(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-3814
(P2009-3814A)
(43)公開日　平成21年1月8日(2009.1.8)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/30	(2006.01)	G06F	17/30	180A	5B075
G06N 3/00	(2006.01)	G06F	17/30	210A	
		G06F	17/30	340Z	
		G06F	17/30	210D	
		G06N	3/00	560A	

審査請求 未請求　請求項の数 14　ＯＬ　外国語出願　（全 22 頁）

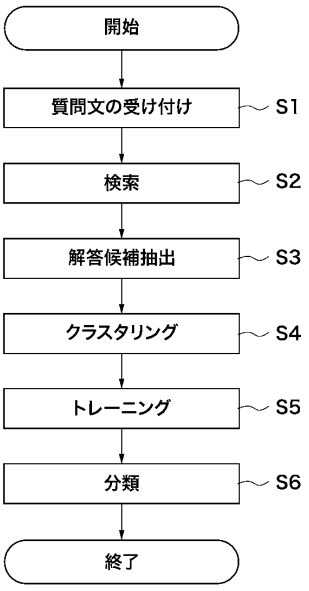
(21)出願番号　特願2007-165692 (P2007-165692)
(22)出願日　平成19年6月22日 (2007.6.22)

(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100107250
弁理士　林　信之
(74)代理人　100120868
弁理士　安彦　元
(72)発明者　呉　友政
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　柏岡　秀紀
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

Ｆターム(参考)　5B075 KK02 ND03 NK32 NR12 PP02 PP12 PP24 PQ02 PR06 UU40

(54)【発明の名称】質問応答方法及びシステム

(57)【要約】
【課題】オープンドメインな質問応答システムを実現する上で、その解答の正答率を向上させる。
【解決手段】質問文から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、検索した各ウェブページから、アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとを有する。
【選択図】図２



【特許請求の範囲】
【請求項１】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、
上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、
上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、
上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、
上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとを有すること
を特徴とする質問応答方法。
【請求項２】
上記分類ステップは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項１記載の質問応答方法。
【請求項３】
上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項１又は２記載の質問応答方法。
【請求項４】
上記ＳＢＦＳは、上記キーワードの重複度、上記キーワードの非重複度、上記キーワードのbi-gramの一致度、シーソラスの重複度、上記キーワードと上記解答候補との間の規格化距離の何れか１以上に基づくものであり、
上記ＢＭＦＳは、人名が一致しているか否か、地域名が一致しているか否か、組織名が一致しているか否か、時を示す単語が一致しているか否か、数量を示す単語が一致しているか否か、語源が一致しているか否か、解答候補が質問文のbi-gramと一致したbi-gramを有するか否か、解答候補がネームディエンティティタイプを要求されているか否かの何れか１以上に基づくものであり、
上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数
を特徴とする請求項３記載の質問応答方法。
【請求項５】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析手段と、
上記質問文解析手段により抽出されたキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索手段と、
上記検索ステップにより検索された各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出手段と、
上記解答候補抽出手段により抽出された各解答候補に基づいて、候補選択スニペットを

クラスタに割り当て、これをトレーニングデータとする、クラスタリング手段と、
上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類手段とを備えること
を特徴とする質問応答システム。
【請求項６】
上記分類手段は、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項５記載の質問応答システム。
【請求項７】
上記分類手段は、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項５又は６記載の質問応答システム。
【請求項８】
上記分類手段は、上記ＳＢＦＳを、上記キーワードの重複度、上記キーワードの非重複度、上記キーワードのbi-gramの一致度、シーソラスの重複度、上記キーワードと上記解答候補との間の規格化距離の何れか１以上に基づくものとし、
上記ＢＭＦＳを、人名が一致しているか否か、地域名が一致しているか否か、組織名が一致しているか否か、時を示す単語が一致しているか否か、数量を示す単語が一致しているか否か、語源が一致しているか否か、解答候補が質問文のbi-gramと一致したbi-gramを有するか否か、解答候補がネームディエンティティタイプを要求されているか否かの何れか１以上に基づくものとし、
上記ＷＷＦＳを、以下のＩＳＦ値により重み付けすること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数。
を特徴とする請求項７記載の質問応答システム。
【請求項９】
上記質問文解析手段と、上記検索手段と、上記解答候補抽出手段と、上記クラスタリング手段と、上記分類手段とを備える制御装置と、当該制御装置に対して通信網を介して情報を送受信可能な複数のユーザ用端末装置とを備え、
上記ユーザ用端末装置は、ユーザからの上記質問文の入力を受け付け、これを通信網を介して上記制御装置における上記質問文解析手段へと送信するとともに、当該制御装置における上記分類手段から出力される上記応答を上記通信網を介して受信し、これをユーザに表示すること
を特徴とする請求項５～８のうち何れか１項記載の質問応答システム。
【請求項１０】
ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、
上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、
上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、
上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、

上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとをコンピュータに実行させること
を特徴とするプログラム。
【請求項１１】
上記分類ステップは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項１０記載のプログラム。
【請求項１２】
上記分類ステップは、上記トレーニングデータと上記質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、上記トレーニングデータと上記質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、上記トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出すること
を特徴とする請求項１０又は１１記載のプログラム。
【請求項１３】
上記ＳＢＦＳは、上記キーワードの重複度、上記キーワードの非重複度、上記キーワードのbi-gramの一致度、シーソラスの重複度、上記キーワードと上記解答候補との間の規格化距離の何れか１以上に基づくものであり、
上記ＢＭＦＳは、人名が一致しているか否か、地域名が一致しているか否か、組織名が一致しているか否か、時を示す単語が一致しているか否か、数量を示す単語が一致しているか否か、語源が一致しているか否か、解答候補が質問文のbi-gramと一致したbi-gramを有するか否か、解答候補がネームディエンティティタイプを要求されているか否かの何れか１以上に基づくものであり、
上記ＷＷＦＳは、以下のＩＳＦ値により重み付けされていること
$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$
ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数。
を特徴とする請求項１２記載のプログラム。
【請求項１４】
請求項１０～１３のうち何れか１項記載のプログラムが記録されていることを特徴とする記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
ユーザから入力された質問文に対して応答を出力可能な質問応答方法及びシステム、プログラム並びに記録媒体に関する。
【背景技術】
【０００２】
近年におけるインターネットの普及に伴い、ユーザは、検索エンジンを利用して所望のウェブページをサーチし、そこから知見を得ることが可能となってきている。特にユーザが知りたい情報について検索エンジンを介してウェブページを検索する場合、検索クエリーとして、キーワードを入力することにより、当該キーワードに関連するスニペットを持つウェブページが自動抽出され、これを検索リストとして表示されることになる。ユーザは、かかる検索リストに表示されたウェブページ一覧から所望のウェブページにアクセスし、知りたい情報を取得することが可能となる。
【０００３】
ところで、現在におけるウェブページの検索方法では、検索リストに表示されたウェブ

ページ一覧から、所望の情報が記載されている、真のウェブページをユーザ自身が順次アクセスしながら見つけ出す必要があり、労力の負担が増大し、また検索に要する時間が長期化してしまうという問題点があった。

【０００４】

このため、このような検索エンジンを介して検索リストを表示する代替として、ユーザ自身が知りたい情報を自然な文章として端末を介して入力し、かかる質問文に対する応答を直接出力する質問応答システムが従来から望まれていた。このため、かかる質問応答システムに関する研究も従来より行われていた。

【０００５】

従来の質問応答システムは、４つのカテゴリーに分類することができる。

【０００６】

先ず、質問文の全てのキーワードと、解答候補との間で類似性を示す距離を求め、これに基づいて解答候補から正解を選び出すモデルが提案されている。しかし、このモデルでは、質問と解答候補が依拠する文章とが表面上一致していないだけで、正解を出すことができなくなるという問題点があった。

【０００７】

また、質問文を最初に予め定義したカテゴリーに分類し、これをオフラインの下で学習したアンサーパターンを利用して正解を抽出するモデルも提案されている。しかし、このモデルは、予め定義した何種類かの質問のタイプに対しては高い正確性を出すことが可能であるが、オープンドメインな質問応答のための質問のタイプを定義するのが困難であり、あらゆるタイプの質問に対して対応することができない。

【０００８】

自然言語処理(ＮＬＰ：Natural Language Processing)に基づくモデルは、ユーザの質問を解析し、応答に相当する文を意味的な表現へと繋げ、そして意味的にマッチングするものを解答として見つけ出すものである。このモデルは、ＴＲＥＣ（Text REtrieval Conference）のワークショップにおいてよく実演されるものであるが、ＮＬＰツールの高パフォーマンスに大きく依存するものである。このため、処理時間が長時間に亘るとともに、作業量の増加が無視できない。

【０００９】

さらに、マシーンラーニングに基づくモデルも研究されている（例えば、特許文献１参照。）が、未だその有用性は確立されていない。

【００１０】

即ち、これらの開示技術は、手入力された質問と解答のペアが所定量必要になるという問題点がある。また、マシーンラーニング技術に特化した上記ペアを収集するのは多大な労力が必要になり、コスト増につながる。

【非特許文献１】Jun Suzuki,Yutaka Sasaki,Eisaku Maeda.SVM Answer Selection for Open-Domain Question Answering,In Proc.of Coling-2002,pp974～980(2002).

【発明の開示】

【発明が解決しようとする課題】

【００１１】

そこで、本発明は、上述した問題点に鑑みて案出されたものであり、オープンドメインな質問応答システムを実現する上で、その解答の正答率を向上させることが可能な質問応答方法及びシステム、プログラム並びに記録媒体を提供することにある。

【課題を解決するための手段】

【００１２】

本発明を適用した質問応答方法は、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候

補抽出ステップと、上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとを有することを特徴とする。

【００１３】

本発明を適用した質問応答システムは、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析手段と、上記質問文解析手段により抽出されたキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索手段と、上記検索ステップにより検索された各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出手段と、上記解答候補抽出手段により抽出された各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリング手段と、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類手段とを備えることを特徴とする。

【００１４】

本発明を適用したプログラムは、ユーザから入力された質問文を構成する単語から、キーワードと、質問の種別を規定するアンサータイプとを特定する質問文解析ステップと、上記質問文解析ステップにおいて抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する検索ステップと、上記検索ステップにおいて検索した各ウェブページから、上記アンサータイプに基づいて解答候補を順次抽出する解答候補抽出ステップと、上記解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする、クラスタリングステップと、上記トレーニングデータを解析することにより上記クラスタを分類し、更に上記トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する分類ステップとをコンピュータに実行させることを特徴とする。

【発明の効果】

【００１５】

上述した構成からなる本発明では、後述する表１～３の結果から示されるように、オープンドメインな質問応答システムを実現する上で、その解答の正答率を向上させることが可能となる。

【発明を実施するための最良の形態】

【００１６】

以下、本発明を実施するための最良の形態として、ユーザから入力された質問文に対して応答を出力可能な質問応答システムに監視、図面を参照しながら詳細に説明をする。

【００１７】

本発明を適用した質問応答システム１は、図１に示すように、質問文を入力するユーザにより操作されるユーザ端末装置１１と、このユーザ端末装置１１により通信網１２を介してそれぞれアクセス可能なウェブサーバ１３と、このウェブサーバ１３を制御するための制御装置１４とを備えている。

【００１８】

ユーザ端末装置１１は、例えばパーソナルコンピュータ（ＰＣ）等が適用され、質問文を入力するためのマウスやキーボード等からなる操作部と、情報をユーザに対して表示するための、例えば液晶ディスプレイからなる表示部を備える。このユーザ端末装置１１は、ユーザからの質問文の入力を受け付けた場合に、これを通信網１２を介して制御装置１４へと送信する。

【００１９】
通信網１２は、例えばウェブサーバ１３とユーザ端末装置１１とを電話回線を介して接続されるインターネット網を始め、ＴＡ／モデムと接続されるＩＳＤＮ（Integrated Services Digital Network）／Ｂ（broadband）－ＩＳＤＮ等のように、情報の双方向送受信を可能とした公衆通信網等である。
【００２０】
また制御装置１４も同様にコンピュータで構成されるものであり、相互にバスで接続されたＣＰＵ(Central Processing Unit)や、メモリ、固定ディスクと、通信網１２を介してユーザ端末装置１１との間で情報を送受信するための通信インターフェースとを備えている。実際に、本発明に係る質問応答システム１を実行するためのプログラムは、この制御装置１４における固定ディスク等にインストールされることになる。また、このプログラムは、他のＣＤ－ＲＯＭ等に記録された記録媒体として具体化することも可能となる。
【００２１】
制御装置１４は、通信網１２を介してユーザ端末装置１１から受信した質問文を受けて、ウェブサーバ１３へアクセスし、後述するような処理を実行することにより、上記質問文に対する応答を作り出し、通信網１２を介してユーザ端末装置１１へと送信する。ユーザ端末装置１１は、送られてきた応答を液晶ディスプレイからなる表示部を介して表示する。
【００２２】
次に、本発明を適用した質問応答システム１の動作について説明をする。
【００２３】
図２は、質問応答システム１を実行する上でのフローチャートを示している。先ずステップＳ１において、ユーザからの質問文の入力を受け付ける。ちなみに、本発明は、オープンドメインの質問応答の実現を想定しているところ、ユーザは、言語や入力形式に支配されることなく、自然に質問したい内容を文章にし、これを入力していくことになる。このため、ユーザの入力すべき内容について、複雑なルールは特段存在せず、また高精度な言語解析技術も特段必要としない。
【００２４】
この質問文は、文書検索のクエリーとなりえるキーワードの集合と、質問の種別を規程するアンサータイプから構成されることになる。例えば、「いつ潜水艦が沈んだか？」という質問文が入力された場合においてキーワードは、「潜水艦」、「沈む」に相当し、アンサータイプは、「いつ」に相当するものとなる。即ち、このキーワードは、質問文中に含まれる名詞、動詞等を規程する単語であり、アンサータイプは、時、場所、主体、数量等、実際にユーザが知りたいカテゴリを示すものである。また、ステップＳ２は、質問からキーワードを抽出し、また疑問詞に基づいて質問の解答タイプを分類する。
【００２５】
次に、ステップＳ２へ移行し、上記ステップＳ１において抽出したキーワードを検索クエリーとし、当該キーワードに関連するウェブページを検索する。その結果、このステップＳ２においては、キーワードに関係するウェブページが順次検索されてくることになる。このステップＳ２においては、例えば検索エンジンにおけるスニペットの記載に基づいて検索を行うようにしてもよい。ちなみに、このステップＳ２において、上述した例の質問文が入力された場合には、「潜水艦」、「沈む」というキーワードに関係するウェブページが順次検索されてくることになる。
【００２６】
次にステップＳ３へ移行し、ステップＳ２において検索した各ウェブページから、アンサータイプに基づいて解答候補を順次抽出する。即ち、アンサータイプとして、時、場所、主体、数量等の何れかがステップＳ１において抽出されているため、これに関係する解答候補を抽出してくることになる。このステップＳ３においては、例えば検索エンジンにおけるスニペットの記載から解答候補を抽出するようにしてもよい。ちなみに、このステップＳ３において、上述した例の質問文が入力された場合には、「いつ」に相当する時を

表すアンサータイプに基づく解答候補を抽出してくることになる。

【００２７】

次にステップＳ４へ移行し、クラスタリングを行う。このクラスタリングは、解答候補抽出ステップにおいて抽出した各解答候補に基づいて、候補選択スニペットをクラスタに割り当て、これをトレーニングデータとする。同一の解答候補を含むウェブスニペットは、同系列のスニペットとみなし、これを同一のクラスタに属させる。そして、この割り当ての結果を分類ステップのトレーニングデータとしてとして利用する。

【００２８】

即ち、検索エンジンのトップm位までのスニペット $\{s_1,s_2,\cdots,s_m\}$ からｎ個の解答候補 $\{c_1,c_2,\cdots,c_n\}$ を抽出する。これらのスニペットは、それぞれ解答候補 $\{c_i\}$ と少なくとも１の質問キーワード $\{q_i\}$ を保有している。そして、これらスニペット $\{s_1,s_2,\cdots,s_m\}$ は、ウェブサーチ結果のクラスタリングにより、ｎ個のクラスタ $\{C_1,C_2,\cdots,C_n\}$ へと割り当てられることになる。

【００２９】

仮にスニペットがＬ個の異なる解答候補を保有するものであれば、そのスニペットはＬ個の異なるクラスタに割り当てられることになる。また、異なるスニペットの解答候補が互いに同一であれば、これらのスニペットは同一のクラスタに割り当てられることになる。

【００３０】

最終的に、クラスタ $\{C_i\}$ は、解答候補 $\{c_i\}$ の数によって決定される。そしてクラスタC_iのクラスタ名は、解答候補c_iに基づくものとなる。これらクラスタ化された解答候補c_iが上述したトレーニングデータとなる。

【００３１】

次に、ステップＳ５へ移行し、トレーニングデータを解析することにより上記クラスタを分類する。このトレーニングステップＳ５は、トレーニングベクトルの構築や、ＳＶＭ(Support Vector Machine)による分類のために、トレーニングデータから３つのタイプを抽出する。

【００３２】

さらに次にステップＳ６へ移行し、トレーニングデータの解析と同一解析条件の下で上記質問文を解析し、当該質問文の解析結果と最も類似するクラスタを順に抽出し、これを応答として出力する。この分類ステップＳ６は、ユーザの質問から分析したテストベクトルを利用することにより、クラスタの一つに割り当て、質問に対する解答を、質問のクラスタの名称と仮定する。

【００３３】

以下、このステップＳ５とステップＳ６を一つの分類ステップとして説明をしていく。この分類ステップでは、ＳＶＭ(Support Vector Machine)を利用することにより、質問文の解析結果と最も類似するクラスタを順に抽出するようにしてもよい。

【００３４】

また、この分類ステップでは、トレーニングデータと質問文との単語重複度を示すＳＢＦＳ（similarity-based feature set）、トレーニングデータと質問文とのブーリアン重複度を示すＢＭＦＳ(Boolean match-based feature set)、トレーニングデータを構成する解答候補の前後を構成する文字を含めた文字列と上記質問文との類似度を示すＷＷＦＳ(window-based word feature set)に基づいて、当該質問文の解析結果と最も類似するクラスタを順に抽出するようにしてもよい。

【００３５】

ＳＢＦＳとしては、キーワードの重複度、キーワードの非重複度、キーワードのbi-gramの一致度、シーソラスの重複度、キーワードと解答候補との間の規格化距離の何れか１以上に基づくものであってもよい。

【００３６】

ＢＭＦＳは、人名が一致しているか否か、地域名が一致しているか否か、組織名が一致

しているか否か、時を示す単語が一致しているか否か、数量を示す単語が一致しているか否か、語源が一致しているか否か、解答候補が質問文のbi-gramと一致したbi-gramを有するか否か、解答候補がネームディエンティティタイプを要求されているか否かの何れか１以上に基づくものであってもよい。

【００３７】

ＷＷＦＳは、以下のＩＳＦ値により重み付けされていてもよい。

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

ここで、$N(w_j)$は、単語w_jが含まれているウェブページのスニペットの総数、$N(w_j,C_i)$は、単語w_jが含まれているクラスタC_iにおけるスニペットの数である。

【００３８】

最後にステップＳ７に移行し、実際に上記プロセスの下で得た回答をユーザ端末装置１１の表示部を介して表示する。

【００３９】

次に、本発明を用いた質問応答システム１による効果について説明をする。

【００４０】

中国語によるウェブの質問応答（QA）における３種類のデータセットとしてのCTREC04、CTREC05、CTEST05を利用し、本発明を適用したU(unsupervised)-SVMを検証することとした。CTREC04は、TREC2004FACTOIDのテスト問題から翻訳された１７８個の中国語による質問のセットからなる。CTREC05は、TRE2005FACTOIDのテスト問題から翻訳された２７９個の中国語による質問のセットからなる。CTEST05は、中国語で記載されたものを除く、TERCのテスト問題に類似する１７８個の中国語の質問のセットからなる。

【００４１】

実験は、３つの評価項目、即ち、top_1,top_5,mrr_5に基づいて評価を行った。評価結果を表１、２、３に示す。ここで、top_1は、解答の正確性がトップ１位である解答が含まれている割合を示している。top_5は、解答の正確性がトップ５位以内である解答の一つが含まれている割合を示している。mrr_5は、各質問に対する正解の平均相対ランク（１/ｎ）を示しており、ここで最高ランクｎ（ｎ≦５）としている。

【００４２】

各データセット（CTREC04、CTREC05、CTEST05）をU-SVMを用いて解析することにより得られた各評価項目（top_1,top_5,mrr_5）を表１に示す。

【００４３】

【表１】

	TREC04	TREC05	Test05
top_1	60.82%	57.33%	59.09%
mrr_5	71.31%	65.61%	67.34%
top_5	88.66%	80.00%	81.82%

【００４４】

また、表２において、CTrec04と、CTrec05のテストデータをU-SVMとthe Retrieval-M（従来の検索手法）それぞれを用いて解析することにより得られた各評価項目（top_1,top_5,mrr_5）の相対比較を表２に示す。

【００４５】

【表２】

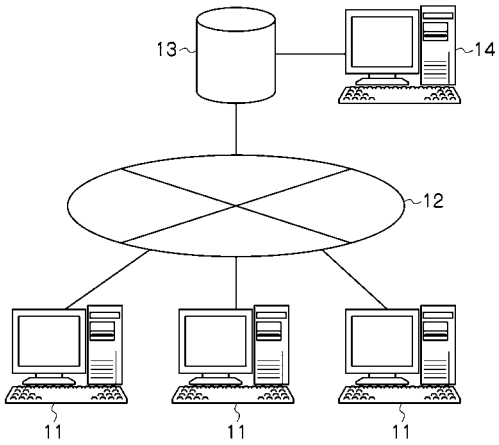
		Retrieval-M	U-SVM
Ctrec04	top_1	27.84%	53.61%
	mrr_5	43.67%	66.25%
	top_5	71.13%	88.66%
Ctrec05	top_1	34.00%	50.00%
	mrr_5	48.20%	62.38%
	top_5	71.33%	82.67%

【００４６】
さらに、Pattern-M（パターン重視による手法）とS-SVM（SVMを利用した教化学習法によるもの）に対するU-SVMのパフォーマンス性を比較するために、CTEST05のデータセットを用いて検証を行った。表３は、U-SVM、Pattern-M、S-SVMの各モデルを用いてCTEST05を解析することにより得られた各評価項目（top_1,mrr_5）の相対比較を示している。
【００４７】
【表３】

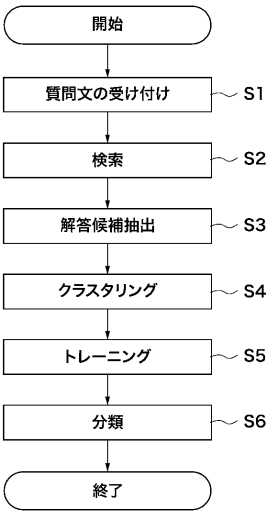
	S-SVM	Pattern-M	U-SVM
top_1	44.89%	53.14%	59.09%
mrr_5	56.49%	61.28%	67.34%
top_5	74.43%	73.14%	81.82%

【００４８】
上述した表１～３の結果から、各モデルによる正答率（パフォーマンスランキング）は、U-SVM＞Pattern-M＞S-SVM＞Retrieval-Mの順となった。
【００４９】
即ち、本発明では、上述した図２に示すフローに基づいて、解答を抽出していくため、正答率を向上させることが可能となる。
【図面の簡単な説明】
【００５０】
【図１】本発明を適用した質問応答システムの構成例を示す図である。
【図２】本発明を適用した質問応答システムの処理手順を示すフローチャートである。
【符号の説明】
【００５１】
１　質問応答システム
１１　ユーザ端末装置
１２　通信網
１３　ウェブサーバ
１４　制御装置

【図１】


13
14
12
11
11
11

【図２】


開始
質問文の受け付け
S1
検索
S2
解答候補抽出
S3
クラスタリング
S4
トレーニング
S5
分類
S6
終了

【外国語明細書】

QUESTION ANSWERING METHOD AND SYSTEM

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

The present invention relates to a question answering method and system capable of outputting an answer to a question sentence input by a user, a program, and a recording medium therefore.

DESCRIPTION OF THE RELATED ART

The recent prevalence of the Internet allows users to search desired web pages to get desired information using a search engine. In searching web pages for desired information using a search engine, particularly, when a user inputs keywords as a query, web pages having snippets relevant to the keywords are automatically extracted and are displayed as a retrieval list. The user can access a desired web site selected from a list of web pages shown in the retrieval list to thereby acquire desired information.

The web site search methods available at present require that a user should find out a proper web site containing desired information by sequentially accessing web pages listed in the list of web pages shown in a retrieval list. This increases the user's burden and takes a longer time for the search.

Accordingly, there is a demand of a question answering system which allows a user to input desired information in the form of a natural language sentence and directly outputs an answer to the question sentence instead of displaying a retrieval list through a search engine. In this respect, researches on such question answering systems have been made conventionally.

The conventional question answering systems can be classified into four categories. First, the retrieval-based model selects a correct answer from candidate answers based on the distance indicating a similarity between a candidate and all keywords in a question sentence. This model cannot provide the correct answer, however, if the question and answer-bearing sentences do not match on the surface.

The pattern-based model classifies a question sentence into predefined categories, and then extracts the correct answer by using answer patterns learned offline. While the pattern-based model can obtain high precision for some predefined types of questions, the model has a difficulty in defining question types in advance for open-domain question answering. That is, this model is not suitable for all types of questions.

The deep NLP (Natural Language Processing)-based model usually parses a user question to transform each sentence equivalent to an answer into a semantic representation, and then provides a semantically matched sentence as the answer. This model, which has performed very well as TREC (Text REtrieval Conference) workshops, heavily depends on high-performance NLP tools. This leads to time consuming processing and a non-negligible intensive labor.

Finally, the machine learning-based model has also been studied (see Jun Suzuki, Yutaka Sasaki, Eisaku Maeda. SVM Answer Selection for Open-Domain Question Answering, In Proc. of Coling-2002, pp 974-980 (2002)), but its availability has not been established yet.

These techniques suffer, however, from the problem of requiring an adequate number of hand-tagged question-answer training pairs. It is too expensive and labor intensive to collect such training pairs for supervised machine learning techniques.

SUMMARY OF THE INVENTION

Accordingly, the present invention has been worked out to cope with the situations, and it is an object of the present invention to provide a question answering method and system capable of improving the performance ranking of an answer, and a program and a recording medium therefor.

A question answering method according to the present invention includes a question parse step of specifying a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user; a retrieval step of retrieving web pages relevant to the keyword extracted in the question parse step using the keyword as a query; a candidate answer extraction step of sequentially extracting candidate answers based on the answer type from the web pages retrieved in the retrieval step; a clustering step of assigning the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data; and a classification step of classifying the clusters by analyzing the training data, parsing the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracting clusters having a highest similarity to a result of parsing the question sentence, and using the extracted clusters as an answer.

A question answering system according to the present invention includes question parse means that specifies a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user; retrieval means that retrieves web pages relevant to the keyword extracted by the question parse means using the keyword as a query; candidate answer extraction means that sequentially extracts candidate answers based on the answer type from the web pages retrieved by the retrieval means; clustering means that assigns the candidate-bearing snippets into the clusters according to the candidate answers extracted by the candidate answer extraction means, and uses a result of assignment as training data; and classification means that classifies the clusters by analyzing the training data, parses the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracts clusters having a highest similarity to a result of parsing the question sentence, and uses the extracted clusters as an answer.

A program according to the present invention allows a computer to execute a question parse step of specifying a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user; a retrieval step of retrieving web pages relevant to the keyword extracted in the question parse step using the keyword as a query; a candidate answer extraction step of sequentially extracting candidate answers based on the answer type from the web pages retrieved in the retrieval step; a clustering step of assigning the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data; and a classification step of classifying the clusters by analyzing the training data, parsing the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracting clusters having a highest similarity to a result of parsing the question sentence, and using the extracted clusters as an answer.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram showing an example of the configuration of a question answering system to which the present invention is adapted; and

Fig. 2 is a flowchart illustrating process procedures of the question answering system according to the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

A question answering system capable of outputting an answer to a question sentence input by a user as a preferred embodiment of the present invention will now be described in detail referring to the accompanying drawings.

As shown in Fig. 1, a question answering system 1 according to the invention includes a user terminal device 11 each of which is operated by a user who inputs a question sentence, a web server 13 which can be accessed by the user terminal devices 11 over a communication network 12, and a control apparatus 14 which controls the web server 13.

A personal computer (PC) or the like, for example, is used as the user terminal device 11. The user terminal device 11 has an operation section including a mouse and a keyboard or the like for inputting a question sentence, and a display section, which is, for example, a liquid crystal display, to present information to the user. Upon reception of a question sentence input by the user, the user terminal device 11 sends the question sentence to the control apparatus 14 over the communication network 12.

The communication network 12 is the Internet to which the web server 13 and the user terminal devices 11 are connected by telephone circuits, as well as public communication networks like ISDN (Integrated Services Digital Network)/B (Broadband)-ISDN which is connected to a TA/modem.

The control apparatus 14 is likewise constituted by a computer, and has a CPU (Central Processing Unit), a memory, a fixed disk, and a communication interface. The CPU, the memory and the fixed disk are connected to one another by a bus. The communication interface ensures transmission and reception of information to and from the user terminal devices 11 over the communication network 12. Actually, a program for achieving the question answering system 1 according to the invention is installed in the fixed disk or the like in the control apparatus 14. The program can be realized as one recorded in another recording medium such as CD-ROM.

Upon reception of a question sentence from the user terminal device 11 over the communication network 12, the control apparatus 14 accesses the web server 13, executes processes (to be described later) to create an answer to the question sentence, and sends the answer to the user terminal device 11 over the communication network 12. The user terminal device 11 displays the sent answer on the display section comprised of a liquid crystal display.

The operation of the question answering system 1 according to the invention will be described next.

Fig. 2 shows a flowchart for achieving the question answering system 1. First, a question sentence input by a user is accepted in step S1. Because the invention is intended to realize open-domain question answering, a user naturally writes the contents of a question into a sentence without being dominated by the language and the input format, and inputs the sentence. This eliminates the need for particularly complicated rules for the contents to be input by the user, or a high-precision special language analysis technique.

The question sentence consists of a set of keywords to be a query for searching documents, and an answer type which defines the type of the question.

When a question sentence "when did the submarine sink?" is input, for example, keywords are "submarine" and "sink" while the answer type is equivalent to "when". That is, the keywords are words that define nouns, verbs, etc. contained in the question sentence, and the answer type indicates a category, such as time, place, subject or quantity, the user actually wants to know. And So this step S2 is extracting the keywords from the question and identifying the answer type of the question according to the interrogative words.

Next, the flow proceeds to step S2 where with the keywords extracted in the step S1 being a query, web pages relevant to the keywords are retrieved. As a result, web pages relevant to the keywords are sequentially retrieved in the step S1. In the step S2, retrieval may be executed based on, for example, the descriptions of snippets of the search engine. When the aforementioned question sentence is input, web pages relevant to the keywords "submarine" and "sink" are sequentially retrieved in the step S2.

The flow then proceeds to step S3 where candidate answers are sequentially retrieved from the individual web pages retrieved in the step S2 based on the answer type. That is, as one of time, place, subject, quantity, etc. has been extracted as the answer type in the step S1, candidate answers relevant to the answer type are to be extracted. In the step S3, candidate answers may be extracted from, for example, the descriptions of snippets of the search engine. When the aforementioned question sentence is input, candidate answers are extracted based on the answer type representing the time equivalent to "when".

Then, the flow proceeds to step S4 to perform clustering. The clustering assigns the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data. The web snippets containing the same candidate answer are regarded as aligned snippets, and thus belong to the same cluster,and using the result of assignment as the training data of the classification step.

Specifically, n candidate answers $\{c_1, c_2, \ldots, c_n\}$ are extracted from top m snippets $\{s_1, s_2, \ldots, s_m\}$ of the search engine. Each of those snippets has candidate answers $\{c_i\}$ and at least one question keyword $\{q_i\}$. The snippets $\{s_1, s_2, \ldots, s_m\}$ are assigned to n clusters $\{C_1, C_2, \ldots, C_n\}$ by clustering of the web search results.

If a snippet has L different candidate answers, the snippet is assigned to L different clusters. If candidate answers of different snippets are identical, those snippets are assigned to the same cluster.

Finally, the cluster $\{C_i\}$ is determined by the number of candidate answers $\{c_i\}$. The cluster name of the cluster C_i is based on the candidate answers c_i. The clustered candidate answers c_i become the training data.

Next, the flow proceeds to step S15 where the training data is analyzed to classify the clusters. The training step S5 is extracting three types of features from the training data to construct the training vectors, and thus to train a SVM classifier.

The flow then proceeds to step S6 to analyze the training data and parse the question sentence under the same analysis condition as used in analyzing the training data, sequentially extract clusters which are most likely the result of parsing the question sentence, and use the extracted clusters as an answer. The

classification step S6 is classifying the user's question into one of the clusters by using the test vector which is parsed from the user question, and assuming the name of the question's cluster as the answer to the question.

In the step S5 and S6, clusters which are most likely the result of parsing the question sentence may be extracted sequentially by using the SVM (Support Vector Machine).

In the step S5 and S6, clusters which are most likely the result of parsing the question sentence may be extracted sequentially based on an SBFS (similarity-based feature set) indicating a word overlap between the training data and the question sentence, a BMFS (Boolean match-based feature set) indicating Boolean matches between the training data and the question sentence, and a WWFS (window-based word feature set) indicating a similarity between a string of characters including characters preceding and following a candidate answer constituting the training data and the question sentence.

The SBFS may be based on at least one of a percentage of matched keywords, a percentage of mismatched keywords, a percentage of matched bi-grams of keywords, a percentage of matched thesauruses, and a normalized distance between candidate and keywords.

The BMFS may be based on at least one of whether person names are matched or not, location names are matched or not, organization names are matched or not, time words are matched or not, number words are matched or not, a root verb is matched or not, a candidate has or does not have bi-gram in snippet matching bi-gram in question, and a candidate has or does not have a desired named entity type.

The WWFS may be weighted with the following ISF value.

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

where $N(w_j)$ is the total number of snippets in a web site containing a word feature wj, and $N(w_j,C_i)$ is the number of snippets in a cluster C_i containing the word feature w_j.

Finally, the U-SVM display the answer actually obtained in the process on the display section of the user terminal device 11.

The effects of the question answering system 1 of the invention will be described.

In the invention, the U-SVM was validated in terms of Chinese web question answering (QA) using three test data sets, CTREC04, CTREC05 and CTEST05. CTREC04 is a set of 178 Chinese questions translated from TREC 2004 FACTOID testing questions. CTREC05 is a set of 279 Chinese questions translated from TREC 2005 FACTOID testing questions. CTEST05 is a set of 178 Chinese questions that are similar to TREC testing questions except that they are written in Chinese.

The experimental results were evaluated in terms of three scores, top_1, top_5 and mrr_5. The top_1 indicates the rate at which the correct answer is included in the top 1 answer. The top_5 indicates the rate at which at least one correct answer is included in the top 1 answer. The mrr_5 indicates the average reciprocal rank (1/n) of the highest rank n (n5) of a correct answer to each question.

Table 1 shows the evaluation items (top_1, top_5, mrr_5) acquired by analyzing the individual data sets (CTREC04, CTREC05, CTEST05) using the U-SVM.

TABLE.1

	TREC04	TREC05	Test05
top_1	60.82%	57.33%	59.09%
mrr_5	71.31%	65.61%	67.34%
top_5	88.66%	80.00%	81.82%

Table 2 shows the comparative performances of the evaluation items (top_1, top_5, mrr_5) acquired by analyzing CTrec04 and CTrec05 test data using the U-SVM and the Retrieval-M.

TABLE.2

		Retrieval-M	U-SVM
Ctrec04	top_1	27.84%	53.61%
	mrr_5	43.67%	66.25%
	top_5	71.13%	88.66%
Ctrec05	top_1	34.00%	50.00%
	mrr_5	48.20%	62.38%
	top_5	71.33%	82.67%

To compare performances of the U-SVM with the Pattern-M and the S-SVM, evaluation was carried out using the CTEST05 data set. Table 3 shows the comparative performances of the evaluation items (top_1, mrr_5) acquired by analyzing CTEST05 using the U-SVM, Pattern-M and S-SVM models.

TABLE.3

	S-SVM	Pattern-M	U-SVM
top_1	44.89%	53.14%	59.09%
mrr_5	56.49%	61.28%	67.34%
top_5	74.43%	73.14%	81.82%

From the results shown in Tables 1 to 3, it follows that the performance ranking of the individual models was: U-SVM > Pattern-M > S-SVM > Retrieval-M.

In other words, the invention can improve the performance ranking for answers are extracted based on the flow shown in Fig. 2.

What is claimed is:

1. A question answering method comprising:

a question parse step of specifying a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user;

a retrieval step of retrieving web pages relevant to the keyword extracted in the question parse step using the keyword as a query;

a candidate answer extraction step of sequentially extracting candidate answers based on the answer type from the web pages retrieved in the retrieval step;

a clustering step of assigning the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data; and

a classification step of classifying the clusters by analyzing the training data, parsing the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracting clusters having a highest similarity to a result of parsing the question sentence, and using the extracted clusters as an answer.

2. The question answering method according to claim 1, wherein the classification step sequentially extracts clusters having a highest similarity to the result of parsing the question sentence by using a SVM (Support Vector Machine).

3. The question answering method according to claim 1 or 2, wherein the classification step sequentially extracts clusters having a highest similarity to the result of parsing the question sentence based on an SBFS (similarity-based feature set) indicating a word overlap between the training data and the question sentence, a BMFS (Boolean match-based feature set) indicating Boolean matches between the training data and the question sentence, and a WWFS (window-based word feature set) indicating a similarity between a string of characters including characters preceding and following a candidate answer constituting the training data and the question sentence.

4. The question answering method according to claim 3, wherein the SBFS is

based on at least one of a percentage of matched keywords, a percentage of mismatched keywords, a percentage of matched bi-grams of keywords, a percentage of matched thesauruses, and a normalized distance between candidate and keywords, the BMFS is based on at least one of whether person names are matched or not, location names are matched or not, organization names are matched or not ,time words are matched or not, number words are matched or not, a root verb is matched or not, a candidate has or does not have bi-gram in snippet matching bi-gram in question, and a candidate has or does not have a desired named entity type, and

the WWFS is weighted with an ISF value given by

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

where $N(w_j)$ is the total number of snippets in a web site containing a word feature w_j, and $N(w_j,C_i)$ is the number of snippets in a cluster C_i containing the word feature w_j.

5. A question answering system comprising:

question parse means that specifies a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user;

retrieval means that retrieves web pages relevant to the keyword extracted by the question parse means using the keyword as a query;

candidate answer extraction means that sequentially extracts candidate answers based on the answer type from the web pages retrieved by the retrieval means;

clustering means that assigns the candidate-bearing snippets into the clusters according to the candidate answers extracted by the candidate answer extraction means, and uses a result of assignment as training data; and

classification means that classifies the clusters by analyzing the training data, parses the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracts clusters having a highest similarity to a result of parsing the question sentence, and uses the extracted clusters as an answer.

6. The question answering system according to claim 5, wherein the classification means sequentially extracts clusters having a highest similarity to the result of parsing the question sentence by using a SVM (Support Vector Machine).

7. The question answering system according to claim 5 or 6, wherein the classification means sequentially extracts clusters having a highest similarity to the result of parsing the question sentence based on an SBFS (similarity-based feature set) indicating a word overlap between the training data and the question sentence, a BMFS (Boolean match-based feature set) indicating Boolean matches between the training data and the question sentence, and a WWFS (window-based word feature set) indicating a similarity between a string of characters including characters preceding and following a candidate answer constituting the training data and the question sentence.

8. The question answering system according to claim 7, wherein the classification means sets the SBFS based on at least one of a percentage of matched keywords, a percentage of mismatched keywords, a percentage of matched bi-grams of keywords, a percentage of matched thesauruses, and a normalized distance between candidate and keywords,

sets the BMFS based on at least one of whether person names are matched or

not, location names are matched or not, organization names are matched or not, time words are matched or not, number words are matched or not, a root verb is matched or not, a candidate has or does not have bi-gram in snippet matching bi-gram in question, and a candidate has or does not have a desired named entity type, and

weights the WWFS with an ISF value given by

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

where $N(w_j)$ is the total number of snippets in a web site containing a word feature w_j, and $N(w_j,C_i)$ is the number of snippets in a cluster C_i containing the word feature w_j.

9. The question answering system according to any one of claims 5 to 8, further comprising a control apparatus including the question parse means, the retrieval means, the candidate answer extraction means, the clustering means, and the classification means, and a plurality of user terminal devices capable of transmitting and receiving information to and from the control apparatus over a communication network,

wherein each of the user terminal devices accepts the question sentence input by the user, transmits the question sentence to the question parse means in the control apparatus over the communication network, receives the answer output from the classification means in the control apparatus over the communication network, and presents the answer to the user.

10. A program that allows a computer to execute:

a question parse step of specifying a keyword and an answer type which defines a type of a question from words constituting a question sentence input by a user;

a retrieval step of retrieving web pages relevant to the keyword extracted in the question parse step using the keyword as a query;

a candidate answer extraction step of sequentially extracting candidate answers based on the answer type from the web pages retrieved in the retrieval step;

a clustering step of assigning the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data; and

a classification step of classifying the clusters by analyzing the training data, parsing the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracting clusters having a highest similarity to a result of parsing the question sentence, and using the extracted clusters as an answer.

11. The program according to claim 10, wherein the classification step sequentially extracts clusters having a highest similarity to the result of parsing the question sentence by using a SVM (Support Vector Machine).

12. The program according to claim 10 or 11, wherein the classification step sequentially extracts clusters having a highest similarity to the result of parsing the question sentence based on an SBFS (similarity-based feature set) indicating a word overlap between the training data and the question sentence, a BMFS (Boolean match-based feature set) indicating Boolean matches between the training data and the question sentence, and a WWFS (window-based word feature set) indicating a similarity between a string of characters including characters preceding and following a candidate answer constituting the training data and the question sentence.

13. The program according to claim 12, wherein the SBFS is based on at least one of a percentage of matched keywords, a percentage of mismatched keywords, a percentage of matched bi-grams of keywords, a percentage of matched the sauruses, and a normalized distance between candidate and keywords,

the BMFS is based on at least one of whether person names are matched or not, location names are matched or not, organization names are matched or not, time words are matched or not, number words are matched or not, a root verb is matched or not, a candidate has or does not have bi-gram in snippet matching bi-gram in question, and a candidate has or does not have a desired named entity type, and

the WWFS is weighted with an ISF value given by

$ISF(w_j,C_i)=(N(w_j,C_i)+0.5)/(N(w_j)+0.5)$

where $N(w_j)$ is the total number of snippets in a web site containing a word feature w_j, and $N(w_j,C_i)$ is the number of snippets in a cluster C_i containing the word feature w_j.

14. A recording medium recording the program as set forth in any one of claims 10 to 13.

ABSTRACT OF THE DISCLOSURE

Disclosed is an open-domain question answering method and system which improve the performance ranking of an answer. The method includes a question parse step of specifying a keyword and an answer type which defines the type of a question, a retrieval step of retrieving web pages relevant to the extracted keyword using the keyword as a query, a candidate answer extraction step of sequentially extracting candidate answers based on the answer type, a clustering step of assigning the candidate-bearing snippets into the clusters according to the candidate answers extracted in the candidate answer extraction step, and using a result of assignment as training data, and a classification step of classifying the clusters by analyzing the training data, parsing the question sentence under a same analysis condition as used in analyzing the training data, sequentially extracting clusters having a highest similarity to the result of parsing the question sentence, and using the extracted clusters as an answer.

FIG.1

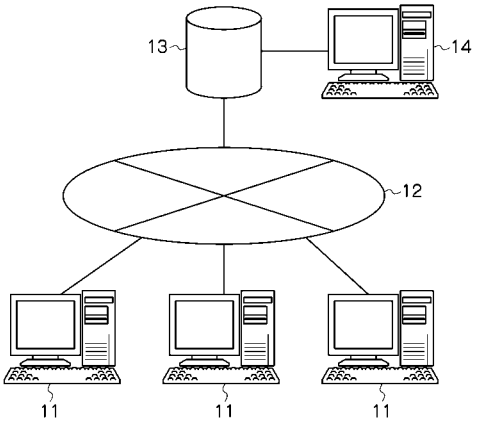
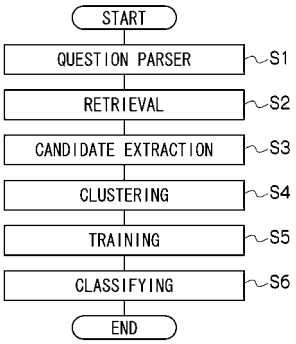

13
14
12
11
11
11

FIG.2


START
QUESTION PARSER
S1
RETRIEVAL
S2
CANDIDATE EXTRACTION
S3
CLUSTERING
S4
TRAINING
S5
CLASSIFYING
S6
END