自然言語処理　134－4
（２００１．５．３１）

知識辞書構築支援ツールの開発

市村 由美,　酢山 明弘,　櫻井 茂明,　折原 良平

(株) 東芝　研究開発センター　知識メディアラボラトリー
〒 212-8582 神奈川県川崎市幸区小向東芝町 1
Tel.(044)549-2240,　Fax.(044)520-1308
yumi.ichimura@toshiba.co.jp

テキストマイニング用知識辞書の構築支援を目的として，知識辞書構築支援ツール CADDIE (Computer-Aided Dictionary Design Intensive Environment) を開発した．辞書構築プロセスは，分析要件の定義を行う上流工程と，定義した分析要件に基づき辞書を作成する下流工程とからなる．本ツールは，言語的専門知識を有しない知識エンジニアの下流工程における作業を支援するもので，表現リストからの辞書作成とタグ付けによる辞書検証の機能を備えており，Web ブラウザから操作することを特徴としている．実データを用いた実験では，ツールを利用することで，244.51 時間を要する工程を 155.31 時間に短縮でき，36.5%の効率改善が見られた．

キーワード：知識辞書，構築支援ツール，テキストマイニング，情報抽出

Knowledge Dictionary Development Tool

Yumi Ichimura,　Akihiro Suyama,　Shigeaki Sakurai,　Ryohei Orihara

Knowledge Media Laboratory, Corporate R&D Center, TOSHIBA Corp.,
1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki 212-8582, Japan
Tel.+81-44-549-2240,　Fax.+81-44-520-1308
yumi.ichimura@toshiba.co.jp

We propose a knowledge dictionary development tool CADDIE (Computer-Aided Dictionary Design Intensive Environment). A knowledge dictionary is created through two processes: An upper process where elements for analysis are defined, and a lower process where expressions are registered into the dictionary for each element. This tool helps knowledge engineers who are not specialists in linguistics with the work in the lower process. It is a Web application and has two main functions: to register entries into a dictionary using expression lists generated from a text, and to verify the dictionary showing a result of annotation on a text with the dictionary. This tool reduced the amount of time required by the main part in the lower process from 244.51 hours to 155.31 hours in our experiments with real data, which was a 36.5% of increase in efficiency of the work.

keywords: knowledge dictionary, development tool, text mining, information extraction

1 はじめに

電子化された文書はますます増加しているが，膨大な文書の中から欲しい情報を探したり，文書の集合を分析して傾向を掴んだりするための情報アクセス手段はまだ確立されていない．しかし，ナレッジマネジメントに対する世の中の関心の高さを反映して，営業日報やコールセンターへの問い合わせなど，大量の文書データを分析して内容を瞬時に把握したいというニーズが高まりつつある．そのための自然言語処理技術としてテキストマイニングと呼ばれる分野が注目されている [4].

那須川の研究 [5] は，コールセンターへの問い合わせ事例を分析対象としている．従来のクラスタリング手法では，文書内の名詞句を中心とするキーワードを文書の特徴として扱うものが多いのに対し，述語概念に対して，問題，要望，質問などのカテゴリを付与して，意図を含んだ情報を抽出している．渡部の研究 [6] は，連想検索によってアイデアを広げていく発散的思考の支援を目指したもので，単語間の距離に基づき，単語間の連想関係を 2 次元マップ (ネットワーク図) として可視化することにより，文書群全体が持つ特徴や傾向を分析する手法を提案している．

これらの研究に対して，我々は営業日報の分析に取り組んでいる．営業日報は，新聞や論文などと異なり，形態的には未知語や字句の書き誤りが多く，構文的には箇条書きや体言止めを多く含んでいる点が特徴である．このようなクリーンでないテキストから，販促活動と売上の因果関係といった営業活動上有効な情報を分析する手法として，情報抽出に基づくテキストマイニング手法を開発した．また，この手法に基づき，日報分析システムを開発した [1]-[3]．本手法の特徴は，知識辞書[1]を用いて分析対象のテキストから重要な概念を抽出し，その概念に基づいて分析を行う点にある．分析性能は知識辞書に依存するが，辞書構築には言語的専門知識を必要とし，大きなコストがかかる．そこで，知識エンジニアの辞書構築作業を支援するため，知識辞書構築支援ツール CADDIE (Computer-Aided Dictionary Design Intensive Environment) を開発した．

本報告では，辞書構築プロセスにおける本ツールの位置付け，ツールの機能概要，ツールの利用効果について述べる．

[1] 知識辞書のことを「情報抽出ルール」とも呼んでいるが，ここでは「知識辞書」という用語を用いる．

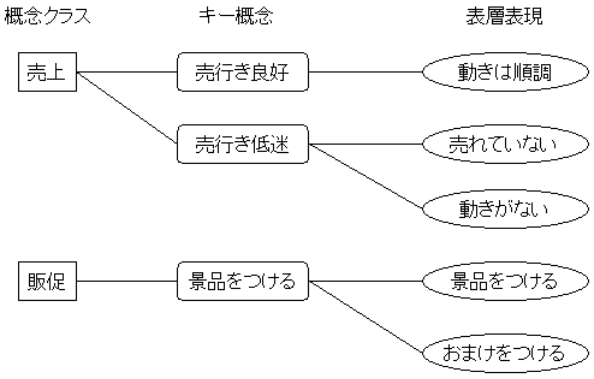

図 1: 知識辞書の記述例

2 ツールの位置付け

2.1 テキストマイニングのための知識辞書とは

知識辞書はテキストから重要な情報を抽出するために用いるものである．辞書には，抽出したい概念と，その概念を表す表現のペアを登録する．抽出したい概念を「キー概念」，概念を表す表現を「表層表現」と呼ぶ．また，多数のキー概念を効率よく扱うため，内容的にまとまりのあるキー概念をグルーピングしておく．このグループを「概念クラス」と呼ぶ．

図 1 に辞書の記述例を示す．概念クラス，キー概念，表層表現の 3 階層で構成されており，表層表現には形態素解析結果に対応する正規表現を記述する．ただし，ここでは説明を簡単にするために，品詞情報や正規表現のメタ文字は省略している．

たとえば，テキスト中に「動きは順調」という表現が出現すれば，該当箇所に「売行き良好」というキー概念タグを付与する．また,「売れていない」や「動きがない」という表現が出現すれば，該当箇所に「売行き低迷」というキー概念タグを付与する．

知識辞書の利用により，表記の多様性が吸収され,「売れていない」や「動きがない」のように表層の表現は異なるが同じ意味をもつものは，ひとつの概念として抽象化されるので，精度の高い分析が可能になる．

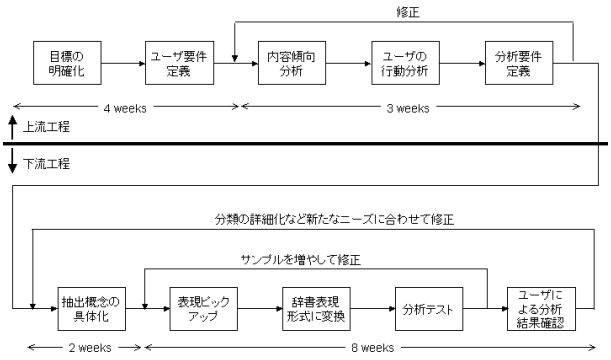

図 2: 辞書構築のプロセス

2.2 辞書構築のプロセス

知識辞書は，分析の目的やテキストの内容に応じて用意する．図 2 に，辞書構築プロセスのフローを示す．辞書構築は，(1) ユーザへのヒアリングを通じて分析要件の定義を行う上流工程と，(2) その分析要件に基づいてテキストから必要な表現を辞書に登録する下流工程，の 2 つのプロセスからなる．

上流工程では，目標の明確化，ユーザ要件の定義，内容の傾向分析，ユーザの行動分析，分析要件の定義を行う．この工程には 7 週間を要し，分析対象となる業務に関する知識や，コンサルテーションに関するスキルが求められる．

一方，下流工程では，抽出概念の具体化，表現のピックアップ，辞書表現形式への変換，分析テスト，ユーザによる分析結果の確認を行う．この工程には 10 週間を要し，言語的専門知識が求められる．

2.3 辞書構築の支援

言語処理の専門家ではない知識エンジニアにとって，2.2 節で述べたプロセスの下流工程では，つぎのような問題点がある．

- 抽出すべき表現を選出するのが難しい．
- 正規表現を正確に記述するのが難しい．
- 抽出結果を確認する有効な手段がない．
- 以前の抽出結果との差分を確認する有効な手段がない．

辞書構築支援ツール CADDIE は，これらの工程の支援を目的に開発されたものである．

3 ツールの機能

3.1 機能概要

CADDIE は，Windows NT 上で動作する Web アプリケーションであり，以下の 5 つの機能を有している．

作成 単語，共起，句のリストから辞書に入れたい表現を探して，キー概念と対応づけて辞書登録する．

検証 作成された辞書を用いて，実際のテキストにキー概念タグを付与しながら，登録洩れや誤りがないかチェックする．

編集メイン 作成された辞書を編集する．上記の「作成」「検証」における辞書編集ではサポートされていない，拡張編集機能がある．

形態素解析 指定されたテキストの形態素解析結果を確認する．

差分表示 2 つの知識辞書同士の差分を作成する．

以下，メインとなる作成機能と検証機能について，詳しく述べる．

3.2 作成機能

図 6 に作成画面を示す．左側に作成中の知識辞書が表示され，右側に指定されたテキストから自動作成された表現リストが表示される．このリストは,「単語」,「共起」,「句」ボタンをクリックすることで切り替えられる．

表現リスト上で登録したい表現を選択し，知識辞書上で登録先のキー概念を選択し,「登録」ボタンをクリックすると，選択された表現が選択されたキー概念の正規表現として登録される．同時に，表現リスト上のステイタス欄には「Registered」と表示され，その表現が登録済みであることがわかるようになっている．また，表現リスト上で不要な表現を選択し,「不要」ボタンをクリックすると，選択された表現のステイタス欄に「Garbage」と表示され，その表現が削除済みであることがわかるようになっている．

この機能により，表現リストからの選択操作だけで辞書作成が行えるので，抽出すべき表現の選出や正規表現の記述に関する困難を解決できる．

一方，知識辞書上では，ノードの追加，削除，編集が行える．なお，編集メイン機能では，上記

の 3 つに加えて，ノードの移動，複写，削除したノードの閲覧，ノードのプロパティ編集が行えるようになっている．

3.3 検証機能

図 7 に検証画面を示す．左側に作成中の知識辞書が表示され，右側に指定されたテキストのタグ付け結果が表示される．テキストにはキー概念タグが埋め込まれており，初期状態では，すべてのキー概念に該当する箇所が色付け表示されている．知識辞書上で検証したい表現を指定すると，指定された表現に該当する箇所のみが色付け表示される．また，検証したい概念クラスやキー概念を指定すると，その子ノードに含まれる表現に該当する箇所のみが色付け表示される．

この機能により，意図する概念が洩れなく抽出されているか，意図しない表現が抽出されていないか，抽出結果を確認しながら辞書を調整していくことができる．また，検証ウインドウは複数起動できるので，以前の辞書によるタグ付け結果の差分を確認することも容易になる．

知識辞書上での編集機能は作成機能と同様である．

4 ツールの利用効果

4.1 下流工程の詳細

ツールの利用効果を測定するため，図 2 に示す辞書構築プロセスにおける下流工程の詳細を述べる．図 3〜5 にフローチャートを示す．図において，各プロセス中に埋め込まれている (Y) のような文字列は，そのプロセスの所要時間，a_iや b_iは繰り返し回数を表している．

辞書構築メインプロセスの処理フローを図 3 に示す．必要な概念クラスをすべて創出してから (ステップ S02〜S05)，各概念クラスに対する処理を行う (ステップ S08)．

ステップ S08 の処理の詳細フローを図 4 に示す．各概念クラスに対する処理では，その概念クラスを辞書に登録し (ステップ S10)，その概念クラスで必要なキー概念をすべて創出してから (ステップ S12〜S15)，各キー概念に対する処理を行う (ステップ S18)．

ステップ S18 の処理の詳細フローを図 5 に示す．各キー概念に対する処理では，そのキー概念を辞書に登録し (ステップ S20)，対応する表現の候補集合を選出してから (ステップ S21)，各表現に対する処理を行う (ステップ S24)．各表現に対する処理では，その表現を正規表現に変換して (ステップ S28)，辞書に登録し (ステップ S29)，登録した正規表現で意図するキー概念が正しく抽出できるかどうかのチェックを行う (ステップ S30〜S33)．

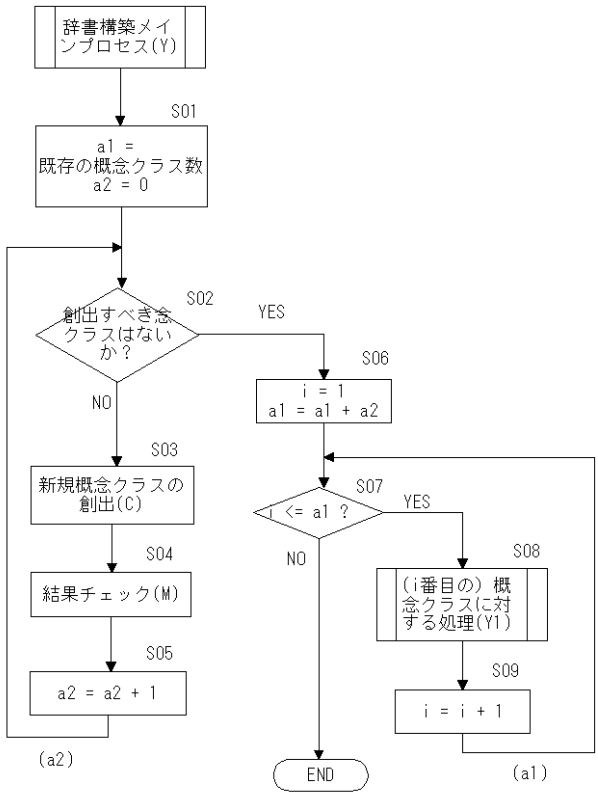

図 3: 辞書構築の詳細フロー (1)

4.2 所要時間の定式化

辞書構築メインプロセスの所要時間 (Y) を，以下のように定式化する．

$$Y = a_1 Y_1 + a_2(C + M) \tag{1}$$

$$Y_1 = a_3 Y_2 + a_4(K + M) + C_r \tag{2}$$

$$Y_2 = b_1(L + a_5 Y_3 + M) + K_r \tag{3}$$

$$Y_3 = R + E + P_l + b_2(P + M) \tag{4}$$

ここで，アルファベット大文字は各プロセスの所要時間，a_iは辞書内容に依存する係数，b_iは知識エンジニアのスキルに依存する係数である．変数の詳細を表 1 に示す．

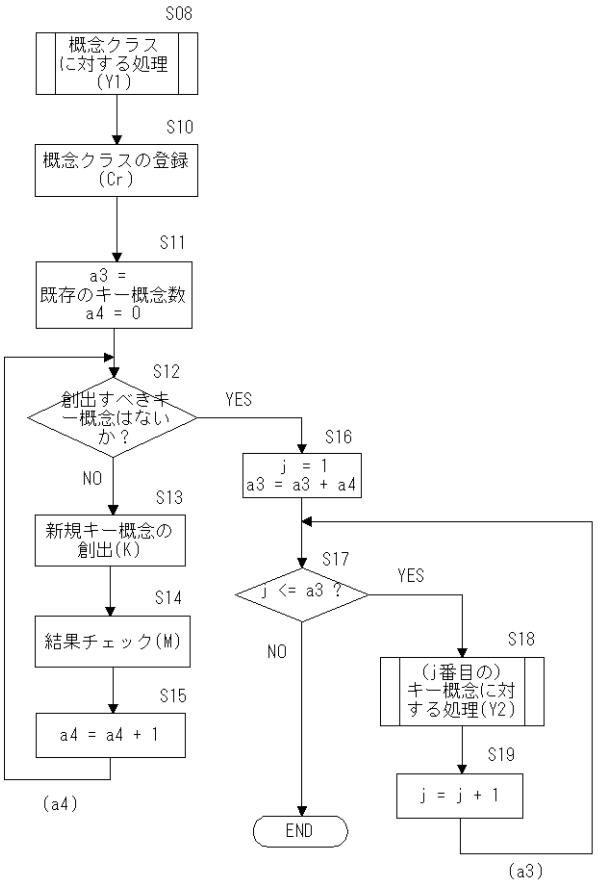

図 4: 辞書構築の詳細フロー (2)

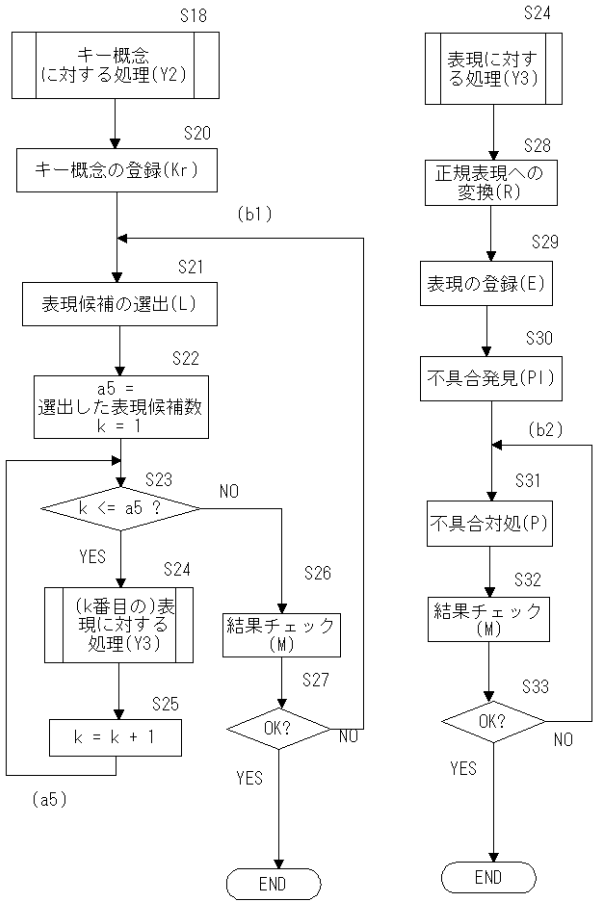

図 5: 辞書構築の詳細フロー (3)

4.3 係数の決定

4.2 節で述べた定式化に対して，実例 [1]-[3] に基づき，表 2 のように係数を決定した．

- 概念クラス総数は 12 個であったので，$a_1 = 12$ とした．
- 概念クラスのうち，9 割はユーザから与えられ，残り 1 割を新規に創出したので，$a_2 = a_1 \times 0.1 = 0.12$ とした．
- キー概念総数は 290 個であったので，1 個の概念クラスあたりのキー概念数を $a_3 = 290/12 = 24.2$ とした．
- 1 個の概念クラス内のキー概念のうち，1 割はユーザから与えられ，残り 9 割を新規に創出したので，$a_4 = a_3 \times 0.9 = 21.8$ とした．
- 表現総数は 1000 個であったので，1 個のキー概念あたりの表現数を $a_5 = 1000/290 = 3.45$ とした．
- b_1 と b_2 は知識エンジニアのスキルに依存するトライ数で，経験に基づき決定した．

4.4 所要時間の測定とツールの利用効果

ツールを利用した場合と利用しない場合との各プロセスの所要時間を測定し，つぎに，それらの値と表 2 の係数を式 (1)〜(4) に代入して Y を計算した．表 3 に結果を示す．ツールの利用により所要時間が短縮された箇所は，太字で記述してある．

図3〜5 に示した辞書構築メインプロセスに要する時間は，ツールなしの場合で 244.51 時間，ツールありの場合で 155.31 時間であった．すなわち，ツールの利用により 89.2 時間短縮でき，36.5%の効率改善が見られた．

効率化の要因はつぎの 2 点にある．

(1) 登録作業を機械化した．従来はエディタで文字列を入力していたが，ツールではリストからの選択操作で登録が行えるようになった．

表 1: 変数の詳細

変数	意味
Y	辞書構築メインプロセスの処理時間
Y_1	1 個の概念クラスの処理時間
Y_2	1 個のキー概念の処理時間
Y_3	1 個の表現の処理時間
C	1 個の新規概念クラスの創出時間
C_r	1 個の概念クラスの登録時間
K	1 個の新規キー概念の創出時間
K_r	1 個のキー概念の登録時間
L	1 個のキー概念の表現候補の選出時間
R	1 個の表現の正規表現への変換時間
E	1 個の表現の登録時間
P_l	1 個の表現の不具合発見時間
P	1 個の不具合対処時間
M	結果チェック時間
a_1	概念クラス数
a_2	新規創出概念クラス数
a_3	1 概念クラスあたりのキー概念数
a_4	1 概念クラスあたりの新規創出キー概念数
a_5	1 キー概念あたりの表現数
b_1	1 キー概念あたりのトライ数
b_2	1 表現あたりのトライ数

表 2: 係数の決定

係数	値
a_1	12
a_2	$f_1(a_1) = f_1(12) = 0.1 \times 12 = 0.12$
a_3	$290/12 = 24.2$
a_4	$f_2(a_3) = f_2(24.2) = 0.9 \times 24.2 = 21.8$
a_5	$1000/290 = 3.45$
b_1	2 〜 5
b_2	1 〜 3

表 3: 各プロセスの所要時間

プロセス	ツールなし (秒)	ツールあり (秒)
C	60	60
$\boldsymbol{C_r}$	**30**	**14**
K	60	60
$\boldsymbol{K_r}$	**30**	**14**
$\boldsymbol{L}$	**60**	**45**
R	6	6
$\boldsymbol{E}$	**30**	**14**
P_l	6	6
P	30	30
$\boldsymbol{M}$	**205**	**120**
$\boldsymbol{Y}$	**244.51 (時間)**	**155.31 (時間)**

(表 3 における C_r, K_r, E が短縮された．)

(2) タグ付け結果を確認する手段を提供した．従来はタグ付け結果を確認する有効な手段はなく，結果ファイルを上から順に目で見てチェックしていた．ツールではチェックしたいキー概念に対応する箇所が色付け表示されるので，視認性が大幅に向上した．(表 3 における L, M が短縮された．)

5 まとめ

テキストマイニング用知識辞書の構築支援を目的として，知識辞書構築支援ツール CADDIE を開発した．本ツールは，言語的専門知識を有しない知識エンジニアの下流工程における作業を支援するもので，表現リストからの辞書作成とタグ付けによる辞書検証の機能を備えている．

ツールの利用効果を測定したところ，ツールなしで 244.51 時間を要する工程を 155.31 時間に短縮でき，36.5%の効率改善が見られた．

今後は，現在支援できていないプロセスへの支援として，下流工程における知的な支援を検討していく予定である．

参考文献

[1] 市村由美, 中山康子, 赤羽俊男, 三好みよ子, 関口寿一, 藤原庸祐. “営業日報を対象としたテキストマイニング —成功事例および機会損失情報の抽出—”. 人工知能学会 第 14 回全国大会, 26-06, 2000.

[2] 市村由美, 中山康子, 赤羽俊男, 三好みよ子, 関口寿一, 藤原庸祐. “営業日報を対象としたテキストマイニングのための知識辞書の構築”. 情報処理学会 第 61 回全国大会, 5N-7, 2000.

[3] 市村由美, 中山康子, 赤羽俊男, 三好みよ子, 関口寿一, 藤原庸祐. “日報分析システムの開発”. 電子情報通信学会 技術研究報告 NLC2000-26, pp.31-38, 2000.

[4] 市村由美, 長谷川隆明, 渡部勇, 佐藤光弘. “テキストマイニング — 事例紹介”. 人工知能学会誌, Vol.16, No.2, pp.192-200, 2001.

[5] 那須川哲哉. “コールセンターにおけるテキストマイニング”. 人工知能学会誌, Vol.16, No.2, pp.219-225, 2001.

[6] 渡部勇. “ビジュアルテキストマイニング”. 人工知能学会誌, Vol.16, No.2, pp.226-232, 2001.

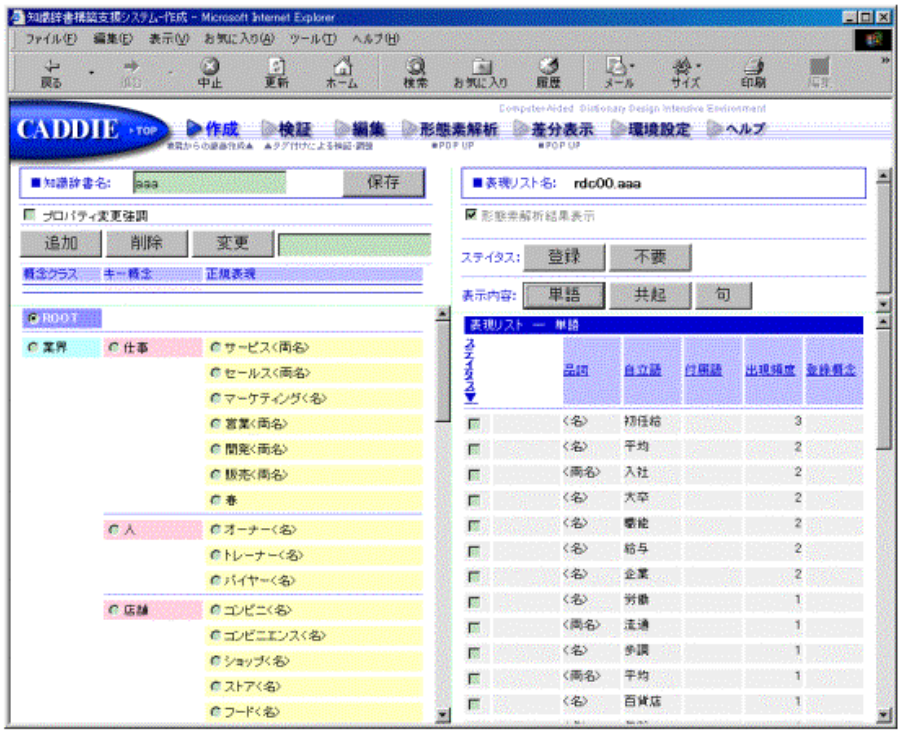

図 6: 作成画面

図 7: 検証画面