

(19) 日本国特許庁(JP)

(12) 公表特許公報(A)

(11) 特許出願公表番号

特表2005-535007

(P2005-535007A)

(43) 公表日 平成17年11月17日(2005.11.17)

(51) Int. Cl. <sup>7</sup>	F I	テーマコード (参考)
<b>G06F 17/30</b>	G06F 17/30	220Z 5B075
<b>G06F 17/27</b>	G06F 17/30	170A 5B091
<b>G06F 17/28</b>	G06F 17/27	Z
	G06F 17/28	X

審査請求 有 予備審査請求 有 (全 54 頁)

(21) 出願番号	特願2004-508041 (P2004-508041)	(71) 出願人	504431854
(86) (22) 出願日	平成14年5月28日 (2002. 5. 28)		ナシブニイ、ウラジミル・ウラジミロビッチ
(85) 翻訳文提出日	平成17年1月21日 (2005. 1. 21)		ロシア国、115573モスクワ、ウーエル・シビロフスカヤ、デー44/27、カーペー75
(86) 国際出願番号	PCT/RU2002/000258		
(87) 国際公開番号	W02003/100659	(71) 出願人	504431865
(87) 国際公開日	平成15年12月4日 (2003. 12. 4)		ナシブナヤ、ガリナ・アナトリエバ
(81) 指定国	AP (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), EA (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), EP (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OA (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW	(74) 代理人	100081352 弁理士 広瀬 章一

最終頁に続く

(54) 【発明の名称】 文書検索システム用の知識抽出のための自己学習システムの合成方法

(57) 【要約】

本発明はインターネットに基づくデータ検索システムの開発に使用することができる。本発明は、知識の自動的形成と異なる種々の言語での電子形態で提示されたテキスト系文書からの該知識の抽出およびテキスト系データとユーザー・リクエストの知的プロセッシングを可能にする。本発明の方法は、確率的に指標付けされたインテリジェンスシステムの形態で文法および意味解析のルールを含んだシステムのための自己学習メカニズムを用意し、確率的に指標付けされた辞書のデータベースおよび言語テキストのインデックス表を形成し、テキスト系文書の解析および確率的指標付けを行い、対応する知識を形成することからなる。確率的に指標付けされたユーザー・リクエストを複数の新たなリクエストに変換し、変換されたリクエストのワード・グループを含んでいるテキスト系文書の断片を選択する。該断片を確率的に指標付けされた意味構造と、それに基づくシステムの短い応答を形成するのに使用する。受領した短い応答のリクエストに対する適合性は、それに基づいて問い合わせ文章を形成し、該問い合わせ文章をリクエストと比較することによりチェックされる。

## 【特許請求の範囲】

## 【請求項 1】

下記工程を含む、検索システムに使用するためテキスト系文書からの或る所定の自然言語での知識抽出用の自己学習システムの合成方法：

確率的情報インデックスの2進信号の単一組み合わせの適用に基づき、確率的に指標付けされた人工知能システムの形態の自己学習メカニズムを用意し；

確率的に指標付けされたテキスト断片類の等価変換と論理的結論とを使用し、かつ該断片類からリンクされた意味構造を形成し、それらの意味構造をプロダクションルールの形態での表示のために確率的に指標付けすることによって、文法および意味解析のルールに対してシステムを自動的に命令し；

10

該言語での電子形態の言語文書の形態素解析および確率的指標付けを、形態素解析ルールに対するシステムへの自動命令と同時に行い；

該言語での、あるテーマに関係する、電子形態のテキスト系文書の形態素および統語解析、ならびに確率的指標付けを、統語解析ルールに対するシステムへの自動命令と同時に行い；

その所定テーマに関係する、電子形態の確率的に指標付けされたテキスト系文書の意味解析を、意味解析ルールに対するシステムへの自動命令と同時に行い；

所定の自然言語でユーザーの1つのリクエストを形成し、それを1つの問い合わせ文章として確率的指標付けした後に、それを電子形態に変換し；

確率的に指標付けされた形態のユーザーのリクエストを、該ユーザーのリクエストに等価の、1セットの新たな複数のリクエストに変換し；

20

該新たな複数リクエストの全部のワード・コンビネーションを含む、確率的に指標付けされた電子形態のテキスト系文書の断片類を、ユーザー・リクエストに基づいて予備選択し；

該確率的に指標付けされたテキスト系文書の断片類から1つの確率的に指標付けされた意味構造を生成させ；

この意味構造に基づいて、確率的に指標付けされたテキスト系文書の断片間のリンクを与える論理的結論とテキストの等価変換とにより、システムから簡単な応答を生成させ；

該簡単な応答から問い合わせ文章を生成させ、生成した問い合わせ文章をユーザーのリクエストと比較することにより、該簡単な応答のユーザー・リクエストとの適合性をチェックする；

30

ここで、生成した問い合わせ文章がユーザーのリクエストと同一である場合には、該簡単な応答のユーザー・リクエストとの適合性を認め、該簡単な応答を所定の自然言語でユーザーに提示する。

## 【請求項 2】

下記工程を含む、検索システムに使用するための、テキスト系文書からの任意の所定の自然言語での知識抽出用の自己学習システムの合成方法：

確率的指標付けのために確率的情報インデックスの2進信号の単一組み合わせを適用することと、文法および意味解析手順の記述を含むある基本言語での言語テキスト断片の探索とに基づき、確率的に指標付けされた人工知能システムの形態の自己学習メカニズムを用意し、そして確率的に指標付けされた言語テキスト断片類の等価変換と論理的結論とを使用し、かつ該断片類からリンクされた複数の意味構造を形成し、該意味構造をプロダクションルールの形態での表示のために確率的指標付けすることによって、文法および意味解析のルールに対してシステムを自動的に命令し；

40

その所定基本言語での電子形態の言語文書の形態素解析および確率的指標付けを、同時に形態素解析ルールに対してシステムを自動命令しながら行い、所定の各外国言語についての確率的に指標付けされた辞書類のデータベースおよび言語テキスト・インデックスの表と、基本言語および所定の各外国言語についてのプロダクションルールを含む形態素解析の知識ベースとを構築し；

検索システムから、所定の各外国言語での、あるテーマについての、電子形態のテキス

50

ト系文書の形態素および統語解析、ならびに確率的指標付けを行って、該文書をテキスト系文書のインデックスの表として表示し、かつ該文書を確率的に指標付けされたテキストのベースに格納し、その間に同時に、基本言語での確率的に指標付けされた言語テキストを用いて統語解析ルールに対してシステムの自動命令を行い、そして基本言語および所定の各外国言語についての統語解析の知識ベースを構築し；

その所定テーマについて、該電子形態の確率的に指標付けされたテキスト系文書の意味解析を、意味解析ルールに対するシステムへの自動命令と同時に行って、基本言語および所定の各外国言語についての意味解析の知識ベースを構築し；

ある自然外国言語でユーザーの1つのリクエストを形成し、それを、1つの疑問語ワード・コンビネーションとユーザーのリクエストの意味を決めるワード・コンビネーションとを含んだ1つの問い合わせ文章として確率的指標付けした後に、それを電子形態に変換し；

確率的に指標付けされた形態のユーザーのリクエストを、該ユーザーのリクエストに等価の、1セットの新たな複数のリクエストに変換し；

該新たな複数リクエストの全部のワード・コンビネーションを含む、確率的に指標付けされた電子形態のテキスト系文書の断片類の、ユーザー・リクエストに基づく予備選択を行い；

該確率的に指標付けされたテキスト系文書の断片類から1つの確率的に指標付けされた意味構造を生成させ；

この意味構造に基づいて、確率的に指標付けされたテキスト系文書の断片間のリンクを与える論理的結論とテキストの等価変換とにより、システムから簡単な応答を生成させ、この応答は、ユーザー・リクエストの意味を規定する確率的に指標付けされたワード・コンビネーションと、ユーザー・リクエストの疑問語ワード・コンビネーションに対応する1つの応答ワード・グループ、とを含んでおり；

応答ワード・グループを、対応する確率的に指標付けされた疑問語ワード・コンビネーションで置換し、生成した問い合わせ文章をユーザー・リクエストと比較することにより、該簡単な応答のユーザー・リクエストとの適合性をチェックする；

ここで、生成した問い合わせ文章がユーザーのリクエストと同一である場合には、該簡単な応答のユーザー・リクエストとの適合性を認め、該簡単な応答を所定の外国言語でユーザーに提示する。

#### 【請求項3】

ユーザーのリクエストに同一の問い合わせ文章を生成することに失敗した場合に、検索システムから、ユーザーのリクエストに適合する応答を探索するための新たなテキスト系文書をリクエストすることをさらに含む、請求項1または2に記載の方法。

#### 【請求項4】

ユーザーのリクエストにより、確率的に指標付けされた意味構造を形成するための論理的結論、ならびに該簡単な応答のより詳しい内容を提供する新たな確率的に指標付けされたテキストを得るための該テキスト系文書断片の必要な等価変換によって、より詳しい情報または特定の知識を含む完全な応答を生成させることをさらに含む、請求項1～3のいずれか1項に記載の方法。

#### 【請求項5】

形態素解析ルールに対してシステムを自動命令する工程が、確率的に指標付けされたテキストにおいて、複数の語のそれぞれの語形の予め決められたセットを選択し、語幹、ならびにその語尾、接頭辞、接尾辞および前置詞の予め決められた1セット、の確率的インデックスを付与し、該インデックスに従って確率的に指標付けされた言語テキストにランダムにアクセスして、それから、該語尾、接頭辞、接尾辞および前置詞のセットを、ある語に対応するスピーチ・パートならびに語の語形変化もしくは活用変化から生ずる語尾、接頭辞、接尾辞および前置詞の完全セットに関連づける、いくつかの断片を選択し、該断片を確率的インデックスによりプロダクションルールの形態に変換し、ここで、各ルールの正しさは、対応する言語テキストからのいくつかの断片に基づく自律導出により得られ

10

20

30

40

50

、そして形態素解析の知識ベースに対するプロダクションルールのインデックスの表を得る、ことを含む、請求項 1 または 2 に記載の方法。

【請求項 6】

形態素解析の知識ベースのルールを用いて各語のスピーチ・パートを決定した後に、言語テキストの確率的指標付けを行う工程が、確率的に指標付けされた辞書のデータベースに、各語の語幹の確率的インデックスならびにその語尾、接頭辞、接尾辞および前置詞の完全セットの確率的インデックスを充填することを含む、請求項 2 ~ 5 のいずれか 1 項に記載の方法。

【請求項 7】

テキスト・インデックスの表を構築する工程が、情報の確率的変換と、語幹、それらの語尾、接頭辞、接尾辞、前置詞、文章、段落およびテキスト題名のインデックスの 2 進信号の単一組み合わせを生成させ、それらのインデックスは確率的に指標付けされたテキストのベースのインデックスの表内に置かれ、該インデックス間のリンクを付与し、このリンクが原テキスト内に明示されて、インデックスの表を用いたテキスト・リカバリを保証することを含む、請求項 2 ~ 6 のいずれか 1 項に記載の方法。

10

【請求項 8】

統語解析のルールに対してシステムを自動命令する工程が、確率的に指標付けされた言語テキスト内で、文章の統語解析の手順を記述する断片を探索し；統語要素および構造と語の予め決められたスピーチ・パートとの間のリンクを規定する確率的に指標付けされた意味構造を得るように論理的結論を取得し；形態素的な語の特徴に関して文章の統語解析を明示するプロダクションルールを導出し、ここで、各ルールの正しさは、対応する言語テキストからのいくつかの断片に基づく自律導出により得られ、生成したルールを統語解析の知識ベースに格納し、確率的に指標付けしてインデックス表の形態で表示することを含む、請求項 1 または 2 に記載の方法。

20

【請求項 9】

意味解析のルールに対してシステムを自動命令する工程が、語幹およびスピーチ・パートの確率的インデックスを参照して、言語テキストのインデックスの表への、正確に規定されていない文章メンバーのリクエストを形成して、特定の文章メンバーと一致させるためにその語が有する意味的特徴を記述するテキスト断片として応答を取得し；そして、該応答に従って、ある所定の語の語幹の確率的インデックスと要求された意味的特徴を用いて、汎用辞書もしくは専門辞書ならびに百科辞典のインデックスの表を参照し；そして、論理的結論によって、その所定の語と要求された意味的特徴とをリンクする確率的に指標付けされた意味構造を特定することを試み；その試みが成功したなら、該文章メンバーが正確に決まったと決定し；リクエストに適合するテキスト断片をプロダクションルールに変換し、ここで、各ルールの正しさは、対応する言語テキストからのいくつかの断片に基づく自律導出により得られ、該ルールを意味解析の知識ベースに格納し、確率的に指標付けして、文章メンバーとしての語の意味解析およびワード・コンピネーション間のリンクに使用されるインデックスの表の形態で表示することをさらに含む、請求項 1 または 2 に記載の方法。

30

【請求項 10】

各テキストのインデックスの表を生成させて、該テキストを形態素的、統語的および意味的に解析し終わった後、文章のそれぞれの中の各語に対応するスピーチ・パートの名称、文章メンバーおよびそれらへの質問の確率的インデックスを生成させ、該インデックスを該テキストのインデックスの表にエントリして、テキスト断片の探索において、語のそれぞれが何のスピーチ・パートおよび文章メンバーに属しているかの自動的な決定を与え、かつ該語に対する質問を示すことをさらに含む、請求項 2 ~ 9 のいずれか 1 項に記載の方法。

40

【請求項 11】

テキストのインデックスの全部の表を生成させた後に、あるテーマについてのインデックスの表を生成させ、この表の行は、語幹の繰り返されない確率的インデックスにより指

50

定され、各列は、特定のテキストの確率的インデックスに対応し；そして特定の語幹インデックスを持つある語を含むテキスト段落の確率的インデックスを該表にエントリし、その所定テーマについてのインデックスの表を、ユーザー・リクエストの予め決められたワード・コンビネーションのセットを含む断片の予備探索に使用することをさらに含む、請求項 2 ~ 10 のいずれか 1 項に記載の方法。

【請求項 12】

ユーザーのリクエストを等価変換する工程は、ほぼ同じ意味を持つ語である類義語を使用し、形態素、統語および意味の各解析の確率的に指標付けされたルールに基づいて、ユーザー・リクエストの問い合わせ文章のワード・コンビネーションの等価構造を得ると共に、それらの間の意味的關係を維持するように、ユーザー・リクエストの意味を保存しながらのスピーチ・パートおよび文章メンバーの置換を行うことを含む、請求項 1 ~ 11 のいずれか 1 項に記載の方法。

10

【請求項 13】

ユーザー・リクエストの全部のワード・コンビネーションを含む意味的にリンクされたテキスト断片を生成させる工程が、該語幹の確率的インデックスに従って、所定テーマに関してテキスト・インデックスの表を参照し、ユーザー・リクエストの全部のワード・コンビネーションを含む段落および対応するテキストの確率的インデックスを選択し、該インデックスに従って、選択されたテキストのそれぞれのインデックスの表を参照し；インデックスの表およびテキストの等価変換に基づいて論理的結論を作成して、ユーザー・リクエストの疑問語ワード・コンビネーション、ならびにユーザー・リクエストの意味を規定し、予備選択された段落が含んでいるユーザー・リクエストの全部のワード・コンビネーションに対応する、応答のワード・グループのインデックスをリンクする確率的に指標付けされた 1 つの意味構造を生成させることを含む、請求項 1 ~ 12 のいずれか 1 項に記載の方法。

20

【請求項 14】

論理的結論により首尾よく生じた、ユーザー・リクエストに対応する確率的に指標付けされた意味構造を、得られたテキスト断片のセットを用いて、ユーザー・リクエストと同一の問い合わせ文章を生成させるための基礎として使用し；そして該問い合わせ文章の生成を、ユーザー・リクエストのテキスト系文書断片の各ワード・コンビネーションの要求された意味的特徴を与えるための該知識ベースからのルールに基づく語幹ならびに語尾、接尾辞、接頭辞および前置詞の確率的インデックスの等価変換により、かつ複数のワード・コンビネーション間の推移的關係に基づく論理的結論を使用して、それらを、ユーザー・リクエストの疑問語ワード・コンビネーションに対応する該応答のワード・グループを含んでいる、ユーザー・リクエストに同一の問い合わせ文章に組み合わせることによって行うことをさらに含む、請求項 13 に記載の方法。

30

【請求項 15】

簡単な応答の正しさは、異なる複数の予備選択した確率的に指標付けされたテキスト系文書断片に基づいて、該応答の確率的に指標付けされた同一の意味構造がいくつか生成することにより保証される、請求項 1 ~ 14 のいずれか 1 項に記載の方法。

【請求項 16】

テキスト系文書のインデックスの表を用いた探索プロセスおよび応答の生成中に、リクエストとそれに適合する簡単な応答とをリンクする指標付けされたテキスト系要素を生成させて「リクエスト-リプライ」型の要素を含む知識ベースを生じさせることにより、システムの自己学習を行い、この知識ベースは、確率的指標付けされた後にインデックス表の形態で提示されて、テキストの文章の文法および意味解析ならびに該指標付けされた知識ベース中に含まれる繰り返されたリクエストに対する応答の生成のために使用されることをさらに含む、請求項 1 ~ 15 のいずれか 1 項に記載の方法。

40

【請求項 17】

簡単な応答に基づき、かつテキスト断片を得る時に使用されたインデックスの表に従って論理的結論を用いてユーザー・リクエストに適合する知識を含んでいる完全な応答を生

50

成させる工程が、応答のワード・グループを文章の語幹の確率的インデックスにリンクする確率的に指標付けされた1つの意味構造を生成させることを含み、そしてこのリンクは、テキスト断片内の簡単な応答の完全な開示を与える推移的關係を保持して、該確率的に指標付けされた意味構造に基づく文章の等価変換を用いて完全な応答のリンクされたテキストが得られる、請求項4～16のいずれか1項に記載の方法。

【請求項18】

確率的に指標付けされた断片の等価変換が、各文章を確率的に指標付けされた複数のワード・コンビネーションの1セットとして提示し、該コンビネーションを、共通語根の語幹、語尾、接頭辞、接尾辞および前置詞の確率的インデックスの等価変換によって、形態素、統語および意味の各解析の知識ベースに格納されたルールを用いて変換して、新たなスピーチ・パートおよび文章メンバーを生じさせることを含み、但し、各文章の確率的に指標付けされた意味構造におけるワード・コンビネーション間のリンクの無矛盾性および新たなテキスト断片を生成させた時の文章間の一致性が条件である、請求項1～17に記載の方法。

10

【請求項19】

確率的に指標付けされた語の辞書中または言語テキスト中に含まれていない新しい語が、テキスト系文書の確率的指標付けのプロセス中に指標付けされたテキスト中に出現した時に、辞書内でのその新語に対する共通語根の語の検索と、形態素解析の知識ベース内での該共通語根の語の新語への等価変換のルールの検索とを行い；等価変換のタイプにより、その新語が属するスピーチ・パートならびに語形変化もしくは活用変化により生ずるその語形の全部を決定し、

20

そして、辞書内に共通語根の語が全く見つからない場合には、テキストから、その新語の特定の語形の1セットを選択し、該語形の語尾、接尾辞および接頭辞に基づいて、確率的に指標付けされた辞書または形態素解析のプロダクションルールを用いて、該新語が属するスピーチ・パートならびに語形変化もしくは活用変化により生ずるその語形の完全セットを決定する、ことをさらに含む、請求項1～18のいずれか1項に記載の方法。

【請求項20】

複数の所定外国言語中のテキスト系文書からの知識の同時抽出を行うことをさらに含み、該同時抽出が、

所定の1つの基本言語に関して形態素、統語および意味の各解析のルール内でのシステムの自動命令を行い；

30

ある基本言語で、確率的に指標付けされた辞書のデータベースと確率的に指標付けされた言語テキストを用いた形態素、統語および意味の各解析の知識ベースとを構築し；

該知識ベースを用いて、所定外国言語のいずれかの自動命令のためのリクエストを自動生成させ；

該リクエストに従って、該外国言語の学習に必要な知識を含んでいる、基本言語の言語テキスト断片を予備選択し；

該テキストの等価変換を行い；

確率的に指標付けされた意味構造を生成させ、該構造に対する論理的結論を作成して、自動生成させたリクエストに適合する応答を生成させ；

40

該応答を所定外国言語のいずれかの形態素、統語および意味の各解析の知識ベースを生成させるために使用し、所定の外国言語のテキスト系文書からの知識の抽出を保証する、ことを含む、請求項2～19のいずれか1項に記載の方法。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、コンピュータサイエンス、情報探索およびインテリジェントシステムに関する。

本発明は、インターネットに基づいて動作する情報探索ならびに他の情報およびインテリジェントシステムの開発に好適に使用することができる。

50

## 【背景技術】

## 【0002】

インターネットは、非常に多くの主題およびトピックに関する巨大な量の恒久的に更新される情報を現在蓄積している。しかし、何百万ものユーザー人口によるインターネットへのアクセスは複雑化している。その原因は、検索システムにおけるデータ検索のための現状技術の効率が不十分なことにある。既知方法として、Yandex, Yahoo, Rambler検索システム用のデータ検索方法がある。これらの既知方法は、インターネットユーザがリクエストしたテキスト系文書(textual documents)をアウトプットする。

## 【0003】

公知のデータ検索方法の主な欠点は次の通りである：

- リクエストが公式化される言語が複雑である；
- テキスト系文書を意味解析し、要求された質問とのそれらの適合性(コンフォーマンス)を確かめるためのメカニズムが欠如している；
- ユーザーのリクエストに示された情報が探索文書中に存在することを正確に決定することが不可能で、かつ多量の情報源からユーザーが必要とする特定の情報および知識を抽出することが不可能である。

## 【0004】

上述した欠点のために、情報探索の手順を実施する時に、有用な情報と一緒に多くの重複した「ノイズ」情報も出力され、既存の検索システムではそれらを容易にフィルターで取り除くことができない。これは、必要な情報を探索するのに要する時間をかなり増大させ、不必要な情報の伝達と処理のために検索システムのチャネルおよびサーバーに負荷をかけすぎる。

## 【0005】

主要な難点は、検索システムにリクエストを送ったユーザーが、要求したデータを含んでいないことがよくある多量のデータを取得してしまうことにある。受け取った全部の文書を、それが必要なデータを含んでいるか否か確認するために吟味する必要性が生ずる。これは不必要な時間および知的作業の無駄につながる。巨大なインターネットのデータ列から、ユーザーが各種の問題を解決するために必要とする特定のデータおよび知識をリアルタイムで獲得することが不可能であることにより、情報の価値と検索システムの効率の両方が著しく低下する。

## 【0006】

データベースからユーザーのリクエストにより知識およびデータを抽出する方法として、下記の小論に記載されたインテリジェント情報 - 論理コンピューティングシステムにおいて実行される方法が公知である：Nasyzny V.V., 人工知能情報工学に基づく開システムデザインの理論の開発、モスクワ、1994 (pp. 85-112)。この方法は、確率情報工学に基づき、論理的結論を用いて効率的な知識探索および知識のリアルタイムでのプロセッシングを提供する。この利点は、従来の人工知能システムで使用されている既存の知識プロセッシング法とは全く異なり、提供されるのが、探索時間と、一方では論理的プロセッシングとの、他方では応答(reply)の準備に要する知識の量との、線形関係である手法により得られる。しかし、該方法は、テキスト系文書から知識を抽出することはできない。なぜなら、この方法は知識ベースからの公式化情報のプロセッシングに向けられており、このプロセッシングは専門家およびエンジニアが、関与する知識に対して行うものであるからである。この欠点のため、該方法は、既存の情報検索システムにおいてテキスト系文書から知識を抽出するためには使用できない。

## 【0007】

別の公知の方法として、下記小論に記載されているテキスト系文書からの知識抽出方法がある：Nasyzny V.V., Nasyznaya G.A., インテリジェント情報検索システムの構築、モスクワ、Promethey-Publisher, 2001。該方法は、多量のテキスト系情報のリアルタイムでの形態素的(morphological)、統語的(構文的、syntactical)および意味的(semantic)な解析を可能にする確率的インテリジェント情報工学に基づく。このシステムは、既存の情

10

20

30

40

50

報検索システムと一緒に、該システム上のインテリジェント上部構造として動作させることができ、またテキスト系文書の確率的指標付け(stochastic indexing、確率による指標付けもしくは索引付け)、情報交換プロトコルおよびユーザー・リクエストのプロセッシングのそれ自体の基準を用いた次世代の検索システムを提供する。現行の検索システムで実施されている方法と比べた該方法の主要な利点は、次の通りである：

- 自然言語でのユーザー・リクエストのプロセッシング；
- ユーザーのリクエストに適合した全情報を確実に含む文書の探索；
- 特定の問題を解決するのに要求される各種主題のデータおよび知識を含む、ユーザーのリクエストに従ったテキスト断片の強調表示(highlighting)。

#### 【0008】

該方法の主要な弱点は、形態素的、統語的および意味的解析のためのインテリジェントシステムの知識ベースが専門家によって充填され、それにはかなりの量の時間と技術費用を必要とすることである。即ち、インターネット内に国内用サブシステムを持つ先進国民のユーザーのニーズを満たすために、テキスト系文書から知識を抽出するための類似システムの創出には長い時間がかかる。従って、該方法は、テキスト系文書から知識を抽出するためのインターネットに基づく多言語システムの創出には使用することができない。この障害により、国内用検索システムのテキスト系情報に基礎を置き、かつ文明社会の最新の要求を考慮して、多様な領域 - 産業、科学、教育、文化および家庭活動 - における質的に新規な情報サービスを提供したいであろう知識産業への移行が妨げられる。

#### 【0009】

辞書類が含んでいない新しい語(ワード)の自動解析の可能性が欠如していることは、該方法のさらに別の弱点であると考えられることができる。そのような語がテキスト系文書に現れた場合、専門家はその新語が属するスピーチ・パート(speech part、言葉の役割)を決定し、かつその形態素的特徴を決定しなければならない。このため、システムは、ある新規トピックに関してテキスト系文書をプロセッシングするために自動的にチューニングすることができない。さらに、知識の効率的な抽出には、種々の文書からのテキスト断片を、かかる断片中の意味的関係の論理的結論を用いた解析によって、かつあるテキスト中のある文章の等価変換(equivalent transformations)によって、包括的にプロセッシングする必要があることにも留意すべきである。このような機能は、該方法では実現されなかった。

#### 【発明の開示】

#### 【0010】

本発明の目的は、世界的なインターネットに基づく知識産業の創出に使用され、上述した欠点のない、検索システムのテキスト系文書から知識を抽出するための自己学習システムを合成する方法を提供することにある。本発明の実施を通して得られる結果は次の通りである：

- 知識ベースに充填(filling-in)するための、種々の言語の電子形態のテキスト系文書からの知識抽出による知識の自動的創出の可能性；
- 新しい語の自動解析と辞書類の更新；
- 知識抽出の効率を改善するためのユーザー・リクエストおよびテキスト系文書の文章の等価変換；
- 文法および意味解析のルールに対する該システムの自己命令(self-instruction)；
- ある外国言語での知識を抽出するためのテキスト系情報およびユーザー・リクエストのインテリジェントプロセッシング。

#### 【0011】

本発明の目的は、下記工程を含む、検索システムに使用するためテキスト系文書からの或る所定の自然言語での知識抽出用の自己学習システムの合成方法において達成される：

確率的情報インデックス(確率による情報の指標もしくは索引)の2進信号の単一(独自)組み合わせ(unique combinations)の適用に基づく、確率的に指標付けされた人工知能システムの形態の自己学習メカニズムを用意し；

10

20

30

40

50

確率的に指標付けされたテキスト断片の等価変換および論理的結論を使用し、かつ該断片からのリンクされた複数の意味構造 (semantic structures) を形成して、それらの意味構造をプロダクションルール (production rules) の形態での表示のために確率的に指標付けすることによって、文法および意味解析のルールに対してシステムを自動的に命令し；

該言語での電子形態の言語文書の形態素解析および確率的指標付けを、形態素解析ルールに対するシステムの自動命令と同時に行い；

該言語での、あるテーマに関係する、電子形態のテキスト系文書の形態素および統語解析 (構文解析)、ならびに確率的指標付けを、統語解析ルールに対するシステムの自動命令と同時に行い；

その所定テーマに関係する、電子形態の確率的に指標付けされたテキスト系文書の意味解析を、意味解析ルールに対するシステムの自動命令と同時に行い； 10

所定の自然言語でユーザーの1つのリクエストを形成し、それを1つの問い合わせ文章 (interrogative sentence) として確率的指標付けした後に、それを電子形態に変換し；

確率的に指標付けされた形態のユーザーのリクエストを、該ユーザーのリクエストに等価の1セットの新たな複数のリクエストに変換し；

該新たな複数リクエストの全部のワード・コンビネーション (語の組み合わせ) を含む、確率的に指標付けされた電子形態のテキスト系文書の断片を、ユーザー・リクエストに基づいて予備選択し；

該確率的に指標付けされたテキスト系文書の断片から1つの確率的に指標付けされた意味構造を生成させ； 20

この意味構造に基づいて、確率的に指標付けされたテキスト系文書の断片間のリンクを与える論理的結論とテキストの等価変換とにより、システムから簡単な応答を生成させ；

該簡単な応答から問い合わせ文章を生成させ、生成した問い合わせ文章をユーザーのリクエストと比較することにより、該簡単な応答のユーザー・リクエストとの適合性をチェックする；

ここで、生成した問い合わせ文章がユーザーのリクエストと同一である場合には、該簡単な応答のユーザー・リクエストとの適合性を認め、該簡単な応答を所定の自然言語でユーザーに提示する。

#### 【 0 0 1 2 】

本発明の目的は、下記工程を含む、検索システムに使用するための、テキスト系文書からの任意の所定の自然言語での知識抽出用の自己学習システムの合成方法において達成される； 30

確率的指標付けのために確率的情報インデックスの2進信号の単一組み合わせを適用することと、文法および意味解析手順の記述を含むある基本言語での言語テキスト断片の探索とに基づく、確率的に指標付けされた人工知能システムの形態の自己学習メカニズムを用意し、そして確率的に指標付けされた言語テキスト断片の等価変換および論理的結論を使用し、かつ該断片からリンクされた複数の意味構造を形成し、該意味構造をプロダクションルールの形態での表示のために確率的指標付けすることによって、文法および意味解析のルールに対してシステムを自動的に命令し；

その所定基本言語での電子形態の言語文書の形態素解析および確率的指標付けを、同時に形態素解析ルールに対してシステムを自動命令しながら行い、所定の各外国言語についての、確率的に指標付けされた辞書類のデータベースおよび言語テキスト・インデックスの表 (テーブル) と、基本言語および所定の各外国言語についてのプロダクションルールを含む形態素解析の知識ベースとを構築し； 40

検索システムから、所定の各外国言語での、あるテーマについての、電子形態のテキスト系文書の形態素および統語解析、ならびに確率的指標付けを行って、該文書をテキスト系文書のインデックスの表として表示し、かつ該文書を確率的に指標付けされたテキストのベースに格納し、その間に同時に、基本言語の確率的に指標付けされた言語テキストを用いて統語解析ルールに対してシステムの自動命令を行い、そして基本言語および所定の各外国言語についての統語解析の知識ベースを構築し； 50

その所定テーマについて、該電子形態の確率的に指標付けされたテキスト系文書の意味解析を、意味解析ルールに対するシステムの自動命令と同時に行って、基本言語および所定の各外国言語についての意味解析の知識ベースを構築し；

ある自然外国言語でユーザーの1つのリクエストを形成し、それを、1つの疑問語ワード・コンビネーション(interrogative word combination)とユーザーのリクエストの意味を決めるワード・コンビネーション(word combinations)とを含んだ1つの問い合わせ文章として確率的指標付けした後に、それを電子形態に変換し；

確率的に指標付けされた形態のユーザーのリクエストを、該ユーザーのリクエストに等価の1セットの新たな複数のリクエストに変換し；

該新たな複数リクエストの全部のワード・コンビネーションを含む、確率的に指標付けされた電子形態のテキスト系文書の断片類の、ユーザー・リクエストに基づく予備選択(仮選択)を行い； 10

該確率的に指標付けされたテキスト系文書の断片類から1つの確率的に指標付けされた意味構造を生成させ；

この意味構造に基づいて、確率的に指標付けされたテキスト系文書の断片間のリンクを与える論理的結論とテキストの等価変換とにより、システムから簡単な応答を生成させ、この応答は、ユーザー・リクエストの意味を規定する確率的に指標付けされたワード・コンビネーションと、ユーザー・リクエストの疑問語ワード・コンビネーションに対応する1つの応答ワード・グループ(word group)、とを含んでおり；

応答ワード・グループを、対応する確率的に指標付けされた疑問語ワード・コンビネーションで置換し、生成した問い合わせ文章をユーザー・リクエストと比較することにより、該簡単な応答のユーザー・リクエストとの適合性をチェックする； 20

ここで、生成した問い合わせ文章がユーザーのリクエストと同一である場合には、該簡単な応答のユーザー・リクエストとの適合性を認め、該簡単な応答を所定の外国言語でユーザーに提示する。

#### 【0013】

好ましくは、本方法は、ユーザーのリクエストに同一の問い合わせ文章を生成することに失敗した場合に、検索システムから、ユーザーのリクエストに適合する応答を探索するための新たなテキスト系文書をリクエストする工程をさらに含む。

#### 【0014】

また、ユーザーのリクエストにより、より詳しい情報または特定の知識を含む完全な応答を、確率的に指標付けされた意味構造を形成するための論理的結論、ならびに該簡単な応答のより詳しい内容を提供する新たな確率的に指標付けされたテキストを得るための該テキスト系文書断片の必要な等価変換により生成させてもよい。

#### 【0015】

本方法において、形態素解析ルールに対してシステムを自動命令する工程は下記を含んでいることが好ましい：確率的に指標付けされたテキストにおいて、複数の語のそれぞれの語形(word forms)の予め決められたセットを選択し、語幹、ならびにその語尾、接頭辞、接尾辞および前置詞の予め決められた1セット、の確率的インデックス(確率による索引もしくは指標)を付与し、該インデックスに従って確率的に指標付けされた言語テキストにランダムにアクセスして、それから、該語尾、接頭辞、接尾辞および前置詞のセットを、ある語に対応するスピーチ・パートならびに語の語形変化もしくは活用変化から生ずる語尾、接頭辞、接尾辞および前置詞の完全セットに関連づける、いくつかの断片を選択し、該断片を確率的インデックスによりプロダクションルールの形態に変換し、ここで、各ルールの正しさは、対応する言語テキストからのいくつかの断片に基づく自律導出により得られ、そして形態素解析の知識ベースに対するプロダクションルールのインデックスの表を得る。 40

#### 【0016】

好ましくは、形態素解析の知識ベースのルールを用いて各語のスピーチ・パートを決定した後に言語テキストの確率的指標付けを行う工程は、確率的に指標付けされた辞書類の 50

データベースに、各語の語幹の確率的インデックスならびにその語尾、接頭辞、接尾辞および前置詞の完全セットの確率的インデックスを充填することを含み、テキスト・インデックスの表を構築する工程は、情報の確率的変換と、語幹、それらの語尾、接頭辞、接尾辞、前置詞、文章、段落(パラグラフ)およびテキスト題名のインデックスの2進信号の単一組み合わせ(unique binary combinations)を生成させ、それらのインデックスは確率的に指標付けされたテキストのベースのインデックスの表内に置かれ、該インデックス間のリンクを付与し、このリンクが原テキスト内で特定されて、インデックスの表を用いたテキスト・リカバリを保証することを含む。

【0017】

本方法において、統語解析のルールに対してシステムを自動命令する工程は、好ましくは下記を含む：確率的に指標付けされた言語テキスト内で、文章の統語解析の手順を記述する断片を探索し；統語要素および構造と語の予め決められたスピーチ・パートとの間のリンクを規定する確率的に指標付けされた意味構造を得るための論理的結論を取得し；形態素的な語の特徴に関して文章の統語解析を明示するプロダクションルールを導出し、ここで、各ルールの正しさは対応する言語テキストからのいくつかの断片に基づく自律導出により得られ、生成したルールを統語解析の知識ベースに格納し、確率的に指標付けしてインデックスの表の形態で表示する。

【0018】

また、意味解析のルールに対してシステムを自動命令する工程は、下記をさらに含んでもよい：語幹およびスピーチ・パートの確率的インデックスを参照して、言語テキストのインデックスの表への、正確に規定されていない文章メンバー(文章構成要素)のリクエストを形成して、特定の文章メンバーと一致させるためにその語が有する意味的特徴を記述するテキスト断片としての応答を取得し；そして、該応答に従って、ある所定の語の語幹の確率的インデックスと要求された意味的特徴を用いて、汎用辞書もしくは専門辞書ならびに百科辞典のインデックスの表を参照し；そして、論理的結論によって、その所定の語と要求された意味的特徴とをリンクする確率的に指標付けされた意味構造を特定することを試み；その試みが成功したなら、該文章メンバーが正確に決まったと決定し；リクエストに適合するテキスト断片をプロダクションルールに変換し、ここで、各ルールの正しさは、対応する言語テキストからのいくつかの断片に基づく自律導出により得られ、該ルールを意味解析の知識ベースに格納し、確率的に指標付けして、文章メンバーとしての語の意味解析およびワード・コンビネーション間のリンクに使用するためのインデックスの表の形態で表示する。

【0019】

本方法は、下記をさらに含んでもよい：各テキストのインデックス表を生成させ、該テキストを形態素的、統語的および意味的に解析し終わった後、文章のそれぞれの中の各語に対応するスピーチ・パートの名称、文章メンバーおよびそれらへの質問の確率的インデックスを生成させ、該インデックスを該テキストのインデックスの表にエントリして、テキスト断片の探索において、語のそれぞれが何のスピーチ・パートおよび文章メンバーに属しているかの自動的な決定を与え、かつ該語に対する質問を示し；そしてさらに、テキストのインデックスの全部の表を生成させた後に、あるテーマについてのインデックスの表を生成させ、この表において、行(ロウ)は、語幹の繰り返されない確率的インデックスにより指定され、各列(カラム)は、特定のテキストの確率的インデックスに対応し；そして特定の語幹インデックスを持つある語を含むテキスト段落の確率的インデックスを該表にエントリし、その所定テーマについてのインデックスの表を、ユーザー・リクエストの予め決められたワード・コンビネーションのセットを含む断片の予備探索に使用する。

【0020】

本方法において、好ましくは、ユーザーのリクエストを等価変換する工程は、ほぼ同じ意味を持つ語である類義語を使用し、形態素、統語および意味の各解析の確率的に指標付けされたルールに基づいて、ユーザー・リクエストの問い合わせ文章のワード・コンビネ

ーションの等価構造を得ると共に、それらの間の意味的關係を維持するように、ユーザー・リクエストの意味を保存しながらのスピーチ・パートおよび文章メンバーの置換を行うことを含み；そしてユーザー・リクエストの全部のワード・コンビネーションを含む意味的にリンクされたテキスト断片を生成させる工程は、下記を含む：該語幹の確率的インデックスに従って、所定テーマに関してテキスト・インデックスの表を参照し、ユーザー・リクエストの全部のワード・コンビネーションを含む段落および対応するテキストの確率的インデックスを選択し、該インデックスに従って、選択されたテキストのそれぞれのインデックスの表を参照し；インデックスの表およびテキストの等価変換に基づいて論理的結論を作成して、ユーザー・リクエストの疑問語ワード・コンビネーション、ならびにユーザー・リクエストの意味を規定し、予備選択された段落が含んでいるユーザー・リクエストの全部のワード・コンビネーションに対応する、応答のワード・グループのインデックスをリンクする確率的に指標付けされた1つの意味構造を生成させる。

10

**【0021】**

また、本方法は下記をさらに含むことが好ましい：論理的結論により首尾よく生じた、ユーザー・リクエストに対応する確率的に指標付けされた意味構造を、得られたテキスト断片のセットを用いて、ユーザー・リクエストと同一の問い合わせ文章を生成させるための基礎として使用し；該問い合わせ文章の生成を、ユーザー・リクエストのテキスト系文書断片の各ワード・コンビネーションの要求された意味的特徴を与えるための該知識ベースからのルールに基づく語幹ならびに語尾、接尾辞、接頭辞および前置詞の確率的インデックスの等価変換により、かつ複数のワード・コンビネーション間の推移的關係(transitive relationships)に基づく論理的結論を使用して、それらを、ユーザー・リクエストの疑問語ワード・コンビネーションに対応する該応答のワード・グループを含んでいる、ユーザー・リクエストに同一の問い合わせ文章に組み合わせることによって行い；ここで、簡単な応答の正しさは、多様な予備選択した確率的に指標付けされたテキスト系文書の断片に基づいて、該応答の確率的に指標付けされた同一の意味構造がいくつか生成することにより保証される。

20

**【0022】**

さらに、本方法は下記をさらに含むことが好ましい：テキスト系文書のインデックスの表を用いた探索プロセスおよび応答の生成中に、リクエストとそれに適合する簡単な応答とをリンクする指標付けされたテキスト系要素を生成させて「リクエスト-リプライ(応答)」型の要素を含む知識ベースを生じさせることにより、システムの自己学習を行い、この知識ベースは、確率的指標付けされた後に、インデックスの表の形態で提示されて、テキストの文章の文法および意味解析ならびに該指標付けされた知識ベース中に含まれる繰り返されたリクエストに対する応答の生成のために使用され；ここで、簡単な応答に基づき、かつテキスト断片を得る時に使用されたインデックスの表に従った論理的結論を用いてユーザー・リクエストに適合する知識を含んでいる完全な応答を生成させる工程は、応答のワード・グループを文章の語幹の確率的インデックスにリンクする確率的に指標付けされた1つの意味構造を生成させることを含み、そして、このリンク(リンクすること)は、テキスト断片内の簡単な応答の完全な開示を与える推移的關係を保持して、該確率的に指標付けされた意味構造に基づく文章の等価変換を用いて完全な応答のリンクされたテキストが得られる。

30

40

**【0023】**

本方法において、確率的に指標付けされた断片の等価変換は、好ましくは下記を含む：各文章を確率的に指標付けされた複数のワード・コンビネーションの1セットとして提示し、該コンビネーションを、共通語根の語幹、語尾、接頭辞、接尾辞および前置詞の確率的インデックスの等価変換によって、形態素、統語および意味の各解析の知識ベースに格納されたルールを用いて変換して、新たなスピーチ・パートおよび文章メンバーを生じさせ、但し、各文章の確率的に指標付けされた意味構造におけるワード・コンビネーション間のリンクの無矛盾性(constancy)および新たなテキスト断片を生成させた時の文章間の一致性(concordance)が条件である。

50

## 【 0 0 2 4 】

さらに、確率的に指標付けされた語の辞書中または言語テキスト中に含まれていない新しい語が、テキスト系文書の確率的指標付けのプロセス中に指標付けされたテキスト中に出現した時には、本方法は下記を含むことが好ましい：辞書内でのその新語に対する共通語根の語の検索と、形態素解析の知識ベース内での該共通語根の語の新語への等価変換のルールを検索とを行い；等価変換のタイプにより、その新語が属するスピーチ・パートならびに語形変化もしくは活用変化により生ずるその語形の全部を決定し、

そして、辞書内に共通語根の語が全く見つからない場合には、テキストから、その新語の特定の語形の1セットを選択し、該語形の語尾、接尾辞および接頭辞に基づいて、確率的に指標付けされた辞書または形態素解析のプロダクションルールを用いて、該新語が属するスピーチ・パートならびに語形変化もしくは活用変化により生ずるその語形の完全セットを決定する。

## 【 0 0 2 5 】

また、本方法は、複数の所定外国言語中のテキスト系文書からの知識の同時抽出を達成するが、それは、所定の1つの基本言語に関して形態素、統語および意味の各解析のルール内でのシステムの自動命令を行い；ある基本言語で、確率的に指標付けされた辞書のデータベースと確率的に指標付けされた言語テキストを用いた形態素、統語および意味の各解析の知識ベースとを構築し；該知識ベースを用いて、所定外国言語のいずれかの自動命令のためのリクエストを自動生成させ、該リクエストに従って、該外国言語の学習に必要な知識を含んでいる、基本言語の言語テキスト断片を予備的に(仮に)選択し、該テキストの等価変換を行い；確率的に指標付けされた意味構造を生成させ、該構造に対する論理的結論を作成して、自動生成させたリクエストに適合する応答を生成させ、該応答を所定外国言語のいずれかの形態素、統語および意味の各解析の知識ベースを生成させるために使用し、所定の外国言語のテキスト系文書からの知識の抽出を保証する、ことにより実行される。

## 【 発明を実施するための最良の形態 】

## 【 0 0 2 6 】

本発明を図1に示した実施例ならびに下記の表：

表1 - 文章フレーム、

表2 - テキスト・インデックス、

表3 - あるトピックに関するテキストのインデックス、

によりさらに説明する。

## 【 0 0 2 7 】

本明細書で用いた用語を下記のように規定する。

知識ベース - 1つのトピックに関する概念、ルールおよび事実の系統的セットを格納する1または2以上の特別に配列されたファイル。

## 【 0 0 2 8 】

疑問語ワード・コンビネーションとは、ワード・コンビネーション(語の組み合わせ)中の主要な語(名詞または動詞)に関連して疑問語(interrogative word)として作用する疑問代名詞または疑問副詞を有するワード・コンビネーションである。

## 【 0 0 2 9 】

文法解析 - 形態素的かつ意味的な解析。

知識とは、テキスト系文書に明確には含まれていない新たなテキスト系情報であり、その情報は、等価変換および論理的結論を用いて、応答の形態でシステムにより自動的に生成し、かつその情報はユーザーのリクエストに適合性がある、そのリクエストに従って対応する問題を解決するためのものである。

## 【 0 0 3 0 】

言語テキストとは、ある言語を学習するための教育方法論的、科学的、参照的(参照辞書類、百科辞典類)ならびに他のテキストである。

論理的結論とは、知識を処理するための技法であって、その技法は頭脳的推理プロセス

10

20

30

40

50

を模倣し、或る種の言語ユニットに基づいて、ある決まった意味を持つ1つの意味構造を合成することを可能にする。

【0031】

形態素解析とは、1つの形態素的コンポジションを決定し、その後スピーチ・パートを相互に関係づける別々の語の特徴を詳述するための、文章中の語の分析的な検討であり；その際には、まずその語(ワード)の、文章中でのその位置に依存しない、恒久的な形態素的な特色が特定され、次にその語形変化または活用変化に関係した語の文法的形態が分析される。

【0032】

語幹とは、1つの語の中でその語彙的(辞書的)意味を持つ部分であり、語形変化(屈折)および活用変化を受けることができる語は、語幹と語尾を有し、他の種類の語は語幹のみを有する。 10

【0033】

検索システムとは、キーワード、トピックなどを参照しながら情報の自動的探索を実施するために動作可能なシステムである。

プロダクショナルルールとは、「...なら(if)(条件)、...である(then)(結論)」という複文として知識を表示する1つの形態であり、ここで条件部はあるトピック・スフィア(topical sphere)中の複数のオブジェクト間の叙述的關係または他の関係を含み、論路的な”and”で結合された、種々のワード・コンビネーションを含み；結論部は、正しい意味論的帰結、または条件の全部のワード・コンビネーションが正しい場合に開始される動作を規定する、1つのワード・コンビネーションまたは1セットの複数のワード・コンビネーションを含んでいる。 20

【0034】

適合性とは、1つの文書がユーザーのリクエストに特定された基準をどの程度満たしているかを定める尺度である。

意味(的)構造とは、異なる文章の別々の言語ユニットの、それらの間の関係のタイプに関する1つの形態の關係であり、この構造は解析されたテキストの或る種の意味論的意味(semantic meaning)を表している。

【0035】

意味解析とは、別々の言語ユニット：1つの文章中の複数の語、複数のワード・コンビネーション、の意義、トピック・スフィアのオブジェクトと現実的現象との間に特定の種類の關係を持つそれらの相關關係の、意味の分析である。 30

【0036】

統語解析(構文解析)とは、1つの文章の統語(構文)構成を決定するためのその分析的検討であり、それにより、別々の語、ワード・コンビネーション、それらのタイプ、ワード・コンビネーションおよび文章中の語どうしの關係の種類、複数の文章の構造、複数の文章の構造のタイプの特徴が特定される。

【0037】

人口知能のシステムとは、その基礎として、論理的結論サブシステム、知識ベース、および他の人口知能のソフト/ハードウェアを含むソフト/ハードウェアシステムであって、人の知的活動を支援するか、または制御プロセスのオペレータに代わるためのものである。 40

【0038】

語の語形変化とは、場合もしくは格による名詞の変化(大部分の名詞にとって単数形/複数形による変化)、ならびに形容詞および他の支配された語の変化 - 性別による変化も、である。

【0039】

ワード・コンビネーションとは、依存の關係 - (文法上の)呼応、支配または隣接(付加) - に基づき、また、この關係により生じた語彙 - 文法間の關係にも基づく、2以上の語により構成される統語単位である。

## 【0040】

語形とは、ある決まった文法的形態で存在する語のことである。

語の活用変化とは、人称、数、時制および語形変化に関する動詞の変化であり、過去時制および単数の仮定法では - 性別、にも関係する。

## 【0041】

等価変換とは、1つの文章意味構造内、または1つのテキストの同じ意味論的意味を表すことができる複数の文章からなる或る1セットにおける、別々の言語ユニット同士の、それらの関係を保持した相互置換である。

## 【0042】

特許請求した方法の実施を、図1に示した検索システムのための知識抽出用の知的自己学習システム(I S L S E K, intellectual self-learning system for extracting the knowledge)の構造および操作の例により詳細に説明する。この確率により指標付けされた人口知能のシステムは、下記を備える：

- 多言語プロセッサ(1);
- テキスト系文書の確率的指標付けおよびテキスト断片の選択を行うサブシステム(2);
- 自己命令および知識抽出モードを制御するためのサブシステム(3);
- 確率的に指標付けされたテキストとプロダクションルールの通訳プログラム(4);
- テキストの等価変換のサブシステム(5);
- 論理的結論のサブシステム(6);
- 基本語と新語の確率的に指標付けされた辞書類のデータベース(7);
- 確率的に指標付けされた言語テキストのデータベース(8);
- 「リクエスト - リプライ」知識ベース(9);
- 一定のテーマに関する確率的に指標付けされたテキスト系文書のデータベース(10)
- 外国語の確率的に指標付けされた辞書類のデータベース(11);
- 形態素解析の知識ベース(12);
- 統語解析の知識ベース(13);
- 意味解析の知識ベース(14);
- 確率的に指標付けされたワード・コンビネーションのデータベース(15)。

## 【0043】

該システムは、確率的変換およびシンボル情報(symbolic information)の指標付け、自己命令モードを制御するためにプロダクションルールのインデックスの表(テーブル)のコンパイル、ならびにテキスト・インデックスの使用に基づいている。このシステムは、選択されたテキスト断片からの知識の抽出ならびにプロダクションルールの形態もしくはユーザー・リクエストへの応答としての知識の表示のために確率的に指標付けされたルールを使用して、テキスト系情報断片、論理的結論およびテキストの等価変換への、確率的インデックスに従ったアクセスを与える。

## 【0044】

I S L S E Kの創出は、言語テキストに基づくテキスト系情報の形態素解析、統語解析および意味解析のルールについてシステムを自己命令するためのメカニズムを与える。該テキストは、汎用辞書類、専門辞書類、類義語辞書類、参考辞書類、所定言語群の文法に関する教育方法論テキスト等を含んでいる。

## 【0045】

ユーザーのシステムとの伝達は多言語の言語(処理)プロセッサ1を介して行われる。言語プロセッサ1は、自然言語で書かれたユーザーのリクエストを入力し、システムが生成した応答を出力する。ユーザーとシステムとの情報交換は、複数の所定言語のどれで行うこともできる。言語プロセッサ1は、自己命令および知識抽出モード制御サブシステム3の指令で、I S L S E Kに接続された検索システムとの相互作用を生ずる。この相互作用の目的は、さらなる処理のために、サブシステム3からの指令で、或るテーマに関する所定言語の検索システムからの新たなテキスト系文書を入力することにある。多言語の言語プロセッサ1はまた、その所定言語での言語テキストの電子形態でのシステムへの入力を

行う。

【0046】

言語テキストの形態素解析および形態素解析ルールに対するシステムの自動命令が、基本辞書のコンパイルならびに確率的に指標付けされた基本語と新語の辞書類のデータベース7への該辞書のエントリの間、自己命令および知識抽出モード制御サブシステム3からの指令で行われる。これらの機能は、テキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2を用いて、言語テキストの指標付けと同時に行われる。

【0047】

確率的に指標付けされた基本辞書のコンパイルのために使用されるのが電子形態の汎用辞書であり、これは多言語の言語プロセッサ1を介してシステムに入力される。該辞書の所定の語形(語の形)に従って、各語のスピーチ・パート、その語幹と語尾セットが決定される。語幹は、テキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2を用いて確率的に指標付けされ、該語幹は、確率的に指標付けされた基本語と新語の辞書類のデータベース7内において、確率的に指標付けされた基本辞書の表(テーブル)における語幹のインデックスの列(カラム)に格納される。

【0048】

該辞書の語の上記プロセッシングの結果として、多言語の言語プロセッサ1は全部の語の語幹および語幹それ自体の確率的なインデックスを生成し、同時に語尾のセットが確率的に指標付けされた基本語と新語の辞書類のデータベース7に格納される。

【0049】

確率的に指標付けされた基本語と新語の辞書類のデータベース7に格納された確率的に指標付けされた基本辞書は、それぞれ決まったスピーチ・パートに対応する多くの表フォーマットを有する。表のヘッダ(見出し)は、形態素的特徴(性、数、格、人称、時制等)の名称と、さらに所定の語が語形変化または活用変化をする場合に生ずるその所定語の語の語形に対応する質問とを含むセルを備える。各語幹はその語の該語形の語尾を含む行(ロウ)に対応する。確率的に指標付けされた基本辞書の表の充填が始まる時には、語のそれぞれのわずかな語形しか、即ち、汎用の語の辞書内に提供されている語形しか知られていないことに留意すべきである。確率的に指標付けされた基本辞書の表の充填のための他の語形およびそれらのそれぞれの語尾、接尾辞および接頭辞は、適当な言語テキストの初期指標付けの後に形態素解析のルールによりシステムの自動的な自己命令のモードにおいて特定される。

【0050】

このメカニズムの主な原理は、テキスト系文書の確率的指標付けの新規な方法の使用であり、この方法はテキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2で行われる。この手順は、シンボル情報の確率的変換ならびに語幹、文章、段落および書誌的参照(bibliographic references)を含むテキスト系文書のタイトルの2進信号の単一組み合わせの形態の確率的インデックスの生成という機能に基づく。語幹の確率的インデックス $\{I_{ij}^{(u)}\}$ 、文章の確率的インデックス $\{I_{ij}^{(p)}\}$ 、段落の確率的インデックス $\{I_{ij}^{(a)}\}$ 、および処理されたテキストのタイトルの確率的インデックス $\{I_{ij}^{(t)}\}$ を生成させる、シンボル情報の確率的変換は、各文章のフレームの生成(表1)およびある所定テキストのインデックス表の生成(表2)と同時に達成される。

【0051】

テキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2内で作成されたこのフレーム(表1)は、10レベル(行)の-slot(セル)を含む。これらの-slotセルがテキストの確率的指標付けの際に、ならびに各文章の形態素、統語および意味の各解析中に充填される。

【0052】

言語テキストの確率的指標付け中に、第1レベルの-slotには、語幹およびそれらの語尾の確率的インデックス $\{I_{ij}^{(u)}\}$ が書き込まれる。第2レベルの-slotは、i番目の文章における語をそれらの文章中での順序に従って含んでいる。前置詞、不変化詞(par

10

20

30

40

50

ticles)、接続詞および句読点が、それらが関連する語のスロットに置かれる。第3レベルのスロットを充填するのに、第1レベルのスロットに書き込まれた語幹およびそれらの語尾の確率的インデックス $\{l_{ij}^{(u)}\}$ が使用される。

【0053】

語幹インデックスに従って、確率的に指標付けされた基本辞書の対応する表の行(ロウ)にアクセスし、その行はある所定の語が関連するスピーチ・パートの決定のために同一インデックスにより指定される。確率的に指標付けされた基本語と新語の辞書類のデータベース7からの該情報は、第2レベルのスロットの語に対応する文章フレーム第3レベルのスロットに書き込まれる。

【0054】

スピーチ・パートの特徴はフレームの第3レベルのスロットに書き込まれ、第4～第10レベルのスロットは、その後のテキストの形態素解析および統語解析の間に充填され、これは形態素解析および統語解析ルールに対するシステムの命令と同時にされる。

【0055】

充填された最初の4スロットレベルを持つテキスト文章フレームに基づいて、テキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2が、その所定テキストのインデックスの表を生成する。

【0056】

表2は、行(ロウ)が語幹の確率的インデックス $\{l_{ij}^{(u)}\}$ により指定され、列(カラム)がテキスト内のそれらの出現の順序での段落インデックス $\{l_j^{(a)}\}$ により指定されるインデックスの表を示し、対応する列および行の交差上に位置するセルは、リストのインデックス $\{l_{ij}^{(s)}\}$ を含む。 $\{l_{ij}^{(s)}\}$ として各リストに指定された情報は、別個のあるファイルに書き込まれ、それは一般に下記のデータを含んでいる：

$\{l_{ij}^{(p)}\}$ は、ある所定の語を含む文章のインデックスであり；

$N_i^{(n)}$ は、その所定の語を含む文章の数であり；

$(u_i u_j)$ は、その所定の語が文章 $(l_{ij}^{(p)} N_i^{(n)})$ 中で有する語尾であり；

$l_{j-1}^{(u)}$ は、その文章中またはテキストの段落中の先行する語のインデックスであり、ここで、

$l_j^{(u)}$ が文章(段落)中の最初の語である場合には、インデックス $l_{j-1}^{(u)}$ の後には終止符が来る。 $l_{j-1}^{(u)}$ は所定段落内の先行する文章または先行する段落の最後の語に対応することができる。 $l_{j-1}^{(u)}$ の後にカンマが来る場合には、それは $l_j^{(u)}$ が分詞もしくは副詞的分詞構文、従属節、または複文中の単文の始まりかもしれないことを意味する；

$l_{(j+1)}^{(u)}$ は、その文章、段落、テキスト中の次の語のインデックスであり、

$l_j^{(u)}$ が文章(段落)中の最後の語である場合には、終止符が $l_{j-1}^{(u)}$ に先行する；

$l_{j-1}^{(u)}$ は、この段落または次の段落の新たな文章の始まりの語に対応していてもよい。 $l_{j-1}^{(u)}$ の前にコンマがくる場合には、それは $l_j^{(u)}$ が副詞的分詞もしくは分詞構文、または複文中の単文の最後かもしれないことを意味する；

$l_j^{(vu)}$ は、その所定の語への文章メンバーに関する質問のインデックスであり；

$l_j^{(pu)}$ は、その所定の語が関連する文章メンバーの指定のインデックスであり；

$l_j^{(vpru)}$ は、 $l_j^{(u)}$ で始まる副詞的分詞もしくは分詞構文または従属文章が対応する質問のインデックスであり；

$l_j^{(pru)}$ は、 $l_j^{(u)}$ で始まる副詞的分詞もしくは分詞構文または従属節が対応する文章メンバーの指定のインデックスである。

【0057】

上記インデックス類およびシンボル(記号)類は、 $l_j^{(a)}$ 段落の $l_j^{(p)}$ 文章の1つにある $l_{ij}^{(u)}$ 語幹を有する語に対応して、所定グループ内のインデックスおよびシンボルの位置を規定する予め決められたフォーマットを有する。インデックスが全く存在しない場合には、対応する位置に「空白」マークが挿入される。所定の語 $l_{ij}^{(u)}$ が、 $l_j^{(a)}$ 段落のn個の $\{l_{ij}^{(p)}\}$ 文章により構成される場合には、リスト内の該グループの量もn

10

20

30

40

50

となろう。

【0058】

$l_{ij}^{(s)}$  リストの最初の6個のインデックスはテキストの確率的指標付け中に生成することに留意されたい。それにより、語幹インデックス  $l_j^{(u)}$  に従って、確率的に指標付けされた基本辞書を参照して、所定の語が関係するスピーチ・パートを常に決定することが可能となる。 $l_{ij}^{(s)}$  リストの残りのデータの決定は、文章の文法解析のルールに対してシステムを自己命令すると同時に行われる、その後の形態素解析および統語解析中に、テキストの文章フレームの4～10レベルの充填後に行われる。

【0059】

文章の文法的解析の記述を含むテキストを包含する、全言語テキストの確率的指標付け後に、該テキストは確率的に指標付けされた言語テキストのデータベース8に格納され、本方法は、確率的に指標付けされた基本語と新語の辞書類のデータベース7の充填と同時にテキスト形態素解析のルールの導出へと進む。 10

【0060】

従って、所定のスピーチ・パートに關係する語の語幹を含んだ確率的に指標付けされた基本辞書の各表から、各語幹およびその語尾または前置詞の予め決められたセットの確率的インデックスが選択される。その後、確率的に指標付けされた言語テキストのデータベース8に該インデックスに従ってアクセスして、該スピーチ・パートのインデックスならびに該語の語尾もしくは前置詞のセットを、対応する語尾、前置詞または語形変化もしくは活用変化により生ずる質問の完全セットに相互に關係づけるテキスト断片を選択する。 20  
その後、このテキスト断片が、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4に入力され、ここで確率的に指標付けされた意味構造を、該断片が含んでいる文章のそれぞれのワード・コンビネーションの1セットとして生成させる：

$$S : \{ (l_{ij}^{(u)} | l_{ij}^{(r)} | l_{ij}^{(z)}) (l_j^{(u)} | l_j^{(r)} | l_j^{(z)}) \}, \quad (1)$$

式中、 $l_{ij}^{(u)} | l_j^{(u)}$  は、それぞれこのワード・コンビネーションの語の主要な語および従属的な語の語幹の確率的インデックスであり； $l_{ij}^{(z)} | l_j^{(z)}$  は、それぞれ該ワード・コンビネーションの語の主要な語および従属的な語のスピーチ・パートの形態素的特徴の確率的インデックスであり；そして記号  $|$  は、このワード・コンビネーションの主要な語と従属的な語との間の關係を決める。 30

【0061】

式(1)により表された確率的に指標付けされた各意味構造の主要なリンクは、所定構造パターン内の關係の意味(解釈)を決める動詞である。異なる文章に含まれている異なる複数の確率的に指標付けされた意味構造(1)の中に或る關係が成立するのは、それらの構造が同一のワード・コンビネーション、それらの類義語、主要な語の繰り返し(重複)を有しているか、もしくは第1の文章のワード・コンビネーションの1つに対応する代名詞の第2の文章中における使用があるか、ならびに主要な語と組み合わせた代名詞を有している場合である。それにより、指標付けされた最初のリクエストデータを含む確率的に指標付けされた意味構造が、その応答の指標付けされたデータを有する確率的に指標付けされた意味構造に適正に關係づけられている文章または文章部分が見つけれられる。動詞の意味を決定するために、動詞の語幹の確率的インデックスに従って、確率的に指標付けされた言語テキストのデータベース8が参照されて、類義語辞書類のインデックスの表へのアクセスがなされる。 40

【0062】

第1および第2の構造が、規定すべきスピーチ・パートを有するワード・コンビネーションを介して相互連結され、このスピーチ・パートに関連する動詞の意味がリクエストまたは推定される応答の動詞と同一または類義である場合には、これらの構造はテキストの等価変換のサブシステム5にエントリされる。サブシステム5は、該2つの意味構造を確率的に指標付けされた単一のプロダクションルールの意味構造に変換し、この意味構造はリクエストを含んだ条件と結論(応答)とを有する。該確率的に指標付けされた単一の意味構造は一般に次のように表される： 50

$$P : |_1^{(su)} |_2^{(su)} |_3^{(su)} \dots |_m^{(su)} \\ |_1^{(su)} |_2^{(su)} |_3^{(su)} \dots |_n^{(su)} \quad (2)$$

式中、 $|_i^{(su)}$  は、式(1)からの対応するワード・コンビネーションの確率的インデックス

$|_1^{(su)} : (|_i^{(u)} |_i^{(r)} |_i^{(z)}) (|_j^{(u)} |_j^{(r)} |_j^{(z)})$  であり、そして は、条件の全部のワード・コンビネーションが正しい(リクエストの最初のデータと合致する)場合に、式(2)の右側部分の論理的結論が式(2)の左側部分の条件から引き出されたという記号としての通常の論理学的意味と解される。各ルールの正しさが、上述した手順に従い、かつ対応する言語テキストから選択された断片の数に基づく、確率的に指標付けされたいくつかの同一の意味構造(2)が独立して生成することにより保証されることに留意されたい。

10

#### 【0063】

等価変換のサブシステム5内で導出された、式(2)の形態の各プロダクションルールは、確率的に指標付けされたテキストおよびプロダクションルールの通訳プログラム4に送られ、そこで式(2)はプロダクションルールのテキスト系フォーマットに変換される。このフォーマットは、「...なら(if, 条件)、...である(then, 結論)」と表現される。指標付けされた形態の導出されたルールが形態素解析の知識ベース11に送られる。確率的に指標付けされたルールを含む知識ベースを合成するための手順は後で詳細に述べる。

#### 【0064】

確率的に指標付けされた基本語と新語の辞書類のデータベース7の充填と同時にテキストの形態素解析のためのルールを導出する際に、最初の確率的に指標付けされた意味構造(1)は、語幹類の確率的インデックスを含んでおり、それらのインデックスはスピーチ・パートならびにその語尾および前置詞の予め決められた1セットを表示する。第2の構造(1)は、同一のスピーチ・パートインデックスを介して第1のものにリンクされ、そのスピーチ・パートの語形変化または活用変化により生ずる語尾、前置詞、質問の完全セットを決定する。

20

#### 【0065】

上述した手順に従って、確率的に指標付けされた言語テキストのデータベース8の言語テキストに対応する類義語辞書の指標付けの表を参照することにより、第1および第2の意味構造の動詞の意味の、リクエストおよび想定される応答とのコンプライアンス(順応性)がチェックされる。その後、第1および第2の意味構造をリンクするワード・コンビネーションが決定される。結果が肯定的である場合は、該テキスト断片の2つの部分がテキストの等価変換のサブシステム5に送られ、その後、確率的に指標付けされたテキストとプロダクションルールの翻訳プログラム4に送られる。その結果、該テキスト断片は、「...なら(if, 条件)、...である(then, 結論)」と表されるプロダクションルールのフォーマットに変換される。このルールの条件部は、語のスピーチ・パートと、辞書フォーマットに提供されて、その語が語形変化または活用変化する時に語形の変化を決める、語尾および前置詞の予め定められたセットとを関連づける、ワード・コンビネーションのインデックスを包含する。結論部は、対応するスピーチ・パートとして、その語の語形変化または活用変化により生じた語尾、前置詞および質問の完全セットを包含する。公式化されたプロダクションルールは、形態素解析の知識ベース11に書き込まれる。スピーチ・パートを決めるルール導出された後、プロセスは、自己命令および知識抽出モードを制御するためのサブシステム3からの指令により、共通語根の語の等価変換のためのルールを合成する段階に進む。その目的には、形態素解析の知識ベース11に予め格納され、そして確率的に指標付けされた基本辞書の表と、1つのスピーチ・パートを別の共通語根スピーチ・パートに基づいて生成させる手順を明示する適当な言語テキスト断片の選択とに基づく、スピーチ・パート変換のための下記的一般ルールが使用される：

30

40

「あるスピーチ・パートを別のものに変換すべきであるなら、  
まず、第1のスピーチ・パートの語幹を切り離し、  
次に、確率的に指標付けされた基本辞書フォーマットを参照して、その語幹が、第1ス

50

ピーチ・パートの語幹との、語根(ひょっとすると2つの語根、多分接頭辞を伴い、もしかするといくつかの有声字および子音字の変化、付加、削除を伴う)を含んだ共通部分を有している第2のスピーチ・パートを探索し; その語根を切り離した後、これらのスピーチ・パートの語幹を用いて、それらの接尾辞を切り離し; 次に、スピーチ・パートの語幹の確率的インデックスに関して言語テキスト・インデックスの表を参照することにより、あるスピーチ・パートを別のスピーチ・パートに変換するための適当な方法を記述している断片を選択し; そして、辞書フォーマットを参照して、第1スピーチ・パートの語幹に対する第2スピーチ・パートの語幹の形成のやり方を決定し(接尾辞の置換、除去、付加); その後、このあるスピーチ・パートを置換する方法が第1スピーチ・パートから第2スピーチ・パートを形成するための要求された方法に対応するか否かを決定し; そして、肯定の場合には、第2スピーチ・パートを、新たに形成されたスピーチ・パートとして採用する」。

10

## 【0066】

一般ルールを使用した特定の語の変換の一部として、対応する特定のルールがその一般ルールに基づいて導出され、その特定のルールは変換されたスピーチ・パート、接尾辞および1つのスピーチ・パートを別のものから形成するための方法を明示する。これは、確率的に指標付けされたテキストおよびプロダクションルールの通訳プログラム4およびテキストの等価変換のサブシステム5において起こる。上述した手順は、まず所定の断片を、プロダクションルールの単一の確率的に指標付けされた意味構造(2)に変換し、次に、「...なら(if, 条件)、...である(then, 結論)」と表されるプロダクションルールのフォーマットに変換する。これらのルールは、確率的指標付けがなされた後に形態素解析の知識ベース11にエントリされる。

20

## 【0067】

次のテキスト系文書の指標付け中に、その語幹が基本辞書中に見られない何か新しい語が現れた場合には、本方法は、その新語のスピーチ・パートと語形変化または活用変化により生じるその語尾類を規定する手順に進む。

## 【0068】

まず、その新語が属するスピーチ・パートを規定するために、その語の少なくとも2つの異なる語形をテキストから選択し; これらの語形を比較することにより、その新語の不変部分、即ち、その語幹と思われる部分、ならびにその語尾を決定する。次に、基本辞書フォーマットが新語の語幹と共通する語根(接頭辞がつくかもしれない)を有する語を何か含んでいるか否かを決定する。語根とは、複数の同語源の語(1つの有声字を含む少なくとも2字からなる)の語幹の分割できない共通部分のことであり、その部分は、それに接頭辞、接尾辞および語尾を付加することにより、共通語根のスピーチ・パートを形成するのに使用される。その手順によると、新語の語幹を基本辞書フォーマットから採取された語の語幹と、比較した二つの語 - 新語と基本辞書の次の語 - の分割できない共通部分が見出されるまで比較することにより、共通語根を切り離す。

30

## 【0069】

その後、形態素解析の知識ベース12が、その新語が関係づけられるスピーチ・パートを決めるのを可能にするルールを選択するため参照される。そのためには、等価変換の適当なルールが適用される。

40

## 【0070】

新語のスピーチ・パートを決めるのに等価変換のルールを使用するために、次のように仮定する: 該等価変換の一般ルールにおける第2のスピーチ・パートは、新語に関係し、かつ未知である; それとの共通語根を有する第1のスピーチ・パートは、既に辞書中に見出されており、従って既知である。次に、ルールに記述された変換を用いて、未知スピーチ・パートの新語の語幹が既知のスピーチ・パートから導出することができるか否かをチェックする。そのために、一般ルールから導出され、形態素解析の知識ベース12に含まれている一群の特定のルールを適用して、既知スピーチ・パートを別のスピーチ・パートに変換する。それらのルールの1つの適用が新語の語幹を生ずるなら、その語が属するスピー

50

ーチ・パートが既知になり、そのルールに特定されているように第2のスピーチ・パートに対応しよう。形態素解析の知識ベース12のプロダクションルールの使用によって、各スピーチ・パートの特徴をより詳細に決めることが可能となる。例えば、ロシア語の文書の形態素解析において、形態素解析の知識ベース12のルールが、新語のスピーチ・パートだけでなく、名詞(実名詞<substantive>、形容名詞)、主格、単数の語尾も決めることができる場合には、結果として、該ルールはその新語が関係する語形変化タイプ(ロシア語の1、2、3型)も決めることができる。これにより、実名詞、形容名詞、序数詞、ある種の代名詞、ならびに分詞の場合には、語形変化により生ずるそれらの語尾の完全セットを正確に特定することが可能となる。この場合、該スピーチ・パートにとって、その新語が有するのと同じ主格、単数での語尾を有する、対応する1つの語を辞書フォーマット内で見つけることで十分であろう。該スピーチ・パートの語尾の完全セットは、新語の語尾のセットに対応し、その語尾類はその語幹と一緒に新語の辞書フォーマットにエントリされよう。その後、語幹の確率的インデックスが形成され、生じた新語の特徴が新語辞書フォーマットに書き込まれる。

10

## 【0071】

新語が動詞である場合には、その後、上述した手順に従って、形態素解析の知識ベース12を参照して、その語幹を切り離した後、そのスピーチ・パートおよび不定形を、適当なルールを用いて決定および発見する。(ロシア語における)不定形の接尾辞(“- ”もしくは“- ”)に基づき、基本辞書フォーマットを参照して、不定形の同じ接尾辞(“- ”もしくは“- ”)を有する動詞を見つける。この動詞の活用変化により生じ、辞書フォーマットにエントリされた該動詞の語尾の完全セットが、新たな動詞の語尾の推定できる完全セットとして採用される。動詞活用変化タイプ(ロシア語の1、2型)のより正確な決定のため、そしてそれぞれその語尾の完全セットを明示するために、テキスト指標付けの間に、その動詞が三人称、複数の形態を有する文章、即ち、“- /- ”(第1活用)または“- /- ”(第2活用)の人称語尾を有する該動詞により表現された述語と等位(同格)の複数形の名詞(代名詞)により表現された主語を有する文章、が見つけられる。該動詞の人称語尾に従って、基本辞書フォーマット内で、同一の三人称、複数の語尾を有する動詞が見つけられる。その動詞の語尾の完全セットが、新たな動詞の語尾の完全セットとして採用され、その語幹と一緒に新語辞書フォーマットに書き込まれる。新たな動詞語幹の確率的インデックスの形成後、該情報が全て新語辞書フォーマットに書き込まれる。

20

30

## 【0072】

テキスト指標付け中、確率的に指標付けされた基本語と新語の辞書類のデータベース7に存在しない異なる語形の新しい語が現れた場合、この新語の語幹とその語尾のある特定のセットが、テキスト系文書の確率的指標付けおよびテキスト断片の選択のサブシステム2における該語形との比較のために切り離される。その後、新語の語幹の確率的インデックスが形成され、一緒にその語尾が確率的に指標付けされた基本語と新語の辞書類のデータベース7内の新語辞書フォーマットの中にエントリされる。この語の語形の該セットが処理され、従って辞書フォーマットがその各種の語尾を収容し終わった後、指標付けされた基本辞書の表にアクセスする。既に充填が行われた該辞書は、汎用語のインデックスおよび語幹、ならびにさらに種々のスピーチ・パートの語尾、および所定の語に関係し、その語形変化または活用変化により生ずるそれらの語尾タイプの全種類を、スピーチ・パートの特徴の表示と一緒に含んでいる。辞書になされたリクエストは、その語の語幹の確率的インデックス、語幹それ自体、ならびにさらに利用可能な全種類の語形語尾を含む。確率的に指標付けされた基本語と新語の辞書類のデータベース7において、辞書フォーマットを用いて、語尾の完全セット内で同じ語尾を有する語が見つけられる。これは、新語が同じ語尾を有する辞書内の語と同じスピーチ・パートに属することを意味する。新語のスピーチ・パートが決まった後、リクエストに含まれている全情報が、確立されたフォーマットに従って新語辞書にエントリされる。同時に、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4およびテキストの等価変換のサブシステム5が、

40

50

上述した手順に従って、該断片を、まず、プロダクションルールの単一の確率的に指標付けされた意味構造(2)に変換し、次いで、「...なら (if, 条件)、...である (then, 結論)」の形態のプロダクションルールのフォーマットに変換する。

【0073】

その結果、その条件部が語尾の予め決められたセットを含み、結論部が条件部に列挙された語尾を有する語のスピーチ・パートの名称を含むプロダクションルールが導出され；さらに辞書フォーマットにおいては、その語の語形変化または活用変化により生ずる語形の変化を規定する語尾の完全セットが導出される。結論部はさらに、このスピーチ・パートが語形変化または活用変化を受けた時のその語形に対する質問を含み、その質問は、辞書フォーマットにより決められる手順に従って配列される。

10

【0074】

かくして、それらの種々の語形で提示された新語を含むテキストのプロセッシングの間に、その新語のスピーチ・パートが自動的に決定され、新語は確率的に指標付けされた基本語と新語の辞書類のデータベース7内の辞書フォーマットにエントリされ、システムは形態素解析ルールに対して命令される。新語が新語辞書フォーマットに存在しない場合にその新語のスピーチ・パートおよび特徴を決定するために、これらのルールは、後で説明する手順に従って確率的に指標付けされて、確率的に指標付けされた基本辞書フォーマットと一緒に使用される、形態素解析の知識ベース12に格納される。

【0075】

言語テキストの形態素解析および確率的指標付けを行い、形態素解析の知識ベース12、確率的に指標付けされた言語テキストのデータベース8、ならびに確率的に指標付けされた基本語と新語の辞書類のデータベース7を作成した後、本方法は、統語解析ルールについてシステムを命令すると同時に行われる、所定テーマに関係するテキストの確率的指標付けに進む。

20

【0076】

統語解析ルールに対するシステムへの自動命令は、確率的に指標付けされた言語テキストのデータベース8内で文章統語解析手順を規定する断片を探索するために、自己命令および知識抽出モード制御サブシステム3からの指令により行われる。まず、上記手順に従った断片を、一般に式(2)の形態をとる、プロダクションルールの確率的に指標付けされた1セットの意味構造に変換する。

30

【0077】

次に、文章統語解析手順を記述する得られたプロダクションルールの確率的に指標付けされた意味構造(2)を使用して、論理的結論サブシステム6が論理的結論を実現して、新たなプロダクションルールの確率的に指標付けされた意味構造を生ずる。これらの意味構造は、語の形態素的特徴に基づく文章統語解析を明示するプロダクションルールの導出中に、統語要素を予め決められたスピーチ・パートにリンクさせる。導出されたルールは、確率的指標付けされ、かつインデックス表の形態で表示される統語解析データベース12に格納される。

【0078】

上述したように、テキスト統語解析は、その実行の手順の決定により開始し、その手順は、所定言語の文法に関する教育方法論テキスト系文書に記述されている。該テキストから統語解析手順を規定する知識を抽出するために、自己命令および知識抽出モード制御サブシステム3がまず確率的に指標付けされた言語テキストのデータベース8へのリクエストをコンパイルして、教育方法論テキストのインデックスの表にアクセスする。所定言語で「統語解析手順」という句を包含している該リクエストに応じて、該テキスト中において、上記の句と統語解析シーケンスを規定する用語を含む段落が見つけれよう。

40

【0079】

適当な教育方法論援助 (educational-methodological aids)からのテキスト断片のプロセッシングが終了すると、例えば、ロシア語の場合、下記のプロダクションルールが生じるかもしれない。

50

## 【0080】

「ある文章を統語解析すべき (to be parsed) なら、統語解析手順は次の通りとなる：ワード・コンビネーション (複雑又は複合関係 <the complex or compound relationship>)、単文 (主語、述語、限定語 <attribute>、目的語、副詞)、単文のタイプ (叙述文、疑問文、命令文)、文章構造 (1 もしくは 2 メンバー、拡張 <extended> または非拡張)、述語 (単純 <simple>、複合準動詞 <composite verbal>、複合名詞的語句 <composite nominal>)、同質の複数メンバーを有する文章、別個の複数メンバーを有する文章、直接語法文章、重文、1 つの従属節を有する複文、複数の従属節を有する複文、接続詞のない重文、異なる種類の関係を有する重文」。ワード・コンビネーションのインデックス  $\{I_{i}^{(su)}\}$  に基づいて、式 (2) としてこのルールを公式化した後、予め決められた長さの単一 2 進組み合わせとして、プロダクションルールそれ自体  $\{I_{i}^{(pp)}\}$  の確率的インデックスが形成される：

$$I_{i}^{(pp)} = F(I_{1}^{(su)} I_{2}^{(su)} \dots I_{m}^{(su)} I_{1}^{(su)} I_{2}^{(su)} \dots I_{n}^{(su)}) \quad (3)$$

F は、プロダクションルールの確率的変換のファンクションである。

## 【0081】

次に、プロダクションルール (3) の結論部に述べられた用語のそれぞれが、確率的に指標付けされた言語テキストのデータベース 8 への適当なリクエストをコンパイルするために、順に開示される。出てくる結果は、ルール  $\{I_{i}^{(pp)}\}$  が含んでいる統語的用语のそれぞれを規定する複数の  $\{I_{ij}^{(pp)}\}$  ルールであろう。条件部または結論部において同一の統語的用语を含むプロダクションルール間の関係を用いて、サブシステム 6 が論理的結論をもたらす。これは、下記のプロダクションルールの論理的関係の順序の形成を生ずることになる：

$$I_{i}^{(pp)} \quad \{I_{i1}^{(pp)}\} \quad \{I_{i2}^{(pp)}\} \quad \{I_{i3}^{(pp)}\} \quad \{I_{ik}^{(pp)}\} \quad (4)$$

ここで、インデックス  $\{I_{ij}^{(pp)}\}$  は、ルール  $I_{i}^{(pp)}$  において予め決められた統語解析のあるレベルに関する 1 セットのルールを意味する。例えば、これは、ワード・コンビネーション (複雑又は複合関係)、単文 (主語、述語、限定語、目的語、副詞)、単文のタイプ (叙述文、疑問文、命令文) 等であることができる。

## 【0082】

かくして、本システムは、統語的用语類を特定の語のスピーチ・パート、それらの特徴に結びつけて、該ルールの下で連続統語解析を実施することを目的とする、演繹的な論理的結論を実現する。例えば、ロシア語の場合、該論理的結論の間に、「主語」という用語について下記のテキスト断片が見出されるかもしれない：「文章中の主語は下記の語により表現されることができる：主格の名詞、主格の代名詞、不定詞、単一のワード・コンビネーション」。得られたテキスト断片を通訳プログラム 4、テキストの等価変換のサブシステム 5、および論理的結論サブシステム 6 に送る。該変換は、式 (2) を用いて、語の形態素的特徴を文章メンバーの名称にリンクする下記の 1 セットのプロダクションルールを与える：

「文章が主格の名詞である語を有するなら、その語は恐らく主語である」。

## 【0083】

「文章が主格の代名詞である語を有するなら、その語は恐らく主語である」。

「文章が不定詞である語を有するなら、その語は恐らく主語である」。

「文章が単一のワード・コンビネーションである語を有するなら、その語は恐らく主語である」。

## 【0084】

任意のワード・コンビネーションおよび別々の文章メンバーを規定するプロダクションルールを形成するためにテキスト断片を抽出する間に、文章の語の形態素的特徴が初期情報として機能する。この初期情報に従って、同一のワード・コンビネーションを通して、該情報を文章メンバーの名称を引用している想定されうる応答にリンクするテキスト断片が選択される。これらのワード・コンビネーションは、初期の形態素的特徴を有する語に対応する。

10

20

30

40

50

## 【0085】

従って、所定の形態素的特徴を持つ語と文章メンバーとの間の関係を規定する選択されたテキスト断片は、それが上述した手順に従って正しいことを条件に、確率的に指標付けされた意味構造(2)に移すことができる。その後、確率的に指標付けされた意味構造(2)が、次のプロダクションルールフォーマットで表されよう：「...なら (if, 条件)、...である (then, 結論)」。該手順は、通訳プログラム4、テキストの等価変換のサブシステム5およびプロダクションルールを用いて行われる。ルールの条件部は、最初の語の形態素的特徴を含み、結論部はその語に対応する文章メンバーの名称およびその語に対応する質問を含む。

## 【0086】

その結果、主要な文章メンバー(主語および述語)、補助的な文章メンバー(限定語、目的語、副詞)、ならびにそれらにより作られたワード・コンビネーションの決定のためのプロダクションルールが導出されよう。述語の決定は、そのタイプ：単純、準動詞、複雑名詞的語句を特定する。最初に、文章の叙述ベースが決められ、ここで、主語と述語が調和され、また他の語の組み合わせおよびそれに対応する関係もそうである。それらは、主語と限定語(修飾語)、述語と目的語、述語と副詞などを含む。

## 【0087】

かくして、文章の統語解析時のテキスト系情報のプロセッシングの間に、主要および従属(補助)的文章メンバーを決定するためのルールについてシステムの自己命令が起こる。それにより導出されたルールは、統語解析の知識ベース13に格納される。その後、統語解析手順に従って、システムは別々の文章メンバーを決定するルールについてのその自己命令を開始する。ここでは、初期データは次の通りである：スピーチ・パート、文章メンバーおよびそれらの特徴、それらは、テキストの変換後に、プロダクションルールの条件部に含められよう。これらのルールの結論部は、1群の別々のメンバーのタイプ、文章メンバーの名称、およびそれらが従う質問を規定する。

## 【0088】

こうして、独立した呼応限定語句(separate agreed attributes)(分詞構文、従属語のついた形容詞)、独立した非等位(non-coordinated)限定語句、独立した同格語句、独立した目的語、独立した副詞等が、それらに対応する質問も含めて記述される。

## 【0089】

その後、自己命令モードが、所定文章中の語が何の文章メンバーであるか、それらが何のワード・コンビネーションおよび独立した文章メンバーグループを構成するかを決定する初期データに基づいて、その単文の統語解析を可能するプロダクションルールを導き出す。結果は、その所定文章が1つの2混成文と1つの単文のいずれであるか(単文なら、その文章が関係するタイプ：不定人称、非人称、主格形など)を決定することができるプロダクションルールであろう。その結果として、類似メンバーを持つ文章、独立した文章メンバー、直接話法文章が選択される。

## 【0090】

次に、選択されたテキスト断片に基づいて、複合文章の統語解析のためのプロダクションルールが導き出される。この場合、該プロダクションルールが含んでいる初期データは、複合文章の中に含まれている単文のタイプと特徴である。このルールの結論は、その所定の複合文章が関係する下記タイプの決定を可能にする：重文、1つの従属節を有する複文、複数の従属節を有する複文、接続詞のない重文、多様な種類の関係を有する重文。このルールの結論はまた、何の質問が所定の複合文章内の単文のそれぞれに対応するかも規定する。

## 【0091】

プロダクションルールを生成させる上述したすべてのレベルが、論理式(4)の形で自己命令および知識抽出モード制御サブシステム3からの指令に従って自己命令モードの初めに構成される文章解析パターンに従う。

## 【0092】

10

20

30

40

50

自己命令モードの実現により得られたプロダクションルールは、統語解析の知識ベース13に格納される。文章統語解析に対するシステムの自己命令は、各文章の解析により所定トピックに関して初期テキストをプロセッシングしているまさにその最中に行われることに留意されたい。該解析は、テキストの各文章のフレームの第5～第10レベルの充填を可能にし、次にこのフレームを使用して所定テキストのインデックス表を充填し(表2)、そのセルのそれぞれの内容である上記リストも充填する。次に、統語解析の知識ベースが確率的に指標付けされ、インデックスの表として表示される。これは、探索結果を得るのにプロダクションルールに対応する条件のインデックスに従ったランダムアクセスのために、文章解析の効率をかなり改善する。

#### 【0093】

知識ベースの確率的指標付けとそれを文章の文法解析に使用する手順について次に詳細に説明する。

「...なら (if, 条件)、...である (then, 結論)」のフォーマットで確率的に指標付けされたテキストとして表された1セットのプロダクションルールとして知識ベースを生成させた後、各プロダクションルールは確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4に送られる。ここで、確率的に指標付けされた意味構造(2)が再びセットアップされ、この意味構造は所定ルールの全部のワード・コンビネーションを含んでいる：

$$S : \{(I_i^{(u)} | I_i^{(r)} | I_i^{(z)}) (I_j^{(u)} | I_j^{(r)} | I_j^{(z)})\} \quad (5)$$

各ワード・コンビネーションには対応するインデックス  $I_i^{(su)}$  が割り当てられ：

$$(I_i^{(u)} | I_i^{(r)} | I_i^{(z)}) (I_j^{(u)} | I_j^{(r)} | I_j^{(z)}),$$

次に、該インデックスに基づいて、プロダクションルールのそれぞれの独自の確率的インデックス  $I_i^{(pp)}$  が式(3)に従って形成される。

#### 【0094】

さらに、慣用のテキスト系文書の指標付けと同様に、テキスト形態で所定の知識ベースについてインデックス表が作成される。段落(パラグラフ)として、インデックス  $I_i^{(pp)}$  を有するプロダクションルールが採用される。従って、プロダクションルールのインデックス表にエントリされるのは、プロダクションルール辞書の語の  $\{I_i^{(u)}\}$  語幹(プロダクションルールが含んでいる複数の重複しない語幹)を含んでいる行(ロウ)である。あるインデックス  $(I_i^{(u)})$  に対応する行の各セルは、ワード・コンビネーションのインデックス  $I_i^{(su)}$  と、プロダクションルール中の所定の語、語の語尾および数詞を含んだルールのインデックス  $(I_i^{(pp)})$ 、ならびに、それぞれ所定ルールにおける先行および後続の語のインデックス  $(I_{i-1}^{(u)})$  および  $(I_{i+1}^{(u)})$  を含んでいる。それは、テキスト系文書の場合と同様に、インデックスに基づいて任意のプロダクションルールのテキストをコンパイルするのを可能にする。下記式が確率的に指標付けされたワード・コンビネーションのデータベース15に表の行として書き込まれる：

$$I_i^{(su)} : (I_i^{(u)} | I_i^{(r)} | I_i^{(z)}) (I_j^{(u)} | I_j^{(r)} | I_j^{(z)})$$

プロダクションルールテキストを参照するための初期データが、解析された文章フレームから取られる。上に説明したように、形態素解析後のこのフレームは、それぞれ語幹のインデックス  $\{I_i^{(u)}\}$ 、文章内容の語、所定の語に対応するスピーチ・パートおよび特徴、ならびにそれらへの質問、を含んでいる4行のレベルを有する。各種組み合わせのまさにこの情報が、プロダクションルールの条件に含まれていて、論理的結論に基づいて、所定のスピーチ・パートが関係する文章メンバーが何であるか(正確にまたは近似的に)を決定することを可能にする。プロダクションルールのインデックス表は、文章フレームの語幹のインデックス  $\{I_i^{(u)}\}$  に従って、ならびにルールの条件部もしくは結論部のワード・コンビネーションの値  $\{I_i^{(su)}\}$  に従って、参照される。

#### 【0095】

プロダクションルールを用いた論理的結論ファンクションを遂行するために、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4を使用する。それにより、プロダクションルールが確率的指標付けされた意味構造(2)に変換される。プロダ

10

20

30

40

50

クシヨナルールの条件部のワード・コンビネーション( $I_i^{(su)}$ )に従って( $I_i^{(su)}$ インデックスに従った確率的に指標付けされたワード・コンビネーションのデータベース15の参照後、および所定のワード・コンビネーションの語幹の確率的インデックス $\{I_i^{(u)}\}$ の決定後に)、文章フレームの対応するセルを検索することができ、該セルから語の名称、スピーチ・パートの特徴およびそれへの質問を読み出すことができる。結論部のワード・コンビネーション $\{I_j^{(su)}\}$ に応じて、文章フレームの5~10レベルの対応するセルが充填されなければならない、それらのセルは文章メンバーの名称、それらのグループ、独立したメンバー、複合文章中の単文のタイプ、明示されるそれらへの質問を決定する。それにより、プロダクシヨナルールはその条件部の全部のワード・コンビネーションに関してチェックされ；そして、論理項"and"によりリンクされた全部の条件部ワード・コンビネーションが正しい(そして、プロダクシヨナルールの条件部ワード・コンビネーションに記述されたすべての特徴およびデータが見出された)なら、その条件は正しいものと見なされる。ルール条件のワード・コンビネーションにおいて決定されたデータを、レベル5~10の文章フレームの対応するセルにエントリする。結論部が仮の結果またはワード・コンビネーションを有する場合には、見出されるべき論理的につながるルールがどれであるかに従って、その探索が、ワード・コンビネーションの語幹のインデックスに関して、対応する知識ベースのインデックスの表を参照することにより行われる。確率的インデックスに基づいて表にランダムアクセスするので、プロダクシヨナルールの全セットを吟味する必要性は回避されよう。こうして、プロセッシングに関係するプロダクシヨナルールの数に対する論理的結論時間の依存の線形性が保証される。知識ベースの参照とプロダクシヨナルールのプロセッシングは、文章フレームの全部のセルに正確なデータを充填するためである。

10

20

## 【0096】

もし統語解析がいくつかの文章メンバーを不正確に決定した場合には、それらを正確に決定するために、本システムはこれらの文章の語の意味解析に進み、同時に意味解析ルールに対してシステムの自己命令のモードが実行される。まず、これは、主語、前置詞つき名詞により表現された限定語句および副詞、副詞的分詞構文等の決定に関する。

## 【0097】

文章メンバーの正確な決定のために、ISLSEKにより導出された機能に基づいて、多様なオブジェクト間の関係の全ての可能な種類を記述する段落および文章をテキストから選択する意味解析が使用される。該機能を実行するためのシステムのリクエストは、統語解析が検討した文章中のスピーチ・パートが何の文章メンバーであるかを正確に決定できなかった時に、自己命令および知識抽出モード制御サブシステム3において自動的に生成させることができる。

30

## 【0098】

そのために使用されるのが、自己命令および知識抽出モード制御サブシステム3、論理的結論サブシステム6、および確率的に指標付けされたテキストとプロダクシヨナルールの通訳プログラム4である。統語解析が文章メンバーを正確に決定できなかった場合のそれらの正確な決定は、複数のテキストから所定の複数オブジェクト間の関係を記述する文章を選択することと、それらの間の関係のタイプを決定することとに基づく。システムリクエストの自動生成と選択された文章の意味解析は、確率的に指標付けされたテキストの通訳プログラム4を使用して、予め決められた複数オブジェクト間の関係の下記のタイプを決定することができる：

40

- 性と相 (gender-aspect) ;
- 集合 (部分 - 全体) ;
- 目的語関係 ;
- 定義的關係 ;
- 副詞的 ;
- 許容、非許容。

## 【0099】

50

副詞的關係はさらに下記の種類に再分類される：

- やり方(手段)；
- 場所；
- 時間；
- 尺度または程度；
- 原因；
- 目的；
- 条件；
- 譲歩。

#### 【0100】

テキストにおいて、オブジェクト間の該關係は、各文章の叙述ベース(この叙述ベースは主語と述語からなる)により記述され、さらに、種々の文章メンバーの間のワード・コンビネーション、そしてまず第一に述語と副詞との間の關係(副詞的關係)または目的語との間の關係(目的語關係)を記述するワード・コンビネーション、によっても記述される。關係のタイプを分類するには、述語とそれに関係する目的語または副詞とを含むワード・コンビネーションが決定的な役割を果たす。それは該2つの文章メンバーの内容であり、それにより、ある文章に存在する、主語とさらに目的語または副詞により述べられる複数のトピック・オブジェクト(topical objects)間の關係のタイプが決定される。属性關係(attributive relations)は、(文法的)呼応または非呼応の限定語句(agreed or non-agreed attributes)を含むワード・コンビネーションを用いて、主語、目的語または副詞の性質を記述する。文章メンバーの解析では、それにより記述された關係のタイプの分類によって、統語解析では不正確な結果を生じてしまう、最も込み入った場合でも、實際上正確に文章メンバーを規定することが可能になる。

#### 【0101】

ワード・コンビネーションの關係タイプの分類のために、制御サブシステム3が発する指令に従って、確率的に指標付けされた言語テキストのデータベース8の参照辞書類のインデックスの表から、通訳プログラム4に、上述した關係のそれぞれの典型的なワード・コンビネーションの確率的インデックスが書き込まれる。意味解析の間に、検討されたワード・コンビネーションのそれぞれが、参照辞書テキストのインデックスの表に従った論理的結論を用いて、かつ確率的に指標付けされた意味構造を生成させることにより、通訳プログラム4に書き込まれたワード・コンビネーションのインデックスの1つと相関される(相互に関係づけられる)。インデックスのテキスト表に従った論理的結論を導出する手順は、後で、システム応答の確率的に指標付けされた意味構造をセットアップするプロセスの記載において説明する。

#### 【0102】

一般に、下記の五つの情報源が語およびワード・コンビネーションの意味解析のために使用される：

- 典型的リクエストをプロセッシングするためのISLSEKの動作中に形成された、「リクエスト-リプライ」型のテキスト系要素を含む知識ベース9(該データベースについては後でより詳しく説明する)；
- ある所定テーマのオブジェクト類とそれらの間の關係についての知識を抽出することができる、参照辞書、百科辞典および汎用および専門用の基本的科学方法論的参考文献のインデックスの表を含む確率的に指標付けされた言語テキストのデータベース8；
- 意味解析に要求される文章メンバーの変換の等価性を保証し、入ってくるリクエストに対する生成応答の適合性を評価するために、文章メンバーの正確な決定のためのルールを含む意味解析の知識ベース14；該ベースについては後でより詳しく説明する；
- スピーチ・パートおよびその等価変換語句を決定するためのルールを含む形態素解析の知識ベース12；
- スピーチ・パートおよびその等価変換語句を決めるためのルールを含む統語解析の知識ベース13。

10

20

30

40

50

## 【0103】

上記知識ベースの最初のもは、ユーザー・リクエストのプロセッシングの間に生成した確率的に指標付けされた簡単な応答に基づいて創出され、「リクエスト・リプライ」型の複数のテキスト系要素を含んでいる。この知識は、ユーザー・リクエストに対する適切な応答(リプライ)の意味的基礎を表し、いくつかの問い合わせ文章を含んでいる。これらの文章のそれぞれは、対応するユーザー・リクエストに同一であり、その文章は、疑問語(または疑問句)の後に、さらにそれに対応する応答のワード・グループを含んでいる。このグループは、1または2以上のワード・コンビネーションを含んでいてもよく、1グループの独立した文章メンバーまたは従属節を表す。該知識の各要素において、応答ワード・グループへの質問は正確に決定され、それにより、ある文章において表されるトピック・オブジェクト間の関係を分類し、従って、ある応答ワード・コンビネーションにおいて主要な語が何の文章メンバーであるかを決定することが可能となる。

10

## 【0104】

言語テキストのデータベースは、複数の確率的に指標付けされたテキスト、参照辞書、百科事典、汎用および専門用の両方の基本的科学方法論的参考文献により表される。それらは、あるトピックにおける汎用語彙および専門用語の詳細な記述を含んでいる。インデックスの表として表示されるこれらのテキスト系材料は、そこに含まれている知識の抽出のため、そしてあるトピックにおける種々の主題の主な性質と、それらを上述した分類システムと関連させることによるそれらの間の関係とを特性表示するために使用される。

## 【0105】

意味解析の知識ベース14は、最初の2つの知識ベースに含まれている論理的結論および情報を使用して、テキストの意味解析に役立てるために自動的に導出されたプロダクションルールからなる。

20

## 【0106】

形態素解析および統語解析の知識ベースは、意味解析におけるテキストの等価変換のために使用される。この等価変換プロセスについて、後でより詳細に説明する。

知識の妥当なプロセッシングを確保するため、上記第1のベースは、そのエントリ(入力)が「リクエスト・リプライ」知識に述べられた語の語幹を含んでいるインデックスの表として表示される。この表の各行(ロウ)は、テキスト・インデックス、所定の文章が述べられている段落インデックスおよび番号、文章内の語数、語の語尾、ならびに文章内のその前および後の語のインデックス、を含むセルを有する。これは、システムのリクエストにより、語幹インデックスを用いて、対応する表の行へのランダクアクセス、要求されたセルのそれからの切り離し、および必要に応じて、対応する「リクエスト・リプライ」の初期テキストのリカバリを可能にする。

30

## 【0107】

この知識ベースによって、文章統語解析において、例えば、副詞から直接目的語または間接目的語をそのタイプの正確な分類と共に識別する、等のために、最も込み入った場合における文章メンバーを決定することが可能となる。そのために、意味解析システムは該知識ベースに対する適当なリクエストを生成する。第1の場合では、主語を特定すべき場合(例えば、The rain soaked the umbrella<雨濡らした傘>または The umbrella the rain soaked<傘濡らした雨>のような文章において)、システムのリクエストに応じて、述語により表現される関係が許容されるオブジェクトが何であるか決定される。かくして、許容される関係に対応するオブジェクトが主語として採用されることが明らかとなる。

40

## 【0108】

知識ベースが該リクエストに対する回答を提供することができない場合は、このトピックに関する第2の知識ベースの複数のテキスト系文書の全部における種々のオブジェクト間の要求された関係を含むワード・コンビネーションを探すために、質問は所定の主題に属するテキストのインデックスの表に向けられよう。

## 【0109】

第2の場合では、知識ベースに対するシステムのリクエストに基づいて、目的語と副詞

50

の両方として考えられるかもしれない文章メンバーが回答しているのが何の質問であるかを決定し、こうして、所定の語が何の文章メンバーであるかを正確に確かめるべきである。そのためには、知識ベースにアドレス指定されたシステムのリクエストにおいて、要求された語およびそれに対する想定される回答を列挙する。知識ベースが対応する「リクエスト - リプライ」を有し、その場合の応答(リプライ)のワード・コンビネーションにおいて、主要な語とそれに対する質問が結果的にシステムリクエストの内容と一致するならば、それは、解析された文章メンバーがその質問に正確に回答していることを意味しよう。従って、システムリクエストのプロセッシングの結果は、その特定の語が何の文章メンバーであるかを正確に決定することを可能にする。例えば、"A man's taking a walk in a park"(公園での男の散歩)または"A man's taking a walk in a suit"(スーツを着た男の散歩)なる文章の解析において、文中のin a park(公園での)またはin a suit(スーツを着た)のワード・コンビネーションが何の文章メンバー(副詞または目的語)であるかを特定する目的で、2つのシステムリクエストが生成されよう。統語解析により既にin a parkが場所の副詞的修飾句であることが不正確に結論づけられているので、第1の場合は、疑問語where?(どこ?)とワード・コンビネーションin a parkを含んでいよう。第2の場合は下記のシステムリクエストが生成されよう: in what?(何を着て?) - in a suit。システムリクエストのプロセッシングがこれらのそれぞれに肯定の回答を生じたならば、それは第1のワード・コンビネーションが正確に副詞であり、第2が目的語であることを意味する。システムのリクエストが誤った表明(例、where? - in a suit)を持って生成された場合には、回答は否定になろう。それは、ワード・コンビネーションin a suitが場所の副詞的修飾句ではないことを意味する。

#### 【0110】

意味解析システムの第1の知識ベースに対するリクエストを生成させる上述の方法は、文章の統語解析がより困難な場合にも適用することができる。例えば、副詞的分詞構文が何のタイプの副詞であるかを決定する場合、または従属節のタイプを特定すべき場合である。そのためには、特別のリクエストが生成され、それは、所定の副詞的分詞構文または従属節を含んでおり、該リクエストに基づいて、それらの類似形が「リクエスト - リプライ」型知識のアレイにおいて探索され、正確さは類義語の程度による。該類似形が該データベースの応答ワード・グループの中に含まれている場合には、それらはテキストのインデックスの表を用いてそれから抽出されよう。これにより、この副詞的分詞構文または従属節が対応する質問を規定し、従って、それが属するタイプを正確に確かめることが可能となる。

#### 【0111】

上記第1の知識ベースがリクエストされた類義語を含んでいない場合には、文章メンバーの正確な決定のために、論理的結論サブシステム6と一緒に上記第2および第3の知識ベースが使用される。上述したように、第3の知識ベースは、意味解析を用いて、複文中の文章メンバー、副詞的分詞構文または従属節タイプの名称を、それへの適当な質問を形成するように特定することを可能にするプロダクションルールを含んでいる。

#### 【0112】

この知識ベースを用いた意味解析の実行の主要なバージョンの1つは、その文章メンバーのそれぞれに固有の意味的属性を、1つの決まった語とある1つの基本語(base word)とを含む1セットのワード・コンビネーションに翻訳することである。この基本語は、ある所定の文章メンバーだけに意味的に結合され、そのメンバーに明らかに対応する(他の文章メンバーとは一緒には使用できない)。解析された初期テキストから、プロダクションルールに記述されたワード・コンビネーションを生成させる際には、形態素解析および統語解析の知識ベースのルールに基づいて、論理的結論を使用して、初期テキストの等価変換を行わなければならないことが多い。

#### 【0113】

要求されたワード・コンビネーションが得られた後、これを、要求されたワード・コンビネーションを含む段落および独立した文章を選択することを可能にするデータベースで

ある、第2の指標付けされたテキストのデータベースを参照するために、その許容性(allowability)についてチェックする。複数のテキスト系文書においてそのワード・コンビネーションが使用されている1または2以上の文章が見つかった場合には、そのワード・コンビネーションのワードの間の関係が許容される。その結果、検討した語は正確にある特定の文章メンバーに属すると考えられる。

【0114】

独立したワード・コンビネーションの代わりに、より複雑な構造(例、複文中における分詞、副詞的分詞構文、従属節)を使用してもよい。こうして、複数のテキスト系文書において、特定のワード・コンビネーションにより表現された意味的知識を、その中のワード間の関係の許容性の決定と一緒に組み合わせることにより、その統語解析が正確な結果を与えない時に、正確な文章メンバーを決定することが可能となる。

10

【0115】

テキスト系文書の文章の形態素および意味解析と統語解析が、それにより得られた文章フレームに基づいて完了したら、テキストのインデックスの表が作成され(表2)、それは表の各セルの内容を決定するリスト $\{I_i^{(s)}\}$ を含んでいる。次に、本方法は要求されたテーマに関係する次のテキストの確率的指標付けに進む。同時に、自己命令が実施され、意味解析の知識ベース14に、上述した手順および確率的に指標付けされた意味構造(2)を用いて、対応するテキスト断片に基づいて導出されたプロダクションルールが供給される。各ルールの正しさは、上述した手順により、対応する言語テキストから取られた多数の断片に基づいて、同一の確率的に指標付けされた意味構造(2)が独立して別々に発生することにより得られることに留意されたい。その後、確率的に指標付けされた意味構造が「...なら(if, 条件)、...である(then, 結論)」として表されるプロダクションルールのフォーマットに移される。これは、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4により、テキストの等価変換のサブシステム5においてなされる。

20

【0116】

そのトピックについて全部のテキストをプロセッシングすると、そのトピックに関するテキストのインデックスの表が生ずる(表3)。その行(ロウ)は、テキスト系文書中に含まれた語幹の繰り返されないインデックス $\{I_i^{(u)}\}$ により指定される。この表の列(カラム)は、文法および意味解析中に処理されたテキストの確率的インデックス $\{I_i^{(t)}\}$ に対応する。この表のセルは、各テキストの段落のインデックス $\{I_i^{(a)}\}$ を含むリストのインデックス $\{I_i^{(s)}\}$ を含んでおり、各テキストのインデックス $\{I_i^{(t)}\}$ は語幹の対応するインデックス $\{I_i^{(u)}\}$ を含んでいる。リストにおけるエントリは、適当なインデックス $\{I_i^{(s)}\}$ に従ってアクセスされる別々のファイルに格納される。

30

【0117】

インデックスの該表が作成され、知識ベースがI S L S E Kの自己命令のモードにおいて生成した後、自己命令および知識抽出モード制御サブシステム3の指令により、本方法は、ユーザーのリクエストに適合するテキスト系文書から知識を抽出するためにユーザー・リクエストをプロセッシングすることに進む。

【0118】

このプロセスは、ユーザーのリクエストと知識がそれから抽出される時のテキスト断片の文章の両方の等価変換を広範に使用する。テキスト文章の変換に適用される手順について次に詳述する。

40

【0119】

I S L S E Kはテキストの等価変換の下記のレベルを与える：

第1のレベルの等価変換は、次のような文章メンバーのグループ内で実施される：主語、述語、目的語、副詞を含んでいるワード・コンビネーション。これは、スピーチ・パートを変化させて、文法的呼応語句を非呼応語句に置換する。このレベルは、a computer network - a network of computer(<コンピュータ・ネットワーク> - <コンピュータのネットワーク>)、computer services - service of computers(<コンピュータ・サービス> - <コンピュータのサービス>)のような用語の変換に相当する。

50

## 【0120】

第2のレベルの等価変換は、自律文章と複文を構成する文章の両方を含む、単文内の文章メンバーの等価変換に相当する。共通語根スピーチ・パートの変換を用いた下記タイプの文章メンバーの置換が行われる：

- 主語を述語と置換する；
- 述語を主語と置換する；
- 目的語を主語と置換する；
- 述語を副詞と置換する、等。

## 【0121】

特定の場合には、スピーチ・パートが変化しないこともある(変化するのは文法の格だけである)。

第3のレベルの等価変換は、複文内の等価変換に相当する。この場合、あるタイプの従属節が、別のタイプの従属節で、または分詞、動詞的副詞の語法で置換されうる。場合によっては、複文が、接続詞をルールで決まる適当な前置詞で置換することによって単文に変換される。

## 【0122】

ワード・コンビネーションにおける文章メンバーの置換を用いた等価変換の例、即ち、呼応語句の非呼応語句による置換と、直接目的語の主語による置換、を次に考える。初期文章は"Software and hardware means protect the computer programs"(ソフトウェアおよびハードウェア手段がコンピュータプログラムを保護する)である。システムにおいて、 $\{I_1^{(p)}\}$ インデックスを持つ初期文章が下記の確率的に指標付けされた意味構造により表示されよう：

$$I_1^{(p)} : I_{12}^{(su)} \quad I_{13}^{(su)} \quad I_{14}^{(su)} \quad (5a)$$

この構造は初期文章の下記のワード・コンビネーションを有する。

- $I_{12}^{(su)} = (\text{software and hardware means}) <\text{ソフトウェアおよびハードウェア手段}>$ ,
- $I_{13}^{(su)} = (\text{protect}) <\text{保護する}>$ ;
- $I_{14}^{(su)} = (\text{the computer programs}) <\text{コンピュータプログラム}>$ 。

## 【0123】

上述した文章メンバーの等価変換により下記のワード・コンビネーションの創出が可能となる：

- $I_{22}^{(su)} = (\text{programs of a computer}) <\text{コンピュータのプログラム}>$ ,
- $I_{23}^{(su)} = (\text{are protected}) <\text{保護される}>$ ;
- $I_{24}^{(su)} = (\text{by software and hardware means}) <\text{ソフトウェアおよびハードウェアにより}>$ 。

## 【0124】

これらの変換は、インデックス  $I_1^{(p)}$  を持っていた初期文章に等価である別の文章を生ずることになり、この文章はインデックス  $I_2^{(p)}$  を持ち、さらに下記の確率的に指標付けされた意味構造も有する：

$$I_2^{(p)} : I_{22}^{(su)} \quad I_{23}^{(su)} \quad I_{24}^{(su)} \quad (5b)$$

この構造に基づいて、下記の文章が生成するであろう："Programs of a computer are protected by software and hardware means"(コンピュータのプログラムはソフトウェアおよびハードウェア手段により保護される)なる文章。この文章は初期文章に等価である。この新たな文章において、主語  $I_{22}^{(su)}$  は初期文章の直接目的語の語句  $I_{14}^{(su)}$  のワード・コンビネーションに対応し、呼応語句は非呼応語句に置換されていることが認められよう。ここで、最初の文章の主語  $I_{12}^{(su)}$  は第2の文章では間接目的語  $I_{24}^{(su)}$  に変換され、述語  $I_{13}^{(su)}$  は受動(再帰)動詞  $I_{23}^{(su)}$  の形態を獲得している。該変換は、テキストおよびユーザー・リクエストの両方の確率的に指標付けされた文章の等価変換に最もよく使用される。

## 【0125】

ユーザーのリクエストは、自然言語でコンパイルされる。その後、ユーザー・リクエ

トは、疑問語と、リクエストの意味を規定し、元のリクエストに等価な、ワード・コンビネーション、とを含む複数の新たなリクエストに変換される。元のリクエストの該等価変換は、類義語、近似意味の語、ならびにスピーチ・パートおよび文章メンバーの置換を用いて実施される。こうして、リクエストの問い合わせ文章のワード・コンビネーションの等価構造を得るために形態素、統語および意味の各解析の確率的に指標付けされたルールを適用することと、ワード・コンビネーション間の意味的リンクを維持することによって、元のリクエストの意味がそのまま保たれる。

#### 【0126】

その後、変換されたユーザー・リクエストに従って、リクエストの全部のワード・コンビネーションを有するテキスト系文書の断片が予備選択される。このリクエストがこれらの要件に合致するテキスト系文書の断片の予備選択の可能性を与えることができない場合には、リクエストの新たな等価変換が行われる。

10

#### 【0127】

種々のテキスト系文書、段落および文章に基づくリクエストのプロセッシングと応答生成のアルゴリズムの手順を次に説明する。言語プロセッサ1においてユーザーのリクエストを受け取ると、リクエストがテキスト系文書の確率的指標付けと切り離しのサブシステム2にエントリされ、そこで語幹の確率的インデックスが形成され、それらの語尾が切り離される。次に、確率的に指標付けされたリクエストが、自己命令および知識抽出モード制御サブシステム3を介して、論理的結論サブシステム6に書き込まれる。そこで、知識ベース12、13のプロダクショナルルールに基づいて、リクエストは形態素解析および統語解析を受ける。こうして、問い合わせ文章フレームが作成される。その後、通訳プログラム4は、その問い合わせ文章を主要な語と従属的な語とを有する1セットのワード・コンビネーションとして、該ワード・コンビネーションに対応する語幹の確率的インデックスと共に提示する。

20

#### 【0128】

$$S : \{ ( |_{i}^{(u)} |_{i}^{(r)} |_{i}^{(z)} ) ( |_{j}^{(u)} |_{j}^{(r)} |_{j}^{(z)} ) \} \quad (6)$$

式中、 $|_{i}^{(u)} |_{j}^{(u)}$  は、あるワード・コンビネーションにおける主要な語と従属的な語の語幹の確率的インデックスである。 $|_{i}^{(r)} |_{j}^{(r)}$  は、該ワード・コンビネーションにおける主要な語および従属的な語のスピーチ・パートの確率的インデックスである。 $|_{i}^{(z)} |_{j}^{(z)}$  は、それぞれ、該ワード・コンビネーションにおける主要な語および従属的な語のスピーチ・パートの形態素的および統語的特性の確率的インデックスである。

30

#### 【0129】

得られたインデックスに基づいて、リクエストの確率的に指標付けされた意味構造が生成し、この構造は一般に下記のように表現される：

$$P : |_{1}^{(su)} |_{2}^{(su)} |_{3}^{(su)} |_{4}^{(su)} |_{5}^{(su)} \quad (7)$$

式中、 $|_{1}^{(su)}$  は疑問語のワード・コンビネーションのインデックスであり、 $|_{2}^{(su)}$  は主語(部)のワード・コンビネーションのインデックスであり； $|_{3}^{(su)}$  は述語(部)のワード・コンビネーションのインデックスであり； $|_{2}^{(su)} |_{3}^{(su)}$  は、主語と述語とを結びつける文章叙述ベース(predicative sentence base)であり； $|_{3}^{(su)} |_{4}^{(su)}$  は、文章中の関係のタイプを決める述語と目的語(副詞)との間の関係であり； $|_{4}^{(su)}$  は目的語(副詞)のワード・コンビネーションのインデックスであり； $|_{5}^{(su)}$  は副詞(目的語)のワード・コンビネーションのインデックスである。

40

#### 【0130】

式(6, 7)のインデックスに従って、予め決められたトピックの確率的に指標付けされたテキストのデータベース10を参照し、予め決められたテーマのテキストのインデックスの表(図4)を使用することにより、疑問語ワード・コンビネーションを含むリクエストの全部のワード・コンビネーションを含んだ1セットの断片が見つけれられる。各テキスト断片は、1または2以上の段落からなるものでよい。

#### 【0131】

該条件を満たす1または2以上のテキストが見つかった場合には、本方法はさらに、こ

50

これらのテキストの段落を、それらのそれぞれのインデックスの表を用いてプロセッシングすることに進む。 $I_1^{(su)}$  インデックス(表のセルの1つのリスト $I_1^{(s)}$ における)を持つテキストの1つのインデックスの表において、質問インデックスとそれに関連する主要な語の語幹インデックスとを有する疑問語ワード・コンビネーションの存在が、所定テキスト中の該段落が回答 $I_0^{(su)}$ のワード・グループを含む、ある文章を含んでいることを示している。このワード・グループは、疑問語ワード・コンビネーションの主要な語にリンクされる： $(I_0^{(su)} \quad I_1^{(su)})$ 。

#### 【0132】

該条件と合致する少なくとも1つのテキストが見つからない場合、本方法は、テキスト段落に含まれていなかった語を類義語および近似意味の語により置換し、かつリクエストの意味を変化させずにスピーチ・パートおよび文章メンバーの置換を行うことによる、ユーザー・リクエストの等価変換へと進む。

10

#### 【0133】

該条件を満たすテキストのさらなるプロセッシングは、所定テキストのインデックスの表を参照しながら行われる。そのためには、データベース10におけるテキスト・インデックスの表を参照することにより疑問語ワード・コンビネーションのインデックス $I_1^{(su)}$ を用いて、応答のワード・グループを含む文章が見つけれられる。このワード・グループはリクエストの疑問語ワード・コンビネーションに対応し、リクエストの主要な語にリンクされる。リクエストのワード・コンビネーション

$$S : \{(I_i^{(u)} \quad I_i^{(r)} \quad I_i^{(z)}) \quad (I_j^{(u)} \quad I_j^{(r)} \quad I_j^{(z)})\}$$

20

が各種テキストの異なる段落

$$V : \{(I_i^{(t)} \quad I_j^{(a)})\}$$

に含まれている場合には、応答の単一の論理的にリンクされたテキストを形成するための必要条件は、少なくとも1つの段落における、リクエストの疑問語ワード・コンビネーションの $I_1^{(t)}$ に対応する応答のワード・グループ $I_0^{(su)}$ と、主語と述語のワード・コンビネーションのインデックスを一般的形態で含んでいる式(7)の叙述ベース $I_2^{(su)} \quad I_3^{(su)}$ 、との存在である。該条件が満たされる場合、切り離された段落のセットをさらなるプロセッシングに用いる。なぜなら、仮に選択された段落に基づいて、応答の単一の論理的にリンクしたテキストを形成するように試みることができるからである。そうでない場合には、本方法はそのテーマについての新たなテキストの入力および指標付けに進むべきである。

30

#### 【0134】

まず、リクエストの全部のワード・コンビネーションを含む1つのテキスト断片を、所定テキストの1または2以上の連続した段落に基づいて形成することができるという、適切な応答が生成するより単純な場合を検討する。この場合、まず応答の確率的に指標付けされた意味構造のベースが下記の式として形成される：

$$P : I_0^{(su)} \quad I_1^{(su)} \quad I_2^{(su)} \quad I_3^{(su)} \quad (8)$$

式中、 $I_0^{(su)}$  は応答ワード・グループのインデックスであり； $I_1^{(su)}$  は疑問語ワード・コンビネーションのインデックスであり、 $I_2^{(su)}$  は主語ワード・コンビネーションのインデックスであり； $I_3^{(su)}$  は述語ワード・コンビネーションのインデックスであり； $I_2^{(su)} \quad I_3^{(su)}$  は文章叙述ベースである。そのためには、所定テキスト断片において、疑問語ワード・コンビネーションの主要な語にリンクされた指標付けされた応答ワード・グループ  $(I_0^{(su)} \quad I_1^{(su)})$  を含む文章を決定した後、叙述ベース  $(I_2^{(su)} \quad I_3^{(su)})$  を含む文章を見つける。

40

#### 【0135】

該ワード・グループは一般に異なる種々の表現を含んでいるので、次に意味構造パターン(8)を創出するために、所定テキスト断片の指標付けされた文章を用いて論理的結論の手順を行う。そのためには、応答ワード・グループを含んでいる  $i$  番目の文章を次のように提示する：

$$P : I_0^{(su)} \quad I_1^{(su)} \quad I_{2i}^{(su)} \quad I_{3i}^{(su)} \quad I_{4i}^{(su)} \quad I_{5i}^{(su)} \quad (9)、$$

50

式中、 $I_{0}^{(su)}$  は応答ワード・グループのインデックスであり； $I_{1}^{(su)}$  は疑問語ワード・コンビネーションのインデックスであり、 $I_{2i}^{(su)}$  は主語ワード・コンビネーションのインデックスであり； $I_{3i}^{(su)}$  は述語ワード・コンビネーションのインデックスであり； $I_{2i}^{(su)} \quad I_{3i}^{(su)}$  は文章叙述ベースであり； $I_{3i}^{(su)} \quad I_{4i}^{(su)}$  は、所定文章中の関係のタイプを規定する述語と目的語(副詞)との間の関係であり； $I_{4i}^{(su)}$  は目的語(副詞)のワード・コンビネーションのインデックスであり； $I_{5i}^{(su)}$  は副詞(目的語)のワード・コンビネーションのインデックスである。

【0136】

式(9)に基づく論理的結論を、推移関係(transitive relationship)を用いて達成するために、 $i$  番目の文章についてトピック コメント型の確率的に指標付けされた意味構造を作成する： 10

$$TR : I_{2i}^{(su)} \quad I_{3i}^{(su)} \quad I_{4i}^{(su)} = I_{2i}^{(su)} \quad I_{4i}^{(su)} \quad (10)、$$

式中、トピックは主語ワード・コンビネーションのインデックス $I_{2i}^{(su)}$ であり、コメントは目的語(副詞)ワード・コンビネーションのインデックス $I_{4i}^{(su)}$ である。

【0137】

リクエストの叙述ベースを含んでいる  $j$  番目の文章は一般に下記の確率的に指標付けされた意味構造を有する：

$$P : I_{2}^{(su)} \quad I_{3}^{(su)} \quad I_{4j}^{(su)} \quad I_{5j}^{(su)} \quad (11)$$

式中、 $I_{2}^{(su)}$  はリクエストの主語ワード・コンビネーションのインデックスであり； $I_{3}^{(su)}$  はリクエストの述語ワード・コンビネーションのインデックスであり； $I_{2}^{(su)} \quad I_{3}^{(su)}$  はリクエストの文章叙述ベースであり； $I_{3}^{(su)} \quad I_{4j}^{(su)}$  は、 $j$  番目の文章中の関係のタイプを規定する述語と目的語(副詞)との間の関係であり； $I_{4j}^{(su)}$  は目的語(副詞)のワード・コンビネーションのインデックスであり； $I_{5j}^{(su)}$  は副詞(目的語)のワード・コンビネーションのインデックスである。その後、式(11)は  $j$  番目の文章の下記のトピック コメント意味構造に変換される： 20

$$TR : I_{2}^{(su)} \quad I_{4j}^{(su)} \quad (12)。$$

【0138】

テキスト系情報の複数の完全文章間には意味的なリンク、従って文法的(統語的)リンクがあることに留意されたい。複数文章の構造的相関関係、即ち、それらの間の統語的リンクの2つの方法がある。最初の方法は連結(連続)リンク、第2の方法は並列リンクと呼ぶことができる。 30

【0139】

連結(concatenated)リンクは、リンクされたテキストにおける思考の連続的な展開を示す。トピック(テーマ)は、始点、思考発展の開始、「所定のもの」であり；コメントは思考の展開、その基礎、コア、「新規なもの」である。

【0140】

連結リンクの統語的性質は、2つの隣接している文章の構造上の相関において表現される。一般に、先行側の文章のいずれかの文章メンバー、例えば、目的語、が後続側の文章の主語になる。連結リンクの最も普通の構造タイプは次の通りである：「目的語 - 主語」、「目的語 - 目的語」、「主語 - 目的語」、「主語 - 主語」等。 40

【0141】

連結リンクにおける文章間の構造的相関は、a) 語彙の繰り返し(重複)(文章の相関したメンバー同士が同一に表現される時)；b) 類義語彙；c) 代名詞、により表現される。

連結リンクは自律文章において最も重要かつ広範に使用されるリンクの方法の1つである。

【0142】

並列リンクは、連結リンクと同様に、接続された文章の構造的相関に存する。しかし、この相関の性質が異なる。文章の並列リンクの主な構造上の特色は次の通りである：a) 構造的な並行(接続された文章同士の1つの共通タイプまたは統語的な近さ)；b) 並列(類似)した語順；c) 全てまたは一部の文章メンバーの同一の文法的表示。 50

【0143】

ある段落におけるリンクの連結および並列の両構造への意味論的な「入口」は、その段落の接続された文章のその最初の文章のトピックまたはテキストの連続する段落の数である。

【0144】

かくして、トピック コメント型の各文章の基本的な意味構造に基づいて、論理的結論により、連結タイプと並列タイプの両方の文章間のリンクを規定するより複雑な意味構造を形成することができる。従って、i番目の文章における応答ワード・グループとj番目の文章におけるリクエスト文章叙述ベースとの間の意味的リンクの必要条件是、それらがその所定テキスト断片の単一の意味構造に含まれているという、論理的結論に基づくグループである。確率的に指標付けされた形態では、この構造は次のように提示されていてもよい：

$$\begin{array}{cccccccc}
 | & 0^{(su)} & | & 1^{(su)} & | & 2_i^{(su)} & | & 4_i^{(su)} & | & 4_i^{(su)} & | & 4_k^{(su)} & \dots & | & 2_m^{(su)} \\
 | & 2^{(su)} & & & & & & & & & & & & & \\
 | & 2^{(su)} & | & 4_j^{(su)} & = & | & 0^{(su)} & | & 1^{(su)} & | & 2_i^{(su)} & | & 4_i^{(su)} & | & 4_k^{(su)} \dots | \\
 2^{(su)} & | & 4_j^{(su)} & & & & & & & & & & & & 
 \end{array}
 \tag{13}$$

該ワード・グループ間の意味的リンクの存在を確認するための論理的結論は、所定トピックに関係する確率的に指標付けされたテキスト系文書データベース10のテキストのインデックスの表に従って導出される。そのために、論理的結論サブシステム6および等価変換のサブシステム5が使用される。論理的結論は疑問語ワード・コンビネーションの主要な語、リクエストの叙述ベースにリンクされた応答のワード・グループを含んでいるi番目の文章で始まり、確率的に指標付けされた意味構造(9)を有する。

【0145】

該文章がインデックスの表に従ってトピック コメント型の意味構造(10)として提示された後、その所定文章のコメントが次の文章のトピックに移っている、次の文章が見つけられる。そのために、所定段落のインデックス $|_j^{(a)}$ とi番目の文章の目的語または副詞である $|_{4_i}^{(su)}$ ワード・コンビネーションのインデックスとに対応する、複数のセルが使用される。該セルに従って、所定ワード・コンビネーションが主語を含んでいる所定段落内の文章の番号を見つける。次に、セル・アドレス情報を用いて、該文章の述語のインデックスと目的語もしくは副詞のワード・コンビネーションの関連するインデックス $|_{4_k}^{(su)}$ を見つける、即ち、式(13)に従って、先行文章に論理的にリンクされる次の文章等のコメントを見つける。論理的結論は、関係 $(|_2^{(su)} |_{4_j}^{(su)})$ により規定される次の文章がリクエストの叙述ベースに対応するインデックス $(|_2^{(su)} |_3^{(su)})$ を含むようになるまで続けられる。

【0146】

論理的結論の導出の間に、何かの文章のコメントインデックス $|_{4_n}^{(su)}$ が次の文章のトピック $|_{2_{n+1}}^{(su)}$ と一致しない場合には、それは、該次の文章が、所定の語の類義語または代名詞のいずれかを使用していることを意味する。前者の場合、このワード・コンビネーションの語幹のインデックス $|_{2_{n+1}}^{(su)}$ に従って、確率的に指標付けされた言語テキストのデータベース8の類義語辞書が参照される。そこで、類義語の語幹 $\{|_s^{(u)}\}$ を見つけ、その語幹から次の文章のコメントのインデックス $|_{4_n}^{(su)}$ を見つけることができる。後者の場合、次の文章のトピックのインデックス $|_{2_{n+1}}^{(su)}$ はワード・コンビネーション $|_{4_n}^{(su)}$ に呼応した代名詞に対応するかもしれない、それをデータベース7の辞書のインデックスの表に従ってチェックする。第1または第2の条件が満たされた場合、探索しているリクエストのワード・コンビネーション[この場合、リクエスト叙述ベースの $(|_2^{(su)} |_3^{(su)})$ ]を含む文章が見つかるまで論理的結論を続ける。こうして、論理的結論の間に、式(13)に従った確率的に指標付けされた意味構造が合成されよう。

【0147】

検討している事例では、リクエストの全部のワード・コンビネーションが1つのテキストの1つの段落または1グループの連続した段落の中に収まっているので、所定テキスト

10

20

30

40

50

断片における論理的結論を続けて、目的語 $l_4^{(su)}$ および副詞 $l_5^{(su)}$ のワード・コンビネーションを含む、リクエストの全部のワード・コンビネーションを含むであろう、単一の確率的に指標付けされた意味構造を生成させるようにする：

$$S : l_0^{(su)} \quad l_1^{(su)} \quad l_{2i}^{(su)} \quad l_{4i}^{(su)} \quad l_{4k}^{(su)} \dots l_2^{(su)} \\ l_4^{(su)} \dots l_{2m}^{(su)} \quad l_4^{(su)} \dots l_{2n}^{(su)} \quad l_5^{(su)} \quad (14)$$

そのためには、所定段落の各種文章が含んでいる全部のリクエストのワード・コンビネーションが意味構造(14)の中に含まれてくるようになるまで、トピック コメントの順序で上述した論理的結論ファンクションを遂行する。該意味構造(14)の合成の必要条件は、リクエストのワード・コンビネーションと、テキスト段落中のそれらに同一のワード・コンビネーションとの、同じ文章メンバーとの対応(一致)であることに留意されたい。従って、テキストの文章におけるリクエストのワード・コンビネーションと同一の、何かのワード・コンビネーションが、他の文章メンバーに関係する場合には、これらの文章に対して、該ワード・コンビネーションが要求される文章メンバーに関係づけられるように等価変換を受けさせる。これらのファンクションは、等価変換のサブシステム5において上記手順に従って遂行される。

10

#### 【0148】

意味構造(14)を生成させた後、本方法はそれを無矛盾性についてチェックすることに進む。そのために、意味構造(14)を生成させた文章のそれぞれが含んでいる述語 $\{l_{3i}^{(su)}\}$ のワード・コンビネーションの、ベース関係との意味的相関、をチェックする。このような関係は、性 - 相関係、「部分 - 全体」もしくは「原因 - 結果(条件 - 結論)」型の関係である。これらの関係が、該インデックスに従って、確率的に指標付けされたテキストのデータベース8を参照することにより決められ、インデックスの参照辞書の表における $\{l_{3i}^{(su)}\}$ 述語の意味論的意味(semantic meanings)が探索される。この工程では、この述語の意味論的意味と、上述したベース関係または通訳プログラム(4)で書かれたそれらの類義語の $\{l_{3i}^{(su)}\}$ インデックス、との同一性がチェックされる。これらの条件が満たされた場合、生成させた意味構造(14)において推移関係が維持されている。こうして、 $l_j^{(su)}$ インデックスを有する探索されたリクエストのワード・コンビネーションはいずれも、生成させている応答文章の中に、トピック コメント型の生成させた意味構造における論理的結論を使用して、 $l_{j-1}^{(su)}$ インデックスを有するワード・コンビネーションの後に移すことができる。この条件が満たされない場合には、その所定段落はユーザー・リクエストに適合すると思われる応答を含んでいない。そのような場合には、本方法は、次の予備選択された段落または後続の1セットの段落の解析に進む。

20

30

#### 【0149】

リクエストのワード・コンビネーション間の意味的関係を決定するための上述した論理的結論手順は、該ワード・コンビネーションがその段落の異なる文章において見出された場合には、ユーザーへの簡単な応答が、応答ワード・グループ、疑問語ワード・コンビネーション、叙述ベースおよび応答が含んでいる他の全部のワード・コンビネーションを含んだ1文章として生成されるようになるまで続く。この発生した簡単な応答は、下記の確率的に指標付けされた意味構造として提示されよう：

$$P : l_0^{(su)} \quad l_1^{(su)} \quad l_2^{(su)} \quad l_3^{(su)} \quad l_4^{(su)} \quad l_5^{(su)} \quad (15)$$

40

式中、 $l_0^{(su)}$ は応答ワード・グループのインデックスであり； $l_1^{(su)}$ は疑問語ワード・コンビネーションのインデックスであり、 $l_2^{(su)}$ は主語ワード・コンビネーションのインデックスであり； $l_3^{(su)}$ は述語ワード・コンビネーションのインデックスであり； $l_2^{(su)} \quad l_3^{(su)}$ は文章叙述ベースであり、 $l_3^{(su)} \quad l_4^{(su)}$ は、所定文章中の関係のタイプを規定する、述語と目的語(副詞)との間の関係であり； $l_4^{(su)}$ は目的語(副詞)ワード・コンビネーションのインデックスであり； $l_5^{(su)}$ は副詞(目的語)ワード・コンビネーションのインデックスである。

#### 【0150】

この簡単な応答の正しさは、テキスト系文書の種々の予備選択された確率的に指標付けされた断片に基づいて、上記手順に従って、同一の確率的に指標付けされた意味構造(15)

50

がいくつか生成することにより保証される。

【0151】

生成した意味構造(15)は、論理的結論が、リクエストの問い合わせ文章と同一の簡単な応答を生じたことを意味する。従って、この応答はユーザーのリクエストに適合している。この応答をある所定言語のテキスト系形態に変換させた後、これを、リクエストに従ってシステムにより提供される知識としてユーザーに出力することができる。

【0152】

ユーザーが、より完全な応答を提供してほしいとリクエストしている場合には、本方法は、簡単な応答が生成された段落に基づいて、テキストの最初の段落の変換と、必要なら後続のテキスト段落の変換、に進む。これは、これらの段落に基づいて、所定テキスト断片内で簡単な応答の詳細化の可能性を与える、単一の確率的に指標付けされた意味構造を得るために行われる。完全な応答を生成させるために行われる上述したファンクションについては後で説明する。

10

【0153】

インデックス表の予備的探索がリクエストの全部のワード・コンビネーションを含むと思われる段落を有するテキストを全く見つけることができなかつた場合には、得られたリクエストのインデックスに従って、それらの断片が全体としてリクエストの全部のワード・コンビネーションを含んでいる複数のテキストを探索する。そのようなテキスト断片のセットが見つからない場合には、それは、確率的に指標付けされたテキスト系文書のデータベース10がユーザーのリクエストに適合性のある適切な応答を生成させることができな

20

【0154】

予備選択時には、リクエストのワード・コンビネーションのインデックス  $S : \{I_i^{(u)}, I_j^{(u)}\}$  に従って、テキストのインデックスの表を用い、各テキストについて、リクエストの全部のワード・コンビネーションを含んでいる下記の段落のセットの形態で複数の断片が選択される：

$$V : \{I_i^{(t)}, I_j^{(a)}\}$$

式中、 $I_i^{(t)}, I_j^{(a)}$  は、ユーザーのリクエストのある種のワード・コンビネーションを含むある所定テキストの、それぞれテキストのインデックスと段落のインデックスである。リクエストのワード・コンビネーションのインデックス  $I_i^{(su)} : \{I_i^{(u)}, I_j^{(u)}\}$  が、テキスト  $I_i^{(t)}$  の少なくとも1つのいずれか1つの段落  $(I_i^{(t)}, I_j^{(a)})$  中に完全に含まれることはないが、1つのテキストの異なる段落または異なるテキストの異なる段落の中に含まれていて

30

$V : \{I_i^{(t)}, I_j^{(a)}\}$  となる場合には、

予備選択されたテキスト断片の段落に基づいて、リクエストの全部のワード・コンビネーションを含んでいて、疑問語ワード・コンビネーションを含んだ、単一の論理的にリンクされたテキスト

$S : \{I_i^{(su)}\}$  を生成させるべきである。

【0155】

このような場合において、ワード・コンビネーション  $S : \{I_i^{(su)}\}$  が、異なるテキスト  $V : \{I_i^{(t)}, I_j^{(a)}\}$  中の異なる段落に含まれている場合に、単一の論理的にリンクされた応答テキストを生成させるための必要条件は、少なくとも1つの段落において、応答ワード・グループのインデックス  $I_0^{(su)}$  と、リクエストの疑問語ワード・コンビネーション  $I_1^{(su)}$  の主要な語と、主語および述語のワード・コンビネーションのインデックスを一般的形態で含んでいる式(15)の叙述ベース  $(I_2^{(su)}, I_3^{(su)})$  とが存在することである。

40

【0156】

該条件が満たされる場合、選択された段落のセットをさらなるプロセッシングに用いる。なぜなら、予備選択された段落に基づいて、単一の論理的にリンクした応答テキストを

50

生成させるように試みることができるからである。そうでない場合には、本方法はその所定テーマに関する新たなテキストのエントリおよび指標付けに進むべきである。

【0157】

該条件が満たされる場合、本方法は論理的にリンクされた該段落の1セットを形成することに進む。そのために、下記条件とのコンプライアンス(順応性)がチェックされる：各ワード・コンビネーションは少なくとも2つの異なる段落中に含まれているべきである：

$$|_{i}^{(su)} \quad (|_{i}^{(t)}, |_{j}^{(a)}), \dots, (|_{k}^{(t)}, |_{l}^{(a)}) \quad (16)$$

該条件が満たされなかった場合、本方法は、1つのリクエストのワード・コンビネーション $|_{i}^{(su)}$ のみを含んでいる段落が、他の予備選択された段落に含まれていて、ベース意味的関係の1つにより $|_{i}^{(su)}$ ワード・コンビネーションにつながっている別のワード・コンビネーション $|_{k}^{(su)}$ を有しているかどうかをチェックする。このチェックのために、自己命令および知識抽出モード制御サブシステム3が、確率的に指標付けされた言語テキストのデータベース8において、下記のトピック コメント型の関係によりつながる該インデックスを含むと思われる文章を探索するためのリクエストを生成させる：

$$|_{k}^{(su)} \quad |_{i}^{(su)} \quad (16a)$$

見つかった文章は、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4に送られ、その関係(16a)が性・相、集合、または原因の関係に対応するか否かをチェックする。

【0158】

条件(16)および(16a)が満たされない場合、このテキスト断片は応答を生成させるのに使用できない、と結論づけられる。

該条件が満たされる場合、本方法は、選択した段落に基づいて単一の意味構造を生成させる可能性についてのチェックに進む。そのために、まず、予備選択された段落を含む各テキストのインデックスの表を用いて、ワード・コンビネーションのインデックスのリストをコンパイルする。これらのワード・コンビネーションのインデックスは適当なインデックスにより指定された段落中に入れられる：

$$(|_{i}^{(t)}, |_{j}^{(a)}) \quad (|_{i}^{(su)}), \dots, (|_{k}^{(su)}) \quad (17)$$

その後、該段落のリスト中の同一のワード・コンビネーションのインデックスを用いて、本方法は、それぞれ所定の段落が関係するのが、段落の何であるかを決定する。該リストに基づいて、各段落インデックスごとに新たなリストがコンパイルされ、その各リストは、同一のワード・コンビネーションのインデックスにより所定の段落につながっている、他方の段落のインデックスを含んでいる。それにより、リストのそれぞれが、他のリストの少なくとも1つが含んでいる少なくとも1つの段落インデックスを含んでいる場合には、リスト間の直接的または推移的な関係を用いて、全部の段落のインデックスを含んでいる単一のリストが生成されよう。この場合、予備選択された段落は、単一テキスト断片の形態の、論理的にリンクされた複数の段落からなる1セットを作ると考えられる。そうでなければ、この段落のセットは、単一テキスト断片を構成するのに要求されるであろう、1つの論理的構造を構成しないと考えられる。このような場合は、該セットはプロセッシングから除外され、本方法は新たなテキスト断片の予備選択に進む。

【0159】

予備選択された複数段落が論理的にリンクされた複数段落の単一構造を構成することが確かめられた後、各テキストのインデックスの対応する表に基づいて単一のテキスト表を生成させる。そのために、該段落が、リクエストの問い合わせ文章内の段落が含んでいるリクエストのワード・コンビネーションの続きの順序により決まる順番で並べられる。得られた1つのテキスト断片を、リクエストの全部のワード・コンビネーション $\{|_{i}^{(u)}\}$ を含んでいる段落の文章間の意味的リンクのタイプを、論理的結論を用いて決定するためのさらなるプロセッシングのために送る。該ファンクションは、上述したアルゴリズムにより得られたテキスト断片に基づいて、リクエストの全部のワード・コンビネーションを含むであろう1つの確率的に指標付けされた意味構造を生成させる目的で行われる。その後、得られた意味構造を適用して、上述したアルゴリズムに従った等価変換および推移的

10

20

30

40

50

関係についての論理的結論を用いて、ユーザー・リクエストに適合した簡単な応答を含んでいる文章の意味構造(15)を生成させる。簡単な応答の正しさは、テキスト系文書の種々の予備選択された確率的に指標付けされた断片に基づいて、上記手順に従って、いくつかの同一の確率的に指標付けされた意味構造(15)が生成することにより保証される。

【0160】

得られた簡単な応答は、疑問語ワード・コンビネーションと一緒に、「リクエスト - リプライ」知識ベース9に格納され、このベースは繰り返される典型的なユーザー・リクエストのプロセッシングと、上述したように指標付けされたテキストの意味解析のために使用される。

【0161】

意味構造をセットアップした後、所定のテキスト断片におけるリクエストのワード・コンビネーション $\{l_{ij}^{(su)}\}$ 間で、要求されたベースの意味的リンクが維持されないとわかった場合には、本方法は、ユーザーへの応答の生成のために新たなテキストの探索に進む。

【0162】

論理的結論が肯定の結果を生ずる場合には、リクエストに適切な簡単な応答を持つ文章を生成させて、これを所定言語のテキストとしてユーザーに出力することになる。この場合に、ユーザーがより完全な応答を要求する場合には、本方法は、後述するアルゴリズムに従って、先に得られたテキスト断片の変換に基づいて、完全な応答を生成させる工程に進む。

【0163】

簡単な回答を生成させるための上述したアルゴリズムの実行は、次のように例示される。受領したユーザー・リクエストの等価変換後に、リクエストが下記のテキスト系形態を得ているとする："What program is used in case of incorrect computer operation termination as a result of voltage failure in the mains?" (電力本線の電圧不全の結果としての間違ったコンピュータ操作の停止の際に何のプログラムが使用されるか?)。これから、全体としては変換されたリクエストの全部のワード・コンビネーションを含んでいる、異なるテキスト系文書からの2つの下記の論理的にリンクされた段落が予備選択される可能性が考えられる。

【0164】

第1の段落は次の通りである："Logical errors may occur on the hard disc. The logical errors are disorders in the file structure. To find out the logical errors, the "Disc check" routine is used. The logical errors occur when computer operation is terminated incorrectly." (論理的エラーがハードディスク上で起こることがある。論理的エラーはファイル構造の不調である。論理的エラーを見つけるには「ディスクチェック」ルーティンが使用される。論理的エラーはコンピュータ操作が間違っして停止される時に起こる。)

第2の段落は次の通りである："In case of voltage failure in the mains, file structure disorders occur on the hard disc. In such case the "Disc check" routine is used." (電力本線の電圧不全の場合、ハードディスク上でファイル構造の不調が起こる。このような場合「ディスクチェック」ルーティンが使用される。)

【0165】

確率的に指標付けされた形態において、リクエストの実際のプロセッシングが起こり、簡単な応答が形成されるが、その時のリクエストは下記の通りであろう：

$$l_{01}^{(p)} : l_{01}^{(su)} \quad l_{02}^{(su)} \quad l_{03}^{(su)} \quad l_{04}^{(su)} \quad l_{05}^{(su)} \quad l_{041}^{(su)} \quad l_{051}^{(su)} \quad (18)$$

下記のワード・コンビネーションが確率的インデックス $l_{0j}^{(su)}$ に対応しよう：

- $l_{01}^{(su)} = (\text{what program})(\text{何のプログラム})$
- $l_{02}^{(su)} = (\text{program})(\text{プログラム})$
- $l_{03}^{(su)} = (\text{is used})(\text{使用される})$

10

20

30

40

50

- |  $_{04}^{(su)}$  = (in case of incorrect termination)(間違った停止の場合)
- |  $_{05}^{(su)}$  = (computer operation)(コンピュータ操作)
- |  $_{041}^{(su)}$  = (as a result of failure)(不全の結果として)
- |  $_{051}^{(su)}$  = (voltage in the mains)(電力本線の電圧)。

## 【0166】

確率的に指標付けされた形態の第1の段落の文章は、次のように提示されよう：

- |  $_{1}^{(p)} : |_{12}^{(su)} |_{13}^{(su)} |_{14}^{(su)}$
- |  $_{2}^{(p)} : |_{22}^{(su)} |_{23}^{(su)} |_{24}^{(su)}$
- |  $_{3}^{(p)} : |_{32}^{(su)} |_{33}^{(su)} |_{34}^{(su)} |_{35}^{(su)}$
- |  $_{4}^{(p)} : |_{42}^{(su)} |_{43}^{(su)} |_{44}^{(su)} |_{45}^{(su)}$  (19)。 10

## 【0167】

下記のワード・コンビネーションが確率的インデックス  $|_{ij}^{(su)}$  に対応しよう：

- |  $_{12}^{(su)}$  = (logical errors)(論理的エラー)
- |  $_{13}^{(su)}$  = (may occur)(起こることがある)
- |  $_{14}^{(su)}$  = (on the hard disc)(ハードディスク上で)
- |  $_{22}^{(su)}$  = (logical errors)(論理的エラー)
- |  $_{23}^{(su)}$  = (is)(である)
- |  $_{24}^{(su)}$  = (disorders in the file structure)(ファイル構造の不調)
- |  $_{32}^{(su)}$  = ("Disc check" routine)(「ディスクチェック」ルーティン)
- |  $_{33}^{(su)}$  = (is used)(使用される) 20
- |  $_{34}^{(su)}$  = (to find out)(見つけるために)
- |  $_{35}^{(su)}$  = (logical errors)(論理的エラー)
- |  $_{42}^{(su)}$  = (logical errors)(論理的エラー)
- |  $_{43}^{(su)}$  = (occur)(起こる)
- |  $_{44}^{(su)}$  = (in case of incorrect termination)(間違った停止の場合)
- |  $_{45}^{(su)}$  = (of computer operation)(コンピュータ操作の)。

## 【0168】

確率的に指標付けされた形態の第2の段落の文章は、次のようであろう：

- |  $_{5}^{(p)} : |_{52}^{(su)} |_{53}^{(su)} |_{54}^{(su)} |_{55}^{(su)} |_{551}^{(su)}$
- |  $_{6}^{(p)} : |_{62}^{(su)} |_{63}^{(su)} |_{64}^{(su)}$  (20)。 30

## 【0169】

下記のワード・コンビネーションが確率的インデックス  $|_{ij}^{(su)}$  に対応しよう：

- |  $_{52}^{(su)}$  = (disorders in the file structure)(ファイル構造の不調)
- |  $_{53}^{(su)}$  = (occur)(起こる)
- |  $_{54}^{(su)}$  = (on the hard disc)(ハードディスク上で)
- |  $_{55}^{(su)}$  = (as a result of failure)(不全の結果として)
- |  $_{551}^{(su)}$  = (voltage in the mains)(電力本線の電圧)
- |  $_{62}^{(su)}$  = ("Disc check" routine)(「ディスクチェック」ルーティン)
- |  $_{63}^{(su)}$  = (is used)(使用される)
- |  $_{64}^{(su)}$  = (to find out)(見つけるために) 40
- |  $_{65}^{(su)}$  = (logical errors)(論理的エラー)
- |  $_{651}^{(su)}$  = (in such case)(そのような場合)。

## 【0170】

該確率的に指標付けされた意味構造に基づき、全部のワード・コンビネーション  $|_{ij}^{(su)}$  を含む確率的に指標付けされた意味構造が、上述した手順に従ってセットアップされよう。基準として、疑問語ワード・コンビネーション  $|_{31}^{(p)}$  に対応する応答ワード・グループ  $|_{32}^{(su)}$  を含んでいる構造  $|_{3}^{(p)}$  が選択される。下記ワード・コンビネーションの同一性(語幹として正確に)を考慮に入れる：

- |  $_{02}^{(su)} = |_{31}^{(su)} = |_{64}^{(su)}$
- |  $_{03}^{(su)} = |_{33}^{(su)}$  50

$$\begin{aligned}
 |_{04}^{(su)} &= |_{44}^{(su)} \\
 |_{05}^{(su)} &= |_{45}^{(su)} \\
 |_{041}^{(su)} &= |_{55}^{(su)} \\
 |_{051}^{(su)} &= |_{551}^{(su)} \\
 |_{12}^{(su)} &= |_{22}^{(su)} = |_{42}^{(su)} \\
 |_{24}^{(su)} &= |_{52}^{(su)}
 \end{aligned}
 \tag{21}$$

その結果、該確率的に指標付けされた構造は下記形態を有するであろう：

$$\begin{aligned}
 |_0^{(p)} : |_{32}^{(su)} & |_{33}^{(su)} & |_{34}^{(su)} & |_{35}^{(su)} & |_{24}^{(su)} & |_{55}^{(su)} & | \\
 |_{551}^{(su)} & |_{35}^{(su)} & |_{44}^{(su)} & & & & \\
 |_{45}^{(su)} & & & & & & 
 \end{aligned}
 \tag{22}$$

10

対応するインデックスの上記同一性の点と、所定意味構造におけるインデックス間の関係が性 - 相および原因の性質を有する点を考慮に入れて、下記の構造が論理的結論を用いて得られよう：

$$\begin{aligned}
 |_0^{(p)} : |_{32}^{(su)} & |_{02}^{(su)} & |_{03}^{(su)} & |_{04}^{(su)} & |_{05}^{(su)} & |_{041}^{(su)} & | \\
 |_{051}^{(su)} & & & & & & 
 \end{aligned}
 \tag{23}$$

こうして、簡単な応答の確率的に指標付けされた意味構造が、下記テキストとして読めるようにセットアップされよう："The "Disk check" routine is used when computer operation is terminated incorrectly as a result of voltage failure in the mains." (電力本線の電圧不全の結果としてコンピュータ操作が間違って停止した時には「ディスクチェック」ルーティンが使用される。)

20

#### 【0171】

得られた簡単な応答は、"Disk check" routine("ディスクチェック"ルーティン)という応答ワード・グループを"What program"(何のプログラム)という疑問語ワード・コンビネーションで置換した後に、"What program is used in case of incorrect computer operation termination as a result of voltage failure in the mains?" (電力本線の電圧不全の結果としての間違ったコンピュータ操作の停止の際に何のプログラムが使用されるか?)なるリクエストと同一になる。これは得られた簡単な応答のリクエストに対する適合性の特徴である。従って、得られた簡単な応答はユーザーに出力されるものである。

#### 【0172】

予備選択された段落または得られたテキスト断片に基づいて完全な回答を生成させるには、簡単な応答の文章の生成における論理的結論に含まれていた文章だけが選択される。該段落またはテキスト断片の文章が論理的リンクにより生じた順番に並べられる。論理的リンクの順序は、リクエストのワード・コンビネーション間の意味的リンクを確認するのに使用されたものと同じである。異なる文章に含まれている、これらのワード・コンビネーションは、応答ワード・グループおよび疑問語ワード・コンビネーションの主要な語を有する文章に含まれているのと同じリクエストのワード・コンビネーションに関連づけられる。文章連結を続けていく順序は、既に生成させてある簡単な応答における該連結に対応するリクエストのワード・コンビネーションの続きの順序により決まる。文章同士の呼応(一致)を与えるために、完全な応答の生成は、一部の文章の、該文章の意味の変化を引き起こさないスピーチ・パートまたは文章メンバーの置換による等価変換を含んでもよい。文章の等価変換が前置詞の置換を必要とする場合には、スピーチ・パートがその前置詞と組み合わせられる時に有していなければならない特徴を考慮に入れてそれを置換する。必要なら、該スピーチ・パートの格を、名詞もしくは形容詞、代名詞もしくは分詞の間でのその新たな前置詞との呼応のために置換することがある。そのために、ある前置詞を格と結びつけるのに適当なルールを適用して、該スピーチ・パートがその所定の前置詞と呼応するようにする。

30

40

#### 【0173】

リクエストの疑問語または疑問語ワード・コンビネーション (how? in what manner?) (いかに? 何のやり方で?)が1文章の簡単な応答を想定しないが、あるプロセスまたは現

50

象の1系列の複数工程または記述を想定するなら、そのような場合には、簡単な応答を、下記の種類の応答のワード・グループを含む開始文章として言い表してもよい："as follows", "thus" (「次の通り」「こうして」)。次に、応答の次の文章が、要求される完全さで応答を含んでいる1系列の複数工程または記述の内容を開示しよう。応答のそのような典型的なワード・グループが存在しない場合、それを付加的に導入して開始文章を生成させることができる。その後、開始文章中の応答ワード・グループが、将来の完全な応答のための出発トピックとして採用されよう。さらに、論理的結論を用いて、ある所定のユーザーの質問に対する完全な応答となる、意味的につながった1セットの複数文章を構成する、1または2以上の段落の1系列の文章が選択される。応答の境界は、論理的にリンクされた複数文章の連続的な連結により決められよう。その連結は、段落の1つの最後の文章のトピックが次の段落の最初の文章のコメントにつながらない場合に、その段落の終結により終わりとなる。開始文章を含んでいる完全な応答を含んだテキスト断片を生成させた後、該断片がユーザーに出力されることになる。

10

## 【0174】

ここに開発された本発明の方法は、ある1つの所定外国言語の検索システムに使用するためにテキスト系文書から知識を抽出するための自己命令システムを合成するのに好適に使用することができる。形態素、統語(構文)および意味の各解析のルールに対するシステムの自動命令が、上述した手順に従って、所定外国言語の確率的に指標付けされた言語テキストを用いて行われる。やはり所定の外国言語で提示される、導出されたルールは、確率的に指標付けされて、形態素的、統語および意味の各解析の対応する知識ベース12~14

20

## 【0175】

該データおよび知識ベースが上述した手順に従って生成した後、ユーザーのリクエストが所定外国言語に変換され、要求されたトピックのテキスト系文書の断片が予備的に選択される。次に、テキスト系文書断片のデータを等価変換に付し; 確率的に指標付けされた意味構造を生成させ、そして該構造を用いて論理的結論を引き出し、所定外国言語のリクエストに適合した簡単な応答を生成させる。

## 【0176】

ここに開発された方法はまた、ある複数の所定外国言語のいずれでもよい検索システムに使用するためにテキスト系文書から知識を抽出するための自己命令システムを合成するのに好適に使用することができる。そのために使用されるのが、確率的指標付けのための確率的インデックスの2進信号の単一(独自の)組み合わせの適用と、ある1つの所定基本言語における、文法および意味解析の記述を含んでいる言語テキスト断片の探索、とに基づき、確率的に指標付けされた人工知能システムの形態の、上述した自己命令メカニズムである。このメカニズムは、所定の複数外国言語のどれかのテキストの確率的に指標付けされた断片の等価変換と、該断片からの論理的結論およびリンクされた意味構造の生成と、プロダクションルールの形態で表示される該構造の確率的指標付けとを行うために、文法および意味の解析のルールに対するシステムの自動的な自己命令を生ずる。

30

40

## 【0177】

まず、上述したメカニズムを用いて、ある基本言語での電子形態の言語テキストの形態素解析および確率的指標付けが、形態素解析ルールに対するシステムの命令と同時にされる。これは、所定複数外国言語のそれぞれについて、確率的に指標付けされた辞書類のデータベース7と、データベース8の言語テキストのインデックスの表、の創出と同時に、ならびに所定の基本言語および所定の複数外国言語のそれぞれについて導出されたプロダクションルールを含んでいる形態素解析の知識ベース12の創出により、行われる。

## 【0178】

次に行われるのは、検索システムからの、所定の複数外国言語のそれぞれにおける電子形態の或る所定トピックに関するテキスト系文書の形態素解析および統語解析、ならびに

50

その確率的指標付けである。その後、ある所定テーマのテキスト系文書のインデックスの表が形成され、それらは確率的に指標付けされたテキスト系文書のデータベース10に格納され、同時に統語解析ルールに対するシステムの自動命令がなされる。該命令は、所定基本言語の確率的に指標付けされた言語テキストを用いて前述した手順に従って実行される。その後、基本言語および所定の複数外国言語のそれぞれについて統語解析の知識ベース13が創出される。

**【0179】**

次に行われるのは、電子形態の、所定基本言語での所定テーマについての確率的に指標付けされたテキスト系文書の意味解析であり、同時に意味解析ルールに対するシステムの自動命令と、基本言語および所定の複数外国言語のそれぞれについて意味解析の知識ベース14の創出とがなされる。

10

**【0180】**

知識ベース11、12の完成後、システムは自動自己命令モードからユーザー・リクエストの処理モードに移る。そうする際に、ユーザーのリクエストがある所定の自然外国言語で生成され、リクエストが、疑問語ワード・コンビネーションとリクエストの意味を規定するワード・コンビネーションとを含んだ1つの問い合わせ文章の形態でその確率的指標付け後に電子形態で提示される。その後、上述した手順を用いて、確率的に指標付けされた形態のユーザーのリクエストが、その所定外国言語での元のリクエストに等価の複数の新たなリクエストに変換される。その後、ユーザーのリクエストに従って、変換されたリクエストの全部のワード・コンビネーションを含む、電子形態の所定外国言語でのテキスト系文書の確率的に指標付けされた断片類を予備選択する。該テキスト系文書断片を用いて、確率的に指標付けされた意味構造を生成させる。生成した確率的に指標付けされた意味構造に基づいて、各種テキストの確率的に指標付けされた要素の間のリンクを与える論理的結論を用いて、かつテキストの等価変換を用いて、リクエストの意味を規定する確率的に指標付けされたワード・コンビネーションと、リクエストの疑問語ワード・コンビネーションに対応する応答ワード・グループ、とを含む簡単な応答を生成させる。この簡単な応答の正しさは、異なる予備選択した確率的に指標付けされたテキスト系文書断片に基づいて、いくつかの同一の確率的に指標付けされた意味構造が生成することにより保証される。

20

**【0181】**

得られた簡単な応答のリクエストに対する適合性は、応答ワード・グループを、対応する確率的に指標付けされた疑問語ワード・コンビネーションで置換して、確率的に指標付けされた1つの問い合わせ文章を生成させ、得られた問い合わせ文章をリクエストと比較することによりチェックされる。これらの文章の比較に基づいて、得られた問い合わせ文章とリクエストとが同一であるという結果になった場合には、その簡単な応答はリクエストに適合しているという決定がなされ、この応答は所定の外国言語でユーザーに提示される。

30

**【0182】**

所定の複数外国言語のいずれかでテキスト系文書からの知識の同時抽出を生ずる自己命令システムの合成方法を適用するという別のバージョンを検討する。この場合、まず、上述した手順に従って、ある所定の基本言語での確率的に指標付けされた言語テキストを使用して、形態素、統語および意味の各解析のルールについて、システムを自動的に命令する。確率的に指標付けされた言語テキストのデータベース8は、選択された基本言語に基づいて所定の複数外国言語のそれぞれを学習するために教育・命令的なハンドブックを備えている。外国語の確率的に指標付けされた辞書のデータベース11に、個々の語の基本言語から所定のすべての外国言語への順方向および逆方向の翻訳を与える辞書が書き込まれる。次に、確率的に指標付けされた辞書類のデータベース7ならびに所定基本言語での形態素、統語および意味の各解析の知識ベース12~14が創出される。その後、自動的自己命令モード制御サブシステム3が、基本言語での言語テキスト断片の予備選択のために該データベースおよび知識ベースへのリクエストを自動的に生成させる。その断片は、所定の

40

50

複数外国言語のそれぞれを学習するのに必要な知識を含んでいよう。次に、そのテキストに等価変換を受けさせ、確率的に指標付けされた意味構造を生成させ、そして自動的リクエストに適合した応答を生成させるように予め定めた構造に従って論理的結論を引き出す。これらの応答を用いて、各外国言語についてテキスト系文書の形態素、統語および意味の各解析のプロダクションルールを導出する。例えば、基本言語がロシア語である場合、自動的に導出されたルールのうち、英語を学習するための統語解析の知識ベースは下記のものを含んでいるかもしれない：

1. 文章の初めに前置詞のない名詞が位置し、  
そして、この名詞が前置詞 of (in, from) を持つ名詞の前に位置し、  
そして、該名詞の後に動詞がくるなら、  
最初の名詞は主語である。

10

## 【0183】

例えば：The work of the engineer is on the table. (その技術者の仕事は机上である)。

2. ワード・コンビネーションが連結動詞(verb-copula)(ある人称形態の to be 動詞)と、形容詞により表現された名詞的部分(nominal part)とからなる場合、

このワード・コンビネーションは複合名詞的述語(composite nominal predicate)である。

## 【0184】

例：The tree is big. (その木は大きい)。

20

導出されたルールは、確率的指標付け後に、形態素、統語および意味の各解析の知識ベース12~14に書き込まれて、ユーザーがリクエストした通りのある外国言語でのテキスト系文書からの知識の抽出を与える。所定のトピックに関して、確率的指標付け辞書類のデータベースと指標付けされたテキスト系文書の表との創出が、対応する外国言語を用いて行われる。要求された外国言語での予め定めたトピックに関するテキスト系文書の意味解析において、意味的リンクのタイプを決定するために、確率的に指標付けされた外国語辞書のデータベース11を用いて、いくつかのワード・コンビネーションが基本言語に翻訳されることに留意されたい。このようなワード・コンビネーションは、論理的結論を用いて、基本言語参照辞書のインデックスの表に従って、そのインデックスが、確率的に指標付けされたテキストとプロダクションルールの通訳プログラム4に書かれている、意味的関係のタイプの1つと相関づけられる。これにより、意味解析を用いて、上述した手順に従って、スピーチ・パートとの語の対応性を特定し、リクエストへの応答の確率的に指標付けされた意味構造を生成させる時にワード・コンビネーション間の関係のタイプを決めることが可能になる。

30

## 【0185】

該データベースと知識ベースとを使用して、ユーザーのリクエストは、自己命令および知識抽出モード制御サブシステム3からの指令により、所定複数外国言語での等価変換を受ける。その後、テキスト系文書の断片が所定トピックに対して予備選択され、それらの等価変換が行われ、確率的に指標付けされた意味構造を生成させ、そして該構造に基づいて論理的結論が導出される。これは、所定複数外国言語のそれぞれにおいてユーザーのリクエストに適合した応答が調製されるのを保証する。

40

## 【0186】

リクエストのプロセッシング中に、あるトピックに対するある外国言語の新たなテキスト系文書をエントリするために検索システムを参照しなければならないことが判明した場合には、自己命令および知識抽出モード制御サブシステム3が多言語の言語プロセッサ1を始動させる。このプロセッサは、基本言語に新たな文章をエントリするための指令を受け、その指令はトピックとその外国言語の名前を特定する。多言語の言語プロセッサ1は、確率的に指標付けされた外国言語辞書のデータベース11を用いて、必要な辞書を選択し、トピック名称を示す語を適当な外国言語に翻訳する。受け取った情報に基づいて、多言語の言語プロセッサ1は、関係するトピックに関する新たな外国言語文書をエントリする

50

ように、検索システムへの所定言語での公式化リクエストを準備する。該文書は、上記プロセッシングとそれらを確認的に指標付けされたテキスト系文書のデータベース10に格納するために、テキスト系文書の確率的指標付けおよびテキスト断片の切り離しのサブシステム2に送られる。

【産業上の利用可能性】

【0187】

検索システムに使用するためテキスト系文書からの知識抽出用の自己命令システムを合成する本発明の方法は、テキストから知識を抽出するための多言語システムを用いて、インターネットに基づく世界的な知識産業の創出に使用することができる。この教示内容の実現は、文明社会の発展の現代の要求を考慮して、多様な分野：産業、科学、教育、文化、の人的活動における質的に新しい情報サービスを提供しよう。本方法の産業上の利用に対するもう1つの有望な方向性は、モバイルシステム(モバイルインターネット)である。この方向性は、インターネットに格納された多量のテキスト系文書から、ユーザーのリクエストによって、ユーザーが必要とする情報の伝達と受領に要する時間を最小限にして、特定の知識およびデータを抽出するのを可能にする、インテリジェント情報検索システムを創出する可能性により支えられよう。ユーザーは、自然言語かつ会話形態でシステムにリクエストを入力することができる。特許請求された方法の産業上の利用の重要な方向性は、各種の主題および問題の領域におけるインテリジェント命令システムの新たな生成を創出することにある。

10

【0188】

20

【表 1】

## 文章フレーム

単文への質問	単文への質問は統語解析知識ベースに基づいて生成させる	
複文または十分が含んでいる単文の種類	単文の特徴	
文章メンバーのグループへの質問	文章メンバーのグループへの質問はあるグループのベースである文章メンバーへの質問に基づいて生成させる	10
文章メンバーのグループの名称	下記のグループ： 述語 目的語 副詞 独立した文章メンバー 挿入語、ワード・コンビネーションおよび挿入構造	20
文章のメンバーへの質問	辞書フォーマット (前置詞類を含む) およびスピーチパートへの質問から文章メンバーへの質問の翻訳の表に従う	
文章メンバーの名称	主語 述語 (単動詞、複合動詞 <composite verbal>、複合名詞 <composite nominal>) 限定語句 (attribute) (呼応、非呼応) 目的語 (直接、間接) 副詞的修飾語句 (方法、場所、時間、尺度もしくはは程度、原因、目的、条件、譲歩)	30
スピーチ・パートへの質問	辞書フォーマットに従う	
スピーチ・パートとそれらの特徴	辞書フォーマットに従う	
語 (ワード)	文章の前後関係における	
語幹の確率的インデックス	特殊なアルゴリズムに従って算出されるか、または辞書フォーマットから切り離される	

10

20

30

40

【表 2】

## テキスト・インデックス

語幹の インデックス	段落のインデックス			
	$I_{\zeta 1}^{(a)}$	$I_{\zeta 2}^{(a)}$	...	$I_{\zeta n}^{(a)}$
$I_{\zeta 1}^{(u)}$	$I_{\zeta 11}^{(s)}$	$I_{\zeta 12}^{(s)}$	...	$I_{\zeta 1n}^{(s)}$
$I_{\zeta 2}^{(u)}$	$I_{\zeta 21}^{(s)}$	$I_{\zeta 22}^{(s)}$	...	$I_{\zeta 2n}^{(s)}$
...	...	...	...	...
$I_{\zeta m}^{(u)}$	$I_{\zeta m1}^{(s)}$	$I_{\zeta m2}^{(s)}$	...	$I_{\zeta mn}^{(s)}$

10

【0190】

【表 3】

## あるトピックに関するテキストのインデックス

語幹の インデックス	段落のインデックス			
	$I_{\zeta 1}^{(t)}$	$I_{\zeta 2}^{(t)}$	...	$I_{\zeta n}^{(a)}$
$I_{\zeta 1}^{(u)}$	$I_{\zeta 11}^{(s)}$	$I_{\zeta 12}^{(s)}$	...	$I_{\zeta 1n}^{(s)}$
$I_{\zeta 2}^{(u)}$	$I_{\zeta 21}^{(s)}$	$I_{\zeta 22}^{(s)}$	...	$I_{\zeta 2n}^{(s)}$
...	...	...	...	...
$I_{\zeta m}^{(u)}$	$I_{\zeta m1}^{(s)}$	$I_{\zeta m2}^{(s)}$	...	$I_{\zeta mn}^{(s)}$

20

【図面の簡単な説明】

【0191】

【図 1】 検索システム用のテキスト系文書からの知識の抽出のためのインテリジェント自己学習システムのブロック図を示す。

30

【 図 1 】

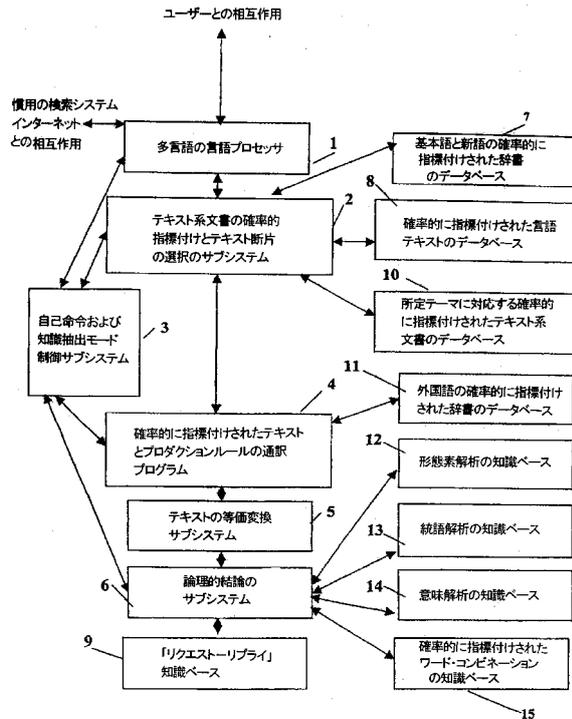


Fig. 1

【 手続 補正書 】

【 提出日 】 平成17年5月23日 (2005.5.23)

【 手続 補正 1 】

【 補正対象書類名 】 明細書

【 補正対象項目名 】 0 0 2 9

【 補正方法 】 変更

【 補正の内容 】

【 0 0 2 9 】

文法解析 - 形態素的かつ統語的な解析。

知識とは、テキスト系文書に明確には含まれていない新たなテキスト系情報であり、その情報は、等価変換および論理的結論を用いて、応答の形態でシステムにより自動的に生成し、かつその情報はユーザーのリクエストに適合性がある、そのリクエストに従って対応する問題を解決するためのものである。

【 手続 補正 2 】

【 補正対象書類名 】 明細書

【 補正対象項目名 】 0 0 6 3

【 補正方法 】 変更

【 補正の内容 】

【 0 0 6 3 】

等価変換のサブシステム5内で導出された、式(2)の形態の各プロダクションルールは、確率的に指標付けされたテキストおよびプロダクションルールの通訳プログラム4に送られ、そこで式(2)はプロダクションルールのテキスト系フォーマットに変換される。このフォーマットは、「...なら (if, 条件)、...である (then, 結論)」と表現される。指標付けされた形態の導出されたルールが形態素解析の知識ベース12に送られる。確率的に指標付けされたルールを含む知識ベースを合成するための手順は後で詳細に述べる。

## 【手続補正 3】

【補正対象書類名】明細書

【補正対象項目名】0065

【補正方法】変更

【補正の内容】

【0065】

上述した手順に従って、確率的に指標付けされた言語テキストのデータベース 8 の言語テキストに対応する類義語辞書の指標付けの表を参照することにより、第 1 および第 2 の意味構造の動詞の意味の、リクエストおよび想定される応答とのコンプライアンス(順応性)がチェックされる。その後、第 1 および第 2 の意味構造をリンクするワード・コンビネーションが決定される。結果が肯定的である場合は、該テキスト断片の 2 つの部分のテキストの等価変換のサブシステム 5 に送られ、その後、確率的に指標付けされたテキストとプロダクションルールの翻訳プログラム 4 に送られる。その結果、該テキスト断片は、「...なら (if, 条件)、...である (then, 結論)」と表されるプロダクションルールのフォーマットに変換される。このルールの条件部は、語のスピーチ・パートと、辞書フォーマットに提供されて、その語が語形変化または活用変化する時に語形の変化を決める、語尾および前置詞の予め定められたセットとを関連づける、ワード・コンビネーションのインデックスを包含する。結論部は、対応するスピーチ・パートとして、その語の語形変化または活用変化により生じた語尾、前置詞および質問の完全セットを包含する。公式化されたプロダクションルールは、形態素解析の知識ベース 12 に書き込まれる。スピーチ・パートを決めるルール導出された後、プロセスは、自己命令および知識抽出モードを制御するためのサブシステム 3 からの指令により、共通語根の語の等価変換のためのルールを合成する段階に進む。その目的には、形態素解析の知識ベース 12 に予め格納され、そして確率的に指標付けされた基本辞書の表と、1 つのスピーチ・パートを別の共通語根スピーチ・パートに基づいて生成させる手順を明示する適当な言語テキスト断片の選択とに基づく、スピーチ・パート変換のための下記的一般ルールが使用される：

「あるスピーチ・パートを別のものに変換すべきであるなら、

まず、第 1 のスピーチ・パートの語幹を切り離し、

次に、確率的に指標付けされた基本辞書フォーマットを参照して、その語幹が、第 1 スピーチ・パートの語幹との、語根(ひょっとすると 2 つの語根、多分接頭辞を伴い、もしかするといくつかの有声字および子音字の変化、付加、削除を伴う)を含んだ共通部分を有している第 2 のスピーチ・パートを探索し；その語根を切り離した後、これらのスピーチ・パートの語幹を用いて、それらの接尾辞を切り離し；次に、スピーチ・パートの語幹の確率的インデックスに関して言語テキスト・インデックスの表を参照することにより、あるスピーチ・パートを別のスピーチ・パートに変換するための適当な方法を記述している断片を選択し；そして、辞書フォーマットを参照して、第 1 スピーチ・パートの語幹に対する第 2 スピーチ・パートの語幹の形成のやり方を決定し(接尾辞の置換、除去、付加)；その後、このあるスピーチ・パートを置換する方法が第 1 スピーチ・パートから第 2 スピーチ・パートを形成するための要求された方法に対応するか否かを決定し；そして、肯定の場合には、第 2 スピーチ・パートを、新たに形成されたスピーチ・パートとして採用する」。

## 【手続補正 4】

【補正対象書類名】明細書

【補正対象項目名】0066

【補正方法】変更

【補正の内容】

【0066】

一般ルールを使用した特定の語の変換の一部として、対応する特定のルールがその一般ルールに基づいて導出され、その特定のルールは変換されたスピーチ・パート、接尾辞および 1 つのスピーチ・パートを別のものから形成するための方法を明示する。これは、確

率的に指標付けされたテキストおよびプロダクションルールの通訳プログラム4およびテキストの等価変換のサブシステム5において起こる。上述した手順は、まず所定の断片を、プロダクションルールの単一の確率的に指標付けされた意味構造(2)に変換し、次に、「...なら (if, 条件)、...である (then, 結論)」と表されるプロダクションルールのフォーマットに変換する。これらのルールは、確率的指標付けがなされた後に形態素解析の知識ベース12にエントリされる。

【手続補正5】

【補正対象書類名】明細書

【補正対象項目名】0130

【補正方法】変更

【補正の内容】

【0130】

式(6, 7)のインデックスに従って、予め決められたトピックの確率的に指標付けされたテキストのデータベース10を参照し、予め決められたテーマのテキストのインデックスの表(図1)を使用することにより、疑問語ワード・コンビネーションを含むリクエストの全部のワード・コンビネーションを含んだ1セットの断片が見つけれられる。各テキスト断片は、1または2以上の段落からなるものでよい。

【手続補正6】

【補正対象書類名】明細書

【補正対象項目名】0190

【補正方法】変更

【補正の内容】

【0190】

【表3】

あるトピックに関するテキストのインデックス

語幹の インデックス	段落のインデックス			
	$I_{\zeta 1}^{(t)}$	$I_{\zeta 2}^{(t)}$	...	$I_{\zeta n}^{(t)}$
$I_{\zeta 1}^{(u)}$	$I_{\zeta 11}^{(s)}$	$I_{\zeta 12}^{(s)}$	...	$I_{\zeta 1n}^{(s)}$
$I_{\zeta 2}^{(u)}$	$I_{\zeta 21}^{(s)}$	$I_{\zeta 22}^{(s)}$	...	$I_{\zeta 2n}^{(s)}$
...	...	...	...	...
$I_{\zeta m}^{(u)}$	$I_{\zeta m1}^{(s)}$	$I_{\zeta m2}^{(s)}$	...	$I_{\zeta mn}^{(s)}$

【手続補正書】

【提出日】平成17年5月24日(2005.5.24)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】0169

【補正方法】変更

【補正の内容】

【0169】

下記のワード・コンビネーションが確率的インデックス  $I_{ij}^{(su)}$  に対応しよう：

$I_{52}^{(su)} = (\text{disorders in the file structure})$  (ファイル構造の不調)

$I_{53}^{(su)} = (\text{occur})$  (起こる)

$I_{54}^{(su)} = (\text{on the hard disc})$  (ハードディスク上で)

$I_{55}^{(su)} = (\text{as a result of failure})$  (不全の結果として)

- $I_{551}^{(su)} = (\text{voltage in the mains})(\text{電力本線の電圧})$   
 $I_{62}^{(su)} = (\text{"Disc check" routine})(\text{「ディスクチェック」ルーティン})$   
 $I_{63}^{(su)} = (\text{is used})(\text{使用される})$   
 $I_{64}^{(su)} = (\text{in such case})(\text{そのような場合})。$

【手続補正2】

【補正対象書類名】明細書

【補正対象項目名】0170

【補正方法】変更

【補正の内容】

【0170】

該確率的に指標付けされた意味構造に基づき、全部のワード・コンビネーション $I_{ij}^{(su)}$ を含む確率的に指標付けされた意味構造が、上述した手順に従ってセットアップされよう。基準として、疑問語ワード・コンビネーション $I_{01}^{(su)}$ に対応する応答ワード・グループ $I_{32}^{(su)}$ を含んでいる構造 $I_3^{(p)}$ が選択される。下記ワード・コンビネーションの同一性(語幹として正確に)を考慮に入れる：

$$\begin{aligned}
 I_{02}^{(su)} &= I_{01}^{(su)} = I_{32}^{(su)} = I_{62}^{(su)} \\
 I_{03}^{(su)} &= I_{33}^{(su)} = I_{63}^{(su)} \\
 I_{04}^{(su)} &= I_{44}^{(su)} \\
 I_{05}^{(su)} &= I_{45}^{(su)} \\
 I_{041}^{(su)} &= I_{55}^{(su)} \\
 I_{051}^{(su)} &= I_{551}^{(su)} \\
 I_{12}^{(su)} &= I_{22}^{(su)} = I_{35}^{(su)} = I_{42}^{(su)} \\
 I_{24}^{(su)} &= I_{52}^{(su)} \quad (21)
 \end{aligned}$$

その結果、該確率的に指標付けされた構造は下記形態を有するであろう：

$$\begin{aligned}
 I_{\zeta 0}^{(p)} : I_{\zeta 32}^{(su)} \rightarrow I_{\zeta 33}^{(su)} \rightarrow I_{\zeta 34}^{(su)} \wedge I_{\zeta 35}^{(su)} \rightarrow I_{\zeta 24}^{(su)} \rightarrow I_{\zeta 55}^{(su)} \wedge I_{\zeta 551}^{(su)} \\
 \downarrow \\
 I_{\zeta 35}^{(su)} \rightarrow I_{\zeta 44}^{(su)} \wedge I_{\zeta 45}^{(su)} \quad (22)
 \end{aligned}$$

対応するインデックスの上記同一性の点と、所定意味構造におけるインデックス間の関係が性 - 相および原因の性質を有する点を考慮に入れて、下記の構造が論理的結論を用いて得られよう：

$$I_0^{(p)} : I_{32}^{(su)} \quad I_{02}^{(su)} \quad I_{03}^{(su)} \quad I_{04}^{(su)} \quad I_{05}^{(su)} \quad I_{041}^{(su)} \quad I_{051}^{(su)} \quad (23)$$

こうして、簡単な応答の確率的に指標付けされた意味構造が、下記テキストとして読めるようにセットアップされよう："The "Disc check" routine is used when computer operation is terminated incorrectly as a result of voltage failure in the mains." (電力本線の電圧不全の結果としてコンピュータ操作が間違って停止した時には「ディスクチェック」ルーティンが使用される。)

## 【 国際調査報告 】

## INTERNATIONAL SEARCH REPORT

International application No. PCT/RU 02/00258
--

A. CLASSIFICATION OF SUBJECT MATTER <b>7</b>		G06F 17/30, G09B 19/00		
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) <b>7</b>				
G06F 17/00, 17/20, 17/21, 17/27, 17/28, 17/30, 17/40, G09B 19/00				
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)				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
A	NASIPNIY V.V. et al. Postroenie intellektualnoi informatsionnoi Poiskovoi systemi. Moskva, Prometei, 2001, p. 3-24	1-20		
A	NASIPNIY V.V. et al. Rasvitie teorii postroenia otkrytikh system na osnove informatsionnoi tekhnologii iskusstvennovo intellekta. Moskva, voennoe isdatelstvo, 1994, p. 36-112	1-20		
A	(NAUCHNO-ISSLEDOVATELSKAYA I PROIZVODSTVENNAYA FIRMA "TEKHINTELL") 27.04.2001	1-20		
A	US 5787234 A (BRUCE G. MOLLOY) Jul. 28, 1998	1-20		
A	US 5454106 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) Sep. 26, 1995	1-20		
<input type="checkbox"/> Further documents are listed in the continuation of Box C. <input type="checkbox"/> See patent family annex.				
* Special categories of cited documents: <table border="0"> <tr> <td style="vertical-align: top;"> <p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"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)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p> </td> <td style="vertical-align: top;"> <p>"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</p> <p>"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</p> <p>"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</p> <p>"&amp;" document member of the same patent family</p> </td> </tr> </table>			<p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"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)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"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</p> <p>"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</p> <p>"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</p> <p>"&amp;" document member of the same patent family</p>
<p>"A" document defining the general state of the art which is not considered to be of particular relevance</p> <p>"E" earlier document but published on or after the international filing date</p> <p>"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)</p> <p>"O" document referring to an oral disclosure, use, exhibition or other means</p> <p>"P" document published prior to the international filing date but later than the priority date claimed</p>	<p>"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</p> <p>"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</p> <p>"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</p> <p>"&amp;" document member of the same patent family</p>			
Date of the actual completion of the international search		Date of mailing of the international search report		
03 February 2003 (03.02.03)		20 February 2003 (20.02.03)		
Name and mailing address of the ISA/RU		Authorized officer		
Facsimile No.		Telephone No.		

## ОТЧЕТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Международная заявка №  
PCT/RU 02/00258

A. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ: G06F 17/30, G09B 19/00		
Согласно международной патентной классификации (МПК-7)		
B. ОБЛАСТИ ПОИСКА:		
Проверенный минимум документации (система классификации и индексы) МПК-7: G06F 17/00, 17/20, 17/21, 17/27, 17/28, 17/30, 17/40, G09B 19/00		
Другая проверенная документация в той мере, в какой она включена в поисковые подборки:		
Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, поисковые термины):		
C. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:		
Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	НАСЫПНЫЙ В.В. и др. Построение интеллектуальной информационной поисковой системы. Москва, Прометей, 2001, стр. 3-24	1-20
A	НАСЫПНЫЙ В.В. Развитие теории построения открытых систем на основе информационной технологии искусственного интеллекта. Москва, Военное издательство, 1994, стр. 36-112	1-20
A	RU 2166208 C2 (НАУЧНО-ИССЛЕДОВАТЕЛЬСКАЯ И ПРОИЗВОДСТВЕННАЯ ФИРМА "ТЕХИНТЕЛЛ") 27.04.2001	1-20
A	US 5787234 A (BRUCE G. MOLLOY) Jul. 28, 1998	1-20
A	US 5454106 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) Sep. 26, 1995	1-20
<input type="checkbox"/> последующие документы указаны в продолжении графы C.		<input type="checkbox"/> данные о патентах-аналогах указаны в приложении
* Особые категории ссылочных документов:		T более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
A документ, определяющий общий уровень техники		X документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну и изобретательский уровень
E более ранний документ, но опубликованный на дату международной подачи или после нее		Y документ, порочащий изобретательский уровень в сочетании с одним или несколькими документами той же категории
O документ, относящийся к устному раскрытию, экспонированию и т.д.		& документ, являющийся патентом-аналогом
P документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета и т.д.		
Дата действительного завершения международного поиска: 03 февраля 2003 (03.02.2003)	Дата отправки настоящего отчета о международном поиске: 20 февраля 2003 (20.02.2003)	
Наименование и адрес Международного поискового органа Федеральный институт промышленной собственности РФ, 123995, Москва, Г-59, ГСП-5, Бережковская наб., 30,1 Факс: 243-3337, телстайп: 114818 ПОДАЧА	Уполномоченное лицо:  О. Ревинский  Телефон № 240-25-91	

Форма PCT/ISA/210 (второй лист)(июль 1998)

---

フロントページの続き

(72)発明者 ナシブニイ、ウラジミル・ウラジミロビッチ

ロシア国、115573モスクワ、ウーエル・シピロフスカヤ、デー44/27、カーベ-75

(72)発明者 ナシブナヤ、ガリナ・アナトリエバ

ロシア国、115573モスクワ、ウーエル・シピロフスカヤ、デー44/27、カーベ-75

Fターム(参考) 5B075 ND03 NK02 NK32 QP01

5B091 AA15 CA02 CA12 EA01