2006－NL－173（7）
2006／5／19

自動通訳に向けた用語自動収集

後藤功雄[†*]　加藤直人[†*]　黒橋禎夫[††**,†]　松原茂樹[†††,†]
[†]ATR 音声言語コミュニケーション研究所
[††]東京大学　[†††]名古屋大学

自動通訳の事前準備として，講演に関連する専門用語を自動的に収集する手法について述べる．用語の自動収集は，情報検索，語の抽出，語の重み付けを組み合わせることにより行う．本手法は講演タイトルを用いて新聞記事から用語を収集する．自動収集の評価は，人手で作成した正解データと比較することで行った．また，実際の講演に出現した用語のうちどの程度をカバーしているかについての確認を行った．

Automatic methods of collecting terms for simultaneous machine interpretation

Isao Goto[†*], Naoto Kato[†*], Sadao Kurohashi[††**,†], and Shigeki Matsubara[†††,†]
[†]ATR Spoken Language Communication Research Laboratories
[††]The University of Tokyo　[†††]Nagoya University

This paper proposes automatic methods of collecting terms for simultaneous machine interpretation of monologues. The proposed methods consist of IR, term extraction, and term weighting. Given titles of monologues, the methods collect terms from newspaper articles. Evaluation was performed by comparing the results with gold standard data that were produced by human. And we checked how many terms in the monologues were covered by the collected terms.

1　はじめに

独話の講演を自動通訳する研究を行っている．翻訳の場合は，未知の専門用語が出現した時点で，時間をかけてその訳を獲得することが可能である．しかし，通訳の場合は，翻訳結果をすぐに出力しなければならないため，そのような作業は困難である．そこで，事前準備として講演に関連する専門用語を対訳辞書に登録しておかなければならない．そのためには，まず原言語側で登録すべき用語を収集することが必要である．本稿では，収集する専門用語の設定と収集手法，収集結果の評価について述べる．

2　収集する用語の設定

ある講演を自動通訳する場合を考える．ここでは，講演原稿は利用できず，講演タイトルだけが分かっている場合を仮定する[1]．講演原稿が利用できないので[2]，講演中にどのような専門用語が出現するかは分からない．そこで，講演タイトルに関連する分野の専門用語を収集する．本稿では，この収集すべき専門用語を，「その分野の概念を表すことば」と設定[1]し，「特徴用語」と呼ぶことにする．

3　特徴用語収集手法

特徴用語の収集は，情報検索，語の抽出，語の重み付けという３つの手法を組み合わせることにより行なう．はじめに，大量の文書から講演タイトルを検索質問として関連する文書を検索する．次に，その関連する文書中の語を抽出する．その後，語を重み付けして順位付けする．上位の語を特徴用語とする．以下，それぞれについて説明する．

3.1　情報検索

検索手法として，Okapi BM25[2]を用いた．キーワードの単位には形態素を用い，機能語は検索質問から除いた．適合性フィードバックは利用しなかった．以下では，講演の話題に一致する文書

---

[*] 現在，NHK 放送技術研究所
[**] 現在，京都大学
[1] タイトル以外に概要などが分かっている場合は，それらを本手法での情報検索の検索質問に利用することで，タイトル以外の情報も活用することができる．
[2] 講演原稿が利用できる場合でも，講演者が原稿から逸脱して話す場合や講演者との質疑応答の中で，原稿に含まれない用語が出現する可能性がある．そのため，講演分野の専門用語を登録しておく必要がある．

表１　語の抽出の例外規則

規則	例
名詞-代名詞，名詞-非自立，名詞-副詞可能，名詞-接尾-人名を除く．ただしカタカナの場合は除かない．	これ（代名詞），こと（非自立），今年（副詞可能），氏（接尾-人名）
未知語のうち，カタカナとアルファベットは名詞と見なし，それ以外は除く．	ペイオフ（未知語），ＯＤＡ（未知語），稽（未知語）
「ら」,「や」の形態素を除く．	ら，や
名詞-数＋名詞-接尾-助数詞を除く．	２２日，４８％
名詞-数のみの形態素からなる用語を除く．	２２，１０５
記号を除く．ただし，カタカナ表現に囲まれた・は除かない．	◇＝「」、。“ ”～
名詞-固有名詞＋名詞-接尾の後は用語を区切る	東京（固有名詞）都（接尾）／渋谷（固有名詞）区（接尾）
用語の先頭が名詞-接尾の場合は，その形態素を除く．	東京（固有名詞）都（接尾）／下（接尾）
固有名詞＋固有名詞以外＋固有名詞の場合は，２番目の固有名詞の前で区切る	神戸（固有名詞）地検（名詞-一般）／姫路（固有名詞）支部（名詞-一般）

（ここでの「除く」とは，その形態素を用語の構成要素として扱わないことを意味する．）

を「話題一致文書」，それ以外の文書を「話題不一致文書」と呼ぶ．

情報検索は，各文書にスコアを付与して文書を順位付けするが，上位の文書のうち，どこまでが話題一致文書であるかを判定することは考慮されていない．そこで，話題一致文書を次のように判定する．

- 話題一致文書判定手法
  閾値 α と最も高い文書のスコア $S_{\max}$ を用いて，文書番号 i のスコア S_i が $S_i \geq \alpha S_{\max}$ の場合は，話題一致文書とし，それ以外は話題不一致文書とする．

我々は，時事的な話題についての講演（e.g., NHK の番組「あすを読む」）を対象としている．この時事的な性質を検索に活用する．例えば，「迷走するペイオフ論議」というタイトルの場合，「迷走」は，時期によらず出現率が一定であると考えられるのに対し，「ペイオフ」は時事的な性質があり，その話題が注目されている時期で出現率が高く，そうでない時期には低いと考えられる．そこで，時事的なキーワードが出現する文書を上位にするために次のようにする．

- 時事的な文書を検索する手法
  スコアの IDF 部分（文献[2]の BM25 の $w^{(1)}$）を計算する際に，検索対象である文書（用語収集用文書）よりも過去に書かれた長期間の文書（IDF 計算用文書）を用いる[3]．

3.2　語の抽出

語の抽出は，話題一致文書中から品詞や文節などの言語的な情報を利用して行う．ただし，自動通訳に用いる機械翻訳システムの基本辞書に登録されている語は順位付けの対象としない．

語の単位は文節を利用して決定する．文節内のうち，機能語を除いてできた単位を語の単位とする．対象を名詞類のみとし，文節内で最も後ろに位置する名詞の形態素とその形態素よりも前に位置する全ての形態素を１つの用語とする．ただし，形態素解析結果の品詞において，未知語が連続または未知語と名詞が連続している部分が文節境界となった場合は，その部分は文節の境界ではないとする．また，表１に示すいくつかの例外規則を設ける．

3.3　語の重み付け

抽出した語に対して重み付けをする．語の重み付けの考え方，重み付けで利用する語の頻度のスムージング，重み付けの指標について説明する．

3.3.1　重み付けの考え方

特徴用語の出現頻度は，話題一致文書中では，文書全体の平均と比較して高いと考えられる．この高さの程度を用語の重みとする．

3.3.2　語の頻度のスムージング

重み付けで利用する語の頻度は，情報検索のスコアに応じて，話題一致文書中の頻度の一部を次のように話題不一致文書中の頻度に割り振ってスムージングする．

- 頻度のスムージング手法
  話題一致文書 i 中の語の頻度に $S_i/S_{\max}$ をかけた値をスムージングした頻度とする．残りの $(1-S_i/S_{\max})$ をかけた値は，話題不一致文書に出現した頻度として扱う．

3.3.3　重み付けの指標

語の重み付けの指標には，文献[3]で結果が良

[3] DF が 0 の場合もあり得るが，その際は DF を 1 とする．

好であった TF・IDF または有意確率を用いる．以下，この２つの指標について説明する．

TF・IDF

話題一致文書（$S_i \geq \alpha S_{\max}$ である全ての i）中での語の頻度 TF と全文書中の語の出現文書数 DF と全文書数 N を用いて，TF×log(N/DF)の値を語の重みとする．

有意確率

話題一致記事中の語数だけランダムに文書全体から語を抽出した場合に，頻度が TF 以上の事象が偶然に起こる確率として統計的仮説検定の有意確率（p 値）を設定する．p 値は TF が大きいほど小さな値になるので，語の重みとしては-log(p 値)を用いる．

仮説は，話題一致文書中での語 t の出現率 p_1 と，文書全体での語 t の出現率 p_0 を用いて，「帰無仮説 H_0: $p_1 = p_0$，対立仮説 H_1: $p_1 > p_0$」となる．検定は，表２の２×２分割表を用いて片側検定を行なえばよい．有意確率は，フッシャーの正確確率検定(Fisher's exact test)[4]により計算[4]することができる．この方法は文献[3]の HGS による順位付けと同じである．

フッシャーの正確確率検定では，小数点以下の頻度を扱えないため，表２の合計は変化させずに，a について小数点以下を切り捨てた場合と切り上げた場合の有意確率を小数点以下の割合で重み付けて平均をとった．

表２　２×２分割表

	t の頻度	t 以外の用語の頻度	合計
話題一致文書	a	b	e
話題不一致文書	c	d	f
合計	g	h	n

4　実験

4.1 実験設定

NHK の独話の番組「あすを読む」から４つの番組を対象として特徴用語を抽出する実験を行った．その番組タイトル（講演タイトル）は，(a):「消費者契約法制定へ」，(b):「定期借家制度導入へ」，(c):「難航する医療保険改革」，(d):「迷走するペイオフ論議」である．「あすを読む」はそのときの時事的なトピックを扱っているので，コーパスには同じようなトピックを扱っている毎日新聞を用いた．この中で，放送日（1999 年）より過去１年分を用語収集用文書として，1991～1995 年の５年分を IDF 計算用文書として用いた．機械翻訳システムの基本辞書として，茶筌[6]の辞書を用いた．これは対訳辞書ではないが，用語収集では原言語側のみを利用するため，ここでは対訳辞書の見出し語と仮定した[5]．形態素解析システムには茶筌[6]を用いた．閾値は経験的に $\alpha = 0.5$ に設定した．

また，語の重み付け手法としては，２種類の文書（話題一致文書と文書全体）の比較による重み付け（TF・IDF，有意確率）のほかに，１種類の文書（話題一致文書）だけを用いる C-value[7]についても実験を行った．ここで，この実験での頻度の数え方は次のようにした．ある文書から語を抽出した結果，「消費者」が１回，「消費者契約法」が１回出現した場合，TF・IDF と有意確率による重み付けでは，「消費者」が１回，「消費者契約法」が１回と数えるが，C-value の場合は部分的な一致も重複して数えて，「消費者」が２回，「消費者契約法」が１回と数える．

4.2 評価手法

評価は人手で作成した正解と比較して行う．正解は，話題一致文書と特徴用語からなる．これらは，次のように作成した．はじめに，用語収集用文書だけを用いて文書を順位付けした．そして，上位 200 から話題一致文書を人手で判別した．また，上位 200 の文書の中から，特徴用語を人手で選択した．正解は，２人の作業によって別々に２組を作成した．自動収集の評価には，recall が重要であるため，２組の正解の論理和（OR）を正解データとして用いた．これらの２組の正解の一致率は，（共通して一致するデータ数／１組の正解データ数）として計算すると，その平均は，話題一致文書は 95%，特徴用語は 53%であった．

4.3 実験結果

- 情報検索の結果

図１に正解と一致した累積文書数と文書順位の番組毎の関係を示す．情報検索の IDF 部分の計算に用語収集用文書だけを用いた場合と IDF 計算用文書を用いた場合を比較している．(d)の結果では，IDF 計算用文書を利用することで，精度が向上していることが分かる．(c)では，「医療」や「保険」は時事的な性質が少ないため，あまり

[4] 階乗の計算が必要であるが，ここでは，階乗とガンマ関数の関係と Lanczos のガンマ関数の近似[5]を利用した．

[5] 具体的には，１形態素からなる語は基本辞書に登録されていると仮定し，２形態素以上からなる語を収集対象とした．なお，解析結果の品詞が未知語の場合でも，１形態素からなる語は基本辞書に登録されていると仮定した．これは正解作成作業の負担軽減のためである．

精度に違いは見られない. 用語収集用文書だけでよい結果が得られている(a), (b)については, 同等の精度であることが分かる.

話題一致文書に判定された文書数は, 閾値 α を 0.5 とした場合に, (a):45, (b):13, (c):188, (d):97 であり, (a), (b)は少なく, (c), (d)は多い. 一方, 図１の m は正解文書のうち最下位の文書順位であり, m も(a), (b)は小さく, (c), (d)は大きい. このように話題一致文書に判定された文書数と m は相関があり, 話題一致文書判定手法は, ある程度有効であるといえる.

- 用語収集の結果

用語収集の結果を図２, 図３に示す. これらの図は, 11 点精度[8]の４番組平均の precision と recall の関係である. ただし, 長さの単位が一致しないために順位付けした語の中に正解が含まれていないものや, 自動判別した話題一致記事中に正解が含まれないものが存在したため, recall が 1 の場合は計算できないので示していない.

図２は, 閾値とスムージングの効果の検証と, TF・IDF と有意確率を比較している.

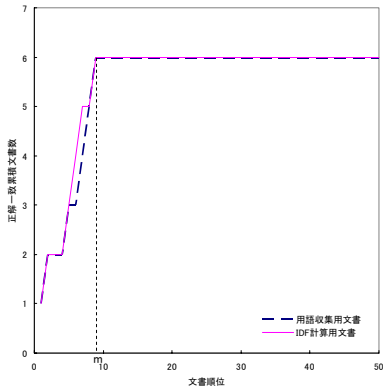
話題一致文書と話題不一致文書の判別を行わず, 情報検索のスコアが 0 より大きい文書を全て話題一致文書とし（$\alpha = 0$ の場合）, 語のスムージングを行って TF・IDF でスコアを計算した場合の結果（閾値なし（TF・IDF））と比べて, 閾値を利用して判別した他の結果は精度が良くなっている. これより, 閾値を与えて判別することが有効であることが分かる.

頻度のスムージング手法を使った場合と使わない場合を比べると, 使ったほうが精度が良い.

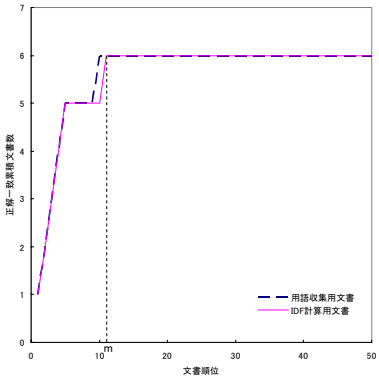
語の重み付け指標に TF・IDF を用いた場合と有意確率を用いた場合では, ほぼ同程度で大きな差は見られない.

図３は, 語の重み付け手法に C-value を用いた場合との比較を行っている. 提案手法ベース（C-value）というのは, 提案手法において, 語の重み付け手法に TF･IDF／有意確率の代わりに C-value を用いた場合である. C-value を用いた場合でも閾値による話題一致文書の判別と語の頻度のスムージングを行ったほうが精度が良くなっている. 話題一致文書だけを用いる C-value よりも, 話題一致文書と文書全体を用いる TF･IDF／有意確率のほうが少し精度が良い.

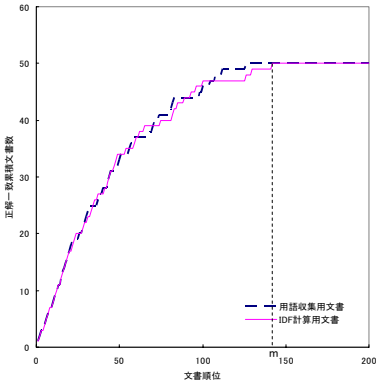
さらに, 図４の(a), (b), (c), (d)に,「あすを読む」の番組中の発話に出現した特徴用語と, 提案手法により収集した語が一致した数を示す. ここでの特徴用語についても, ２形態素以上のものだけを対象としている.「番組中の用語」が示す点は, 番組中の発話に出現した特徴用語の集合であり,



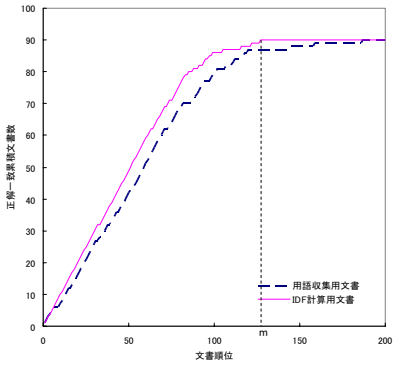
(a) 消費者契約法制定へ



(b) 定期借家制度導入へ



(c) 難航する医療保険改革



(d) 迷走するペイオフ論議

図１ 情報検索結果

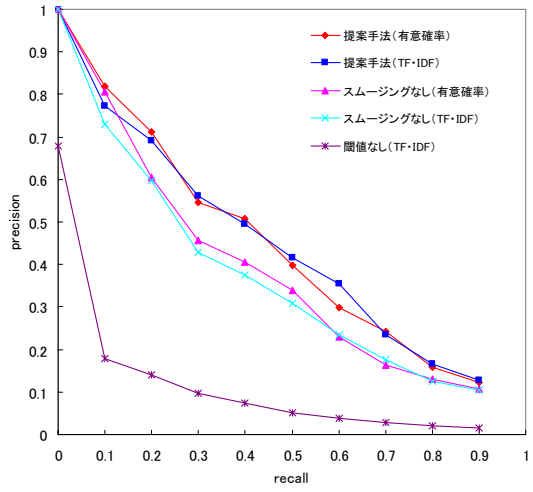


図２　用語一致率（閾値とスムージングの効果）

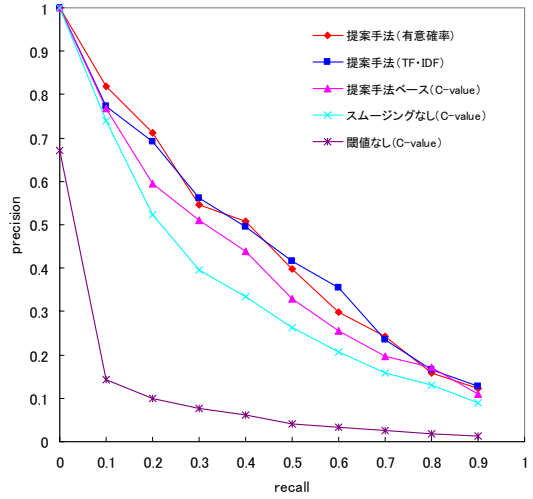
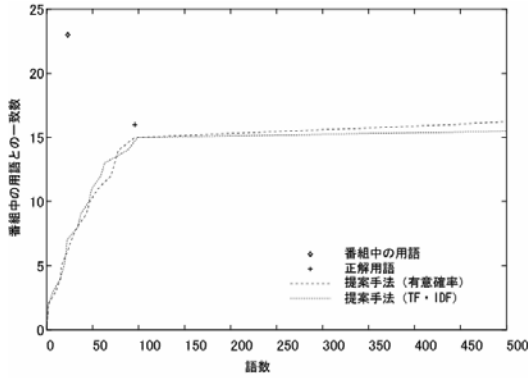


図３　用語一致率（C-value 利用の場合との比較）

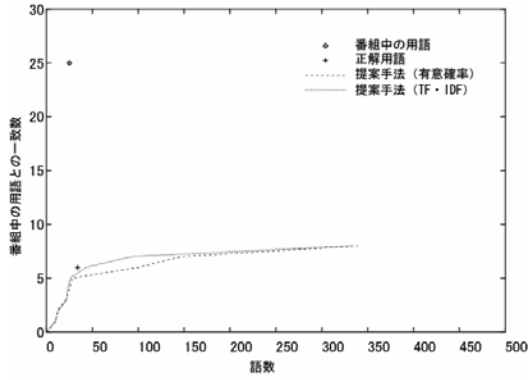
「正解用語」は，新聞から人手で選択し，図２，図３の評価に用いた特徴用語の集合である．例えば上位 200 の語では，30～70%程度の用語をカバーできていることがわかる．

なお，用語抽出の際に，用語の単位を間違うことにより，正解が抽出できないものは 34 件あった．これらは，文節が分かれてしまった場合（e.g.,一時／国有化，貸し渋り／解消）や，固有名詞＋接尾辞が出現した場合（e.g.,日本型／参照価格制度），前後の名詞類と接続して正解より長い単位になってしまった場合，記号を含む場合（e.g.,P&A），用語の一部が誤って機能語として形態素解析された場合（e.g.,ねずみ講）が原因であった．

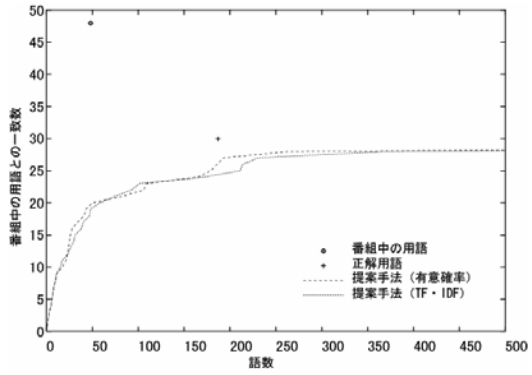
付録に「定期借家制度導入へ」について収集した上位の語を示す．正解と一致した語は，背景色を灰色で示している．



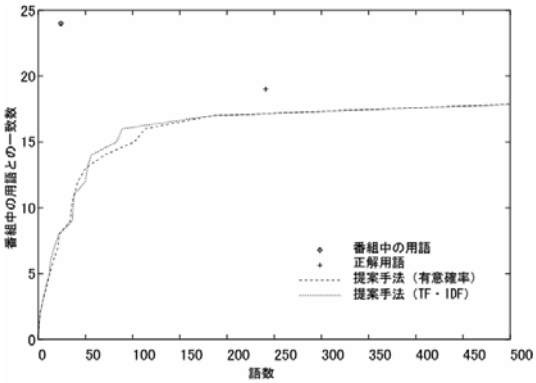
(a) 消費者契約法制定へ



(b) 定期借家制度導入へ



(c) 難航する医療保険改革



(d) 迷走するペイオフ論議

図４　番組中に出現した用語との一致数

5 おわりに

自動通訳の事前準備として, 専門用語を収集する手法について述べた．提案手法は，情報検索, 語の抽出, 語の重み付けを組み合わせて語を収集する. 新聞記事から収集した特徴用語と人手で作成した正解データとの比較を行なった. IDF 計算用文書を用いた情報検索, 閾値を用いた話題一致文書の判別, 情報検索のスコアに応じた頻度のスムージングは効果があることが分かった. 語の重み付けは，話題一致文書だけを用いる C-value よりも話題一致文書と文書全体を用いる TF・IDF ／有意確率の方が結果が良かった．TF・IDF と有意確率の結果にはあまり差が見られなかった.

謝辞

本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである.

参考文献

[1] 後藤ほか，自動通訳のための専門用語収集, 言語処理学会第 12 回年次大会，2006.
[2] S.E. Robertson and S.Walker, "Okapi/Keenbow at TREC-8," TREC-8, 1999.
[3] 久光ほか，組み合わせ的確率モデルに基づく特徴単語選択方法，NL140-12，pp.85-90, 2000.
[4] W.L. Hays, "Statistics," Holt, Rinehart and Winston, Inc., 1988.
[5] W.H. Press et al., "Numerical recipes in C [日本語版]," 技術評論社，1993.
[6] 松本裕治ほか，形態素解析システム『茶筌』version 2.2.9 使用説明書，2002.
[7] K.T. Frantzi et al., "Automatic recognition of multi-word terms: the C-value/NC-value method," International Journal on Digital Libraries, Volume 3, Number 2, 2000.
[8] I.H. Witten et al., "Managing Gigabytes," Van Nostrand Reinhold, p150, 1994.

付録　収集された上位の用語

順位	閾値なし（C-value）	閾値なし（TF・IDF）	提案手法ベース（C-value）	提案手法（TF・IDF）	提案手法（有意確率）
1	介護保険	介護保険	定期借家権	定期借家権	定期借家権
2	委員会	保険料	借家権	借地借家法改正案	借地借家法改正案
3	保険料	高齢者	借地借家法改正案	３党	３党
4	選挙区	金融機関	借家法	借家人	借家人
5	高齢者	介護保険制度	借地借家法	借家法	借家法
6	金融機関	可能性	借家法改正案	議員立法	赤井会長
7	介護保険制度	具体的	３党	赤井会長	議員立法
8	保険制度	東京都	借地借家法改正	法務委員会	法務委員会
9	小選挙区	３党	法務委員会	建設委員会	建設委員会
10	可能性	２１世紀	法改正	借地借家法	住宅政策
11	東京都	被害者	賃貸住宅	定期借家契約	定期借家契約
12	被害者	積極的	建設委員会	経済回復シナリオ	借地借家法
13	審議会	利用者	活性化	住宅政策	新法案
14	制度改革	小渕恵三首相	議員立法	制度整備	定期借家権創設
15	小渕恵三首相	中小企業	２１世紀	新法案	経済回復シナリオ
16	具体的	報告書	定期借家権創設	定期借家権創設	制度整備
17	改正案	定期的	借家人	活性化	活性化
18	比例代表	消費税	今国会	継続審議	継続審議
19	中小企業	定期預金	定期借家契約	今国会	今国会
20	両党	今国会	衆院法務委員会	産業再生	産業再生
21	法改正	現行制度	経済回復シナリオ	賃貸住宅等	契約期間
22	委員長	通常国会	特別措置法案	正当事由	同法案
23	障害者	日本版	経済戦略会議	契約期間	次期国会
24	選挙区制	定期検査	継続審議	同法案	賃貸住宅等
25	報告書	自自公	良質賃貸住宅供給促進特別措置法案	貸しオフィス	正当事由
26	３党	障害者	定期借家権導入	契約期限	早期成立
27	選挙制度	諮問機関	新法案	三井ホーム会長	貸しオフィス
28	情報公開	新制度	同法案	借地法	契約期限
29	幹事長	従業員	赤井会長	賃貸物件	中間とりまとめ
30	自自	必要性	通常国会	早期成立	衆院法務委員会
31	小選挙区比例代表並立制	政治家	住宅政策	中間とりまとめ	三井ホーム会長
32	小渕恵三	同制度	日本経済	ファミリー用	借地法
33	自自公	小選挙区	２１世紀型金融システム	２１世紀	バブル経済
34	法改正案	要介護認定	確定拠出型年金制度	バブル経済	賃貸物件
35	副大臣	問題点	区画整理事業迅速化	中古市場	導入法案
36	容疑者	副大臣	制度改革委員会新設	賃貸契約	ファミリー用
37	協議会	大阪府	設備投資促進税制導入	特別措置法案	通勤定期
38	関係者	総選挙	戦略的パイロットプロジェクト実行	自動更新	３２００市町村
39	恵三首相	不良債権	複数校選択制導入	不良債権	あり方見直し
40	２１世紀	比例代表	未来型社会資本整備	次期国会	介護等
41	積極的	公的資金	潜在成長力	通常国会	金融ルート構築
42	利用者	基本的	制度改革	臨時国会	金融産業
43	要介護	大阪市	不良債権	立ち退き料	区画整理事業迅速化
44	中選挙区	事業者	セーフティーネット	法律施行	財政資金運用
45	事業者	選挙制度	定期券	衆院法務委員会	支援税制整備
46	政府委員制度	基礎年金	賃貸住宅等	通勤定期	私的パートナーシップ
47	通常国会	介護サービス	デビットカードサービス	３２００市町村	試験評価法
48	持ち株会社	小渕首相	中古市場	あり方見直し	社債登録法廃止
49	不良債権	４０１ｋ	契約期間	介護等	重点戦略プロジェクト例
50	政治家	定期借家権	臨時国会	金融ルート構築	制度改革委員会新設