(11) Application No. AU 2011202308 A1

(19) AUSTRALIAN PATENT OFFICE

(54) Title
Systems, methods, and software for assessing ambiguity of medical terms

(51)[6] International Patent Classification(s)
G06F 17/30 (2006.01) 20060101AFI2011052
G06F 17/30 3BHAU

(21) Application No: 2011202308 (22) Application Date: 2011.05.18

(43) Publication Date : 2011.06.09
(43) Publication Journal Date : 2011.06.09

(62) Divisional of:
2006302523

(71) Applicant(s)
Thomson Reuters Global Resources

(72) Inventor(s)
Dozier, Christopher C.; Kondadadi, Ravi; Chaudhary, Mark

(74) Agent/Attorney
Phillips Ormonde Fitzpatrick, 367 Collins Street, Melbourne, VIC, 3000

2011202308 18 May 2011

ABSTRACT

Some known medical terms may function as non-medical terms depending on their particular context. Accordingly, the present inventors devised systems, methods, and software that facilitate determining whether a term that is found in a medical corpus is likely to be a medical term when found in another corpus. An exemplary embodiment receives a term and computes an ambiguity score based on language models for a medical and a non-medical corpus.

<filename>

2011202308 18 May 2011

AUSTRALIA

Patents Act

COMPLETE SPECIFICATION
(ORIGINAL)

Class Int. Class

Application Number:
Lodged:

Complete Specification Lodged:
Accepted:
Published:

Priority

Related Art:

Name of Applicant:

Thomson Reuters Global Resources

Actual Inventor(s):

Christopher C. Dozier, Ravi Kondadadi, Mark Chaudhary

Address for Service and Correspondence:

PHILLIPS ORMONDE FITZPATRICK
Patent and Trade Mark Attorneys
367 Collins Street
Melbourne 3000 AUSTRALIA

Invention Title:

SYSTEMS, METHODS, AND SOFTWARE FOR ASSESSING AMBIGUITY OF MEDICAL TERMS

Our Ref : 914817
POF Code: 498584/489867

The following statement is a full description of this invention, including the best method of performing it known to applicant(s):

60086

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

Systems, Methods, and Software For Assessing Ambiguity of Medical Terms

Copyright Notice and Permission

A portion of this patent document contains material subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent files or records, but otherwise reserves all copyrights whatsoever. The following notice applies to this document: Copyright © 2005-2006, Thomson Global Resources.

Cross-Reference to Related Application

This application claims priority to U.S. provisional application 60/723,483 filed on October 4, 2005. The provisional application is incorporated herein by reference.

Technical Field

Various embodiments of the present invention concern systems, methods, and software for identifying medical content in documents and linking those documents to other documents based on the medical content.

Background

The fantastic growth of the Internet and other computer networks has fueled an equally fantastic growth in the data accessible via these networks. One of the seminal modes for interacting with this data is through the use of hyperlinks within electronic documents.

Hyperlinks are user-selectable elements, such as highlighted text or icons, that link one portion of an electronic document to another portion of the same document or to other documents in a database or computer network. With

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

proper computer equipment and network access, a user can select or invoke a hyperlink and almost instantaneously view the other document, which can be located on virtually any computer system in the world.

Although many hyperlinks are created and inserted into documents manually, recent years have seen development of automated techniques for identifying specific types of document text and linking the identified text using hyperlinks to other related documents. For example, to facilitate legal research, the Westlaw legal research system automatically identifies legal citations and attorney names in text and links the citations to corresponding legal documents in a database and the attorney names to biographical entries in an online directory. For further details, see U.S. Patent 7,003,719 and U.S. Published Patent Application US2003/0135826A1, both of which are incorporated herein by reference.

Although the automated linking technology in the Westlaw system is highly effective for legal citations and names, the present inventors have identified that this technology is not well suited for other types of content, such as medical terms. For example, the inventors recognize that identifying legal citations and entity names within a text is generally simpler than identifying medical terms because terms may function as medical terms in one context and as non-medical terms in another. Legal citations and person names, on the other hand, generally function as legal citations and person names regardless of context.

Accordingly, the present inventors have identified a need for automated methods identifying whether terms are medical terms or non-medical terms.

A reference herein to a patent document or other matter which is given as prior art is not to be taken as an admission or a suggestion that the document or matter was known or that the information it contains was part of the common general knowledge as at the priority date of any of the claims.

Summary

One aspect of the invention provides a computer-implemented method for inserting in a document a hyperlink to a medical document associated with a medical term in the document, the method comprising: receiving a term included within a document; determining an ambiguity score for the term, wherein the ambiguity score is based on first and second language models for respective medical and non-medical corpuses of documents; and determining, based on the ambiguity score, whether to insert a hyperlink in the document to a medical document associated with the term; and inserting the hyperlink in the document if the determination to insert the hyperlink is affirmative.

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

Another aspect of the invention provides a computerized system comprising: an input for receiving a set of terms; a processor for executing code adapted to determine an ambiguity score for a term from the set of terms, wherein the ambiguity score is based on first and second language models for respective medical and non-medical corpuses of documents; and means for determining, based on the ambiguity score, whether to insert a hyperlink in a first document including the term to a medical document associated with the term; and means for inserting the hyperlink in the first document if the determination to insert the hyperlink is affirmative.

Another aspect of the invention provides a computer-readable medium comprising: a code set configured to receive by a computer a term; a code set configured to determine by the computer an ambiguity score for the term, wherein the ambiguity score is based on first and second language models; and a code set configured to output by the computer the ambiguity score.

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

Brief Description of Drawings

Figure 1 is a block diagram of an exemplary system 100 which corresponds to one or more embodiment of the present invention.

Figure 2 is a flow chart of an exemplary method of operating system 100 which corresponds to one or more embodiments of the invention.

Detailed Description of Exemplary Embodiments

The following detailed description, which references and incorporates Figures 1 and 2, describes and illustrates one or more exemplary embodiments of the invention. These embodiments, offered not to limit but only to exemplify and teach the invention, are shown and described in sufficient detail to enable those skilled in the art to make and use the invention. Thus, where appropriate to avoid obscuring the invention, the description may omit certain information known to those of skill in the art.

Exemplary Computer System Embodying the Invention

Figure 1 shows a diagram of an exemplary computer system 100 incorporating a system, method, and software for assessing the ambiguity of terms, such as medical terms. Though the exemplary system is presented as an interconnected ensemble of separate components, some other embodiments implement their functionality using a greater or lesser number of components. Moreover, some embodiments intercouple one or more the components through wired or wireless local- or wide-area networks. Some embodiments implement one or more portions of system 100 using one or more mainframe computers or servers.) Thus, the present invention is not limited to any particular functional partition.

Generally, system 100 includes input terms 110, term-ambiguity calculator 120, and ambiguity scores output 130.

Input terms 110 includes one or more terms, such as a set of terms from a medical database. In the exemplary embodiment, input terms 110 includes terms from the Unified Medical Language System (UMLS). The table below shows

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

that UMLS includes a great number of terms in disease, injury, medical procedure, body part, and drug categories.

Category	Terms	Concepts
Disease	189,712	69,948
Injury	42,141	28,997
Medical procedure	134,179	72,918
Body part	38,041	22,260
Drugs	244,752	129,959

In some embodiment, input terms 110 are terms extracted from one or more input documents, such as an electronic judicial opinion, or other type legal document.

Coupled to database 110 is term-ambiguity calculator 120. Calculator 120 includes one or more conventional processors 121, display device122, interface devices 123, network-communications devices 124, and memory 125. Memory 125, which can take a variety of forms, such as coded instructions or data on an electrical, magnetic, and/or optical carrier medium, includes term-ambiguity software 126. Term-ambiguity software 126 includes various software and data components, for determining or calculating for each input term t and ambiguity score, *Score(term)* defined as

$$Score(term) = \lambda_1 \frac{\log(P(t \mid News\_lang))}{\log(P(t \mid UMLS\_lang))} + \lambda_2 \frac{\log(P(t \mid Legal\_lang))}{\log(P(t \mid UMLS\_lang))}$$

where

$$\log(P(t \mid lang)) = \sum_{i=1}^{n} \log(P(ngram \mid lang))$$

and lamda1 and lamda2 are constants, which in some embodiments are used to normalize or smooth the scoring function. In some embodiments, lambda1 and

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

lamda2 are set to 0.5. The exemplary embodiment uses ngram backoff with Witten Bell smoothing to smooth the language models.

The exemplary scoring function is based on the intuition that medical ngrams, such as "hepatic," occur relatively more often in UMLS than in news or legal and that ngrams such as "drinki" will occur relatively more often in news or legal than in UMLS. Terms having ngrams that are more highly predicted by UMLS than news or legal tend to yield a larger score and thus indicate that the given term is more likely a medical term than not a medical term when found in a news or legal document.

Term-ambiguity calculator 120 outputs a set 130 of one or more ambiguity scores based on the input terms. (Figure 1 shows that the input terms 110 and output scores 130 are also retained in memory 130.) In the exemplary embodiment, the scores are output as a ranked list, with each score associated with corresponding terms. (Note that term may include one or more words.)

The ambiguity scores can be used for a variety of purposes, including for example determining whether it is appropriate to insert a link in a document including a given term back to a ULMS document associated with the term. For example, in the output terms shown the terms having an ambiguity score greater than 1.5 may be considered as clearly being medical terms and thus linked with high confidence back to related ULMS documents. On the other hand, terms such as "word salad" or "anticipatory vomiting" that have lower scores should not generally be linked back to a related ULMS document without contextual corroboration.

Exemplary Operation of System 100

Figure 2 shows a flowchart 200 illustrating an exemplary method of operating system 100. Flow chart 200 includes process blocks 210-230. Though these blocks (and those of other flow charts in this document) are arranged serially in the exemplary embodiment, other embodiments may reorder the blocks, omit one or more blocks, and/or execute two or more blocks in parallel using multiple processors or a single processor organized as two or more virtual machines or subprocessors. Moreover, still other embodiments implement the blocks as one or more specific interconnected hardware or integrated-circuit

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

modules with related control and data signals communicated between and through the modules. Thus, this and other exemplary process flows in this document are applicable to software, firmware, hardware, and other types of implementations.

Block 210 entails receiving a set of terms. In the exemplary embodiment, this entails receiving a set of terms from ULMS or an input news or legal document into memory 126 of term-ambiguity calculator 120. Execution continues at block 220.

Block 220 entails determining one or more ambiguity scores for one or more of the input terms. In the exemplary embodiment this entails computing ambiguity scores according to the definition set forth above for *Score(term)* in equation above, which provides a sum of two conditional probability ratios. Each conditional probability is based on language model of set or corpus of documents. In some embodiments, one of the conditional probability ratios is omitted from the scoring function. Also, in some embodiments, the conditional probability ratios are inverted.

Block 230 entails outputting one or more of the determined ambiguity scores. In the exemplary embodiment, this entails outputting in printed or other human readable form; however, in other embodiments, the output may also be used by another machine, component, or software module, or simply retained in memory.

Conclusion

The embodiments described above are intended only to illustrate and teach one or more ways of practicing or implementing the present invention, not to restrict its breadth or scope. The actual scope of the invention, which embraces all ways of practicing or implementing the teachings of the invention, is defined only by the following claims and their equivalents.

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

The claims defining the invention are as follows:

1. A computer-implemented method for inserting in a document a hyperlink to a medical document associated with a medical term in the document, the method comprising:

receiving a term included within a document;

determining an ambiguity score for the term, wherein the ambiguity score is based on first and second language models for respective medical and non-medical corpuses of documents; and

determining, based on the ambiguity score, whether to insert a hyperlink in the document to a medical document associated with the term; and

inserting the hyperlink in the document if the determination to insert the hyperlink is affirmative.

2. The computer-implemented method of claim 1, wherein the second language model is based on a legal or general news corpus of documents.

3. The computer implemented method of either one of claims 1 or 2, wherein the ambiguity score is based on a ratio of a probability of the term given a non-medical corpus to a probability of the term given a medical corpus.

4. A computerized system comprising:

an input for receiving a set of terms;

a processor for executing code adapted to determine an ambiguity score for a term from the set of terms, wherein the ambiguity score is based on first and second language models for respective medical and non-medical corpuses of documents; and

means for determining, based on the ambiguity score, whether to insert a hyperlink in a first document including the term to a medical document associated with the term; and

means for inserting the hyperlink in the first document if the determination to insert the hyperlink is affirmative.

5. The system of claim 4, wherein the second language model is based on a legal or general news corpus of documents.

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

18 May 2011 2011202308

6. The system of either one of claims 4 or 5, wherein each ambiguity score is based on a ratio of a probability of the term given a non-medical corpus to a probability of the term given a medical corpus.

7. The method of any one of claims 1 to 3, wherein first language model is based on the Unified Medical Language System (UMLS).

8. The method of any one of claims 1 to 3, further comprising:
determining that the ambiguity score meets a specified level.

9. The method of any one of claims 1 to 3, further comprising:
linking the term to a related document.

10. The system of any one of claims 4 to 6, wherein first language model is based on the Unified Medical Language System (UMLS).

11. The system of any one of claims 4 to 6, further comprising:
determining that the ambiguity score meets a specified level.

12. The system of any one of claims 4 to 6, further comprising:
linking the term to a related document.

13. A computer-readable medium comprising:
a code set configured to receive by a computer a term;
a code set configured to determine by the computer an ambiguity score for the term, wherein the ambiguity score is based on first and second language models; and
a code set configured to output by the computer the ambiguity score.

14. The computer-readable medium of claim 13, wherein the first language model is based on a medical corpus of documents and the second language model is based on a legal or general news corpus of documents.

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

18 May 2011 2011202308

15. The computer-readable medium of either one of claims 13 or 14, wherein each ambiguity score is based on a ratio of a probability of the term given a non-medical corpus to a probability of the term given a medical corpus.

16. The computer-readable medium of any one of claims 13 to 15, wherein first language model is based on the Unified Medical Language System (UMLS).

17. The computer-readable medium of any one of claims 13 to 16, further comprising:
determining that the ambiguity score meets a specified level.

18. The system of any one of claims 13 to 17, further comprising:
linking the term to a related document.

19. The method of any one of claims 1 to 3 or 7 to 9, wherein the ambiguity score is based on

$$\lambda_1 \frac{\log(P(t \mid News\_lang))}{\log(P(t \mid UMLS\_lang))} + \lambda_2 \frac{\log(P(t \mid Legal\_lang))}{\log(P(t \mid UMLS\_lang))}$$

where t denotes the term; *News_lang* denotes a news corpus; *UMLS_lang* denotes a medical corpus; *Legal lang* denotes a legal corpus; λ_1 and λ_2 are constants; and

$$\log(P(t \mid lang)) = \sum_{i=1}^{n} \log(P(ngram \mid lang))$$

where *lang* is a placeholder for the corpus of interest; and n denotes the number of words in the term t.

20. The system of any one of claims 4 to 6 or 10 to 12, wherein the means for determining the ambiguity score is adapted to determine the ambiguity score is based on

$$\lambda_1 \frac{\log(P(t \mid News\_lang))}{\log(P(t \mid UMLS\_lang))} + \lambda_2 \frac{\log(P(t \mid Legal\_lang))}{\log(P(t \mid UMLS\_lang))}$$

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

where t denotes the term; *News_lang* denotes a news corpus; *UMLS_lang* denotes a medical corpus; *Legal_lang* denotes a legal corpus; λ_1 and λ_2 are constants; and

$$\log(P(t \mid lang)) = \sum_{i=1}^{n} \log(P(ngram \mid lang))$$

where *lang* is a placeholder for the corpus of interest; and n denotes the number of words in the term t.

SPEC-520071.DOC

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

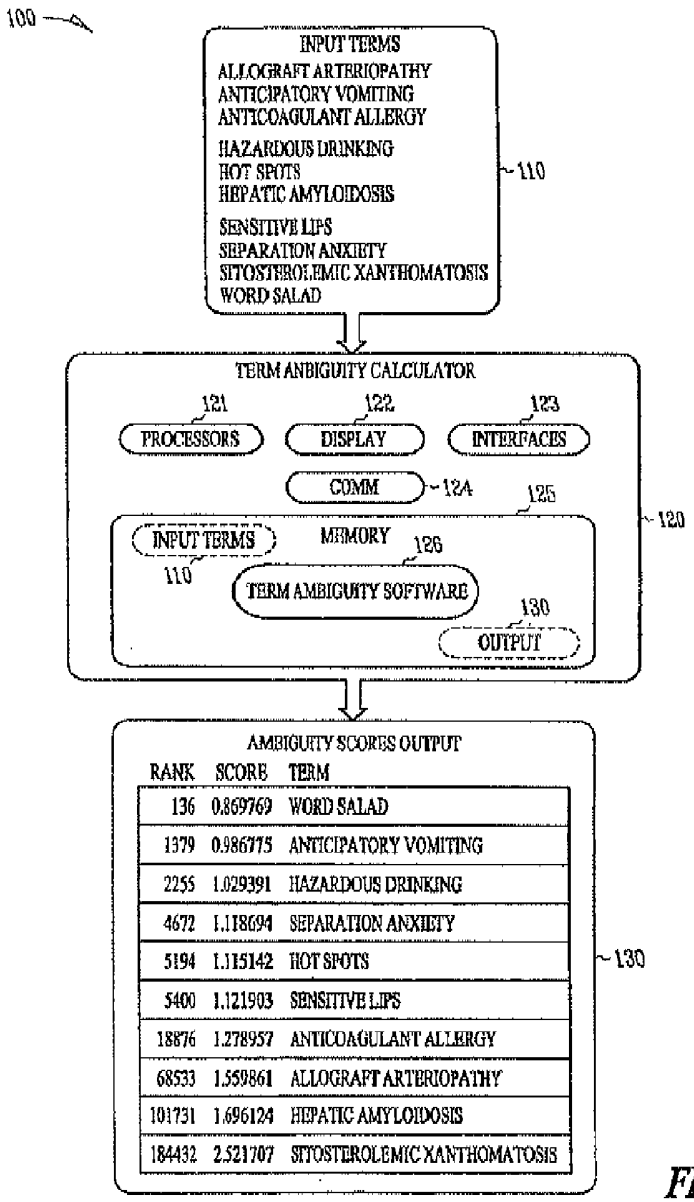

FIG. 1

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18

2011202308 18 May 2011

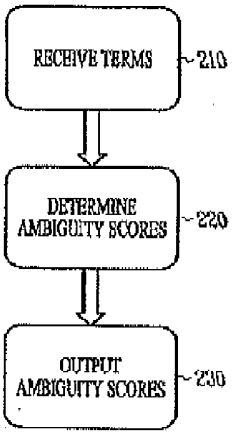

FIG. 2

COMS ID No: ARCS-321657 Received by IP Australia: Time (H:m) 16:46 Date (Y-M-d) 2011-05-18