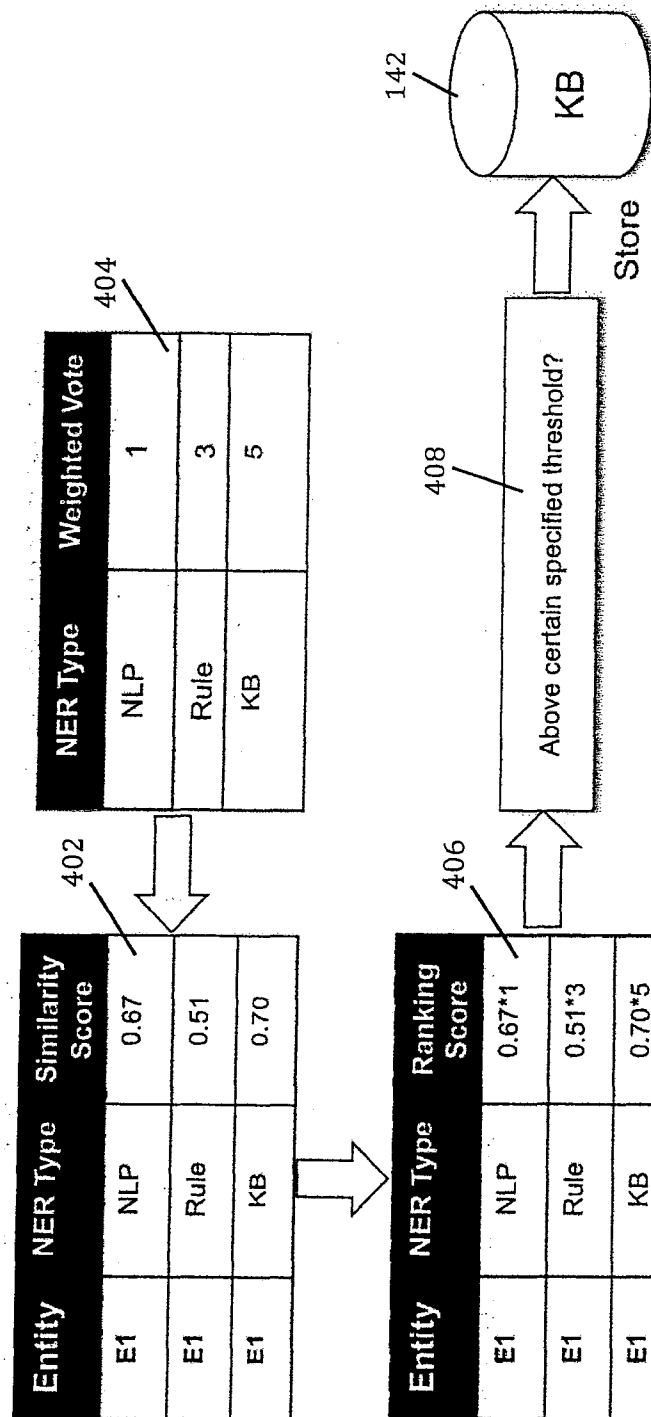


FIG. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/MY2014/000153A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/27
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	KAMEL N NEBHI: "Ontology-Based Information Extraction from Twitter", PROCEEDINGS OF THE WORKSHOP ON INFORMATION EXTRACTION AND ENTITY ANALYTICS ON SOCIAL MEDIA DATA COLING 2012, 9 December 2012 (2012-12-09), pages 17-22, XP055169825, Mumbai, India abstract Sections 1, 2 ----- -/--	1-9



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

18 February 2015

Date of mailing of the international search report

25/02/2015

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Woods, Justin

INTERNATIONAL SEARCH REPORT

International application No
PCT/MY2014/000153

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2009/052277 A1 (EVRI INC [US]; LIANG JISHENG [US]; KOPERSKI KRZYSZTOF [US]; DHILLON NA) 23 April 2009 (2009-04-23) abstract page 6, line 12 - page 15, line 4; figures 3,4 page 22, line 27 - page 24, line 5 page 30, line 22 - page 32, line 4 -----	1,2,5,6
Y	US 2013/158979 A1 (JAKUBIK PAUL A [US]) 20 June 2013 (2013-06-20) abstract paragraphs [0021] - [0036] -----	1-9
Y	NING KANG ET AL: "Using an ensemble system to improve concept extraction from clinical records", JOURNAL OF BIOMEDICAL INFORMATICS, ACADEMIC PRESS, NEW YORK, NY, US, vol. 45, no. 3, 25 December 2011 (2011-12-25), pages 423-428, XP028519146, ISSN: 1532-0464, DOI: 10.1016/J.JBI.2011.12.009 [retrieved on 2012-01-03] abstract Sections 1-3, 5 -----	2,6

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/MY2014/000153

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2009052277 A1	23-04-2009	AU 2008312487 A1	23-04-2009
		CA 2703007 A1	23-04-2009
		EP 2206057 A1	14-07-2010
		US 2009144609 A1	04-06-2009
		US 2014142922 A1	22-05-2014
		WO 2009052277 A1	23-04-2009

US 2013158979 A1	20-06-2013	US 2013158979 A1	20-06-2013
		US 2013158983 A1	20-06-2013
