自動通訳のための専門用語収集

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

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，対訳辞書に登録されていない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分野の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について述べる．

2　収集する専門用語の設定

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．講演原稿が利用できないので[1]，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．

専門用語について検討する．専門用語（専門語）の定義として文献[1]は，１）「一般人が知る・知らない」による分類と，２）「その分野の概念を表すことば」による分類の２つの見方があると述べている．例えば，１）の分類では，「ひらがな」，「ボール」は一般語，「ヲコト点」は専門用語となるが，２）の分類では，「ひらがな」は「ヲコト点」と同様に国語学という分野の専門用語と考え，「ボール」は野球という分野の専門用語と考える．

１）の分類は，機械翻訳で特に役に立つものではない．なぜなら，一般人が知っている語でも，機械翻訳の対訳辞書に登録されていなければ収集すべき語である場合があるためである．それに対して，２）の分類は機械翻訳で必要な用語の基準として有用である．なぜなら，その分野の概念を表すことばは，その分野の話に出現する可能性が高い語であり，対訳辞書に登録されていなければ，一般人が知っているとしても登録すべき語であるからである．そこで，我々は，収集すべき専門用語を，「その分野の概念を表すことば」と設定する．本稿では「その分野の概念を表すことば」を「特徴用語」と呼ぶことにする．

[1] 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．

3　特徴用語収集手法

特徴用語の収集は，情報検索と語の重み付けという２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，その結果を用いて，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．講演タイトル中には，話題の中心ではないキーワードが含まれる場合がある．例えば，「迷走するペイオフ論議」というタイトルの場合，話題の中心は「ペイオフ」であり，「迷走」は話題の中心ではない．このような話題の中心ではないキーワードの出現文書数（DF）が，話題の中心であるキーワードの DF より小さいと検索精度が低下する原因になる．例えば，「ペイオフ」の DF より「迷走」の DF が小さければ，「ペイオフ」より「迷走」が出現する文書が上位になってしまう．ここで，「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[2].

3.2 語の重み付け

語の重み付けは，話題一致文書中の語に対して行う．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．

3.2.1 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．この大きさの程度を用語の重みとする．この方法は，２章で述べた２）の分類基準に適合していると思われる．なぜなら，この方法の場合，例えば，野球の分野の文書中で「ボール」の出現頻度が平均よりも高ければ，野球の分野の専門用語である「ボール」の重みを大きくすることができるためである[3].

3.2.2 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．

3.2.3 重み付けの指標

語の重み付けの指標には，文献[3]で結果が良好であったTF･IDFと有意確率を用いて結果を比較する．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[4].

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１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[5]することができる[6]．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１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１　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

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[7]. 閾値は経験的に $\alpha = 0.5$ に設定した．

評価は人手で作成した正解と比較した．正解は，

---

[2] DF が０の場合もあり得るが, その際は DF を１とする.
[3] 実際には，「ボール」は基本辞書に登録されていて，収集対象にはならないが, もしも基本辞書に登録されていなければ,「ボール」は登録すべき重要な語であろう.
[4] 偶然には起こらない確率“1－p 値”を用語の重みとすることも考えられるが, 引き算の際に丸め誤差が生じるため，負の対数をとるほうが望ましい.
[5] 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.
[6] 有意確率を計算する他の手法にピアソンの χ^2 検定がある．しかし，この手法で計算した有意確率は近似値であるため，フッシャーの正確確率検定を用いる方が望ましい.
[7] 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.

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．さらに，話題一致文書の中から，特徴用語を人手で選択した．選択した用語数は番組あたり平均で 74 となった．特徴用語は，機能語と記号を含まない名詞類から選択した．

特徴用語を自動的に収集する際には，語の長さの単位として文献[7]の基準を用いた．ただし，この基準では，全てのひらがなの形態素を除いているが，ここでは「ら」と「や」だけを除いた．また，茶筌の品詞の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した．

- 情報検索の結果

図２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，あまり精度に違いは見られない．話題の中心から大きく離れた語がない(a)，(b)については，同等の精度であることが分かる．

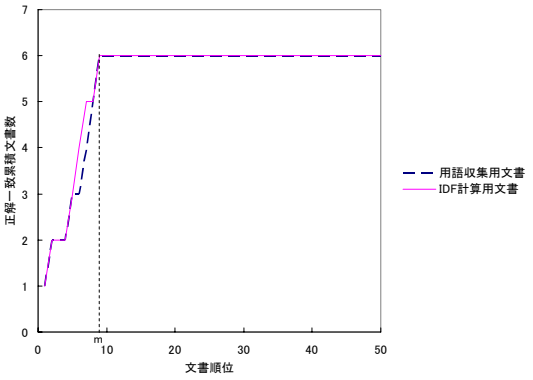
話題一致文書に判定された文書数は，(a):45，(b):13，(c):188，(d):97 であり，(a)，(b)は少なく，(c)，(d)は多い．一方，図２の m は正解と一致した累積文書数が正解文書数となった文書順位であり，(a)，(b)は小さく，(c)，(d)は大きい．このように同じ傾向を示していることから，話題一致文書判定手法は，ある程度有効であるといえる．

- 語の重み付けの結果

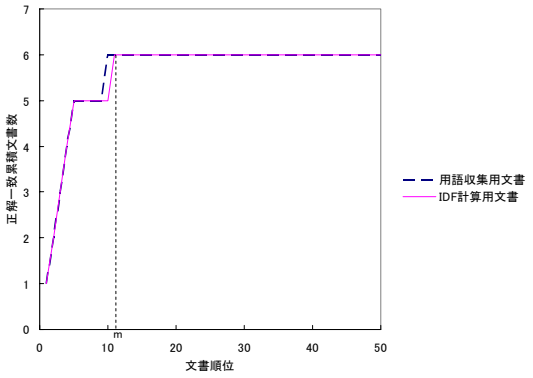
語の重み付けの結果を図３に示す．図３は，11 点精度[8]の４番組平均の precision と recall の関係である．ただし，長さの単位が一致しないために，順位付けした語の中に正解が含まれていないものが存在したが，それらは除いて recall を計算している．
話題一致文書と話題不一致文書の判別を行わず，情報検索のスコアが 0 より大きい文書を全て話題一致文書として（$\alpha = 0$ の場合）TF・IDF でスコアを計算した場合の結果（閾値なし（TF・IDF））と比べて，閾値を利用して判別した他の結果は精度が良くなっている．これより，閾値を与えて判別することが有効であることが分かる．

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．

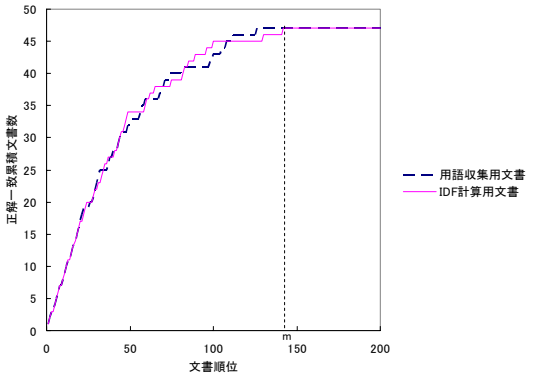
語の重み付け指標に TF・IDF を用いた場合と有意確率を用いた場合では，ほぼ同程度で大きな差は見られない．付録に提案手法による「消費者契約法制定へ」について収集した上位の用語を示す．正解と一致した用語は，背景色を灰色で示している．



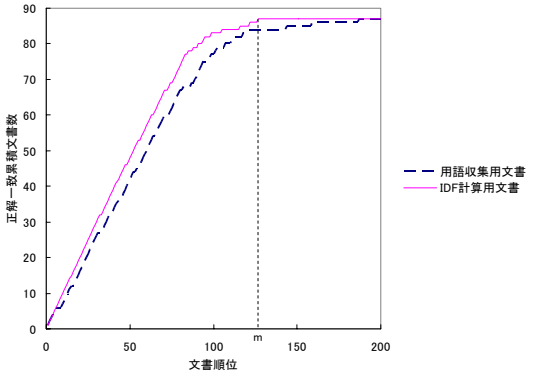
(a) 消費者契約法制定へ



(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２ 情報検索結果

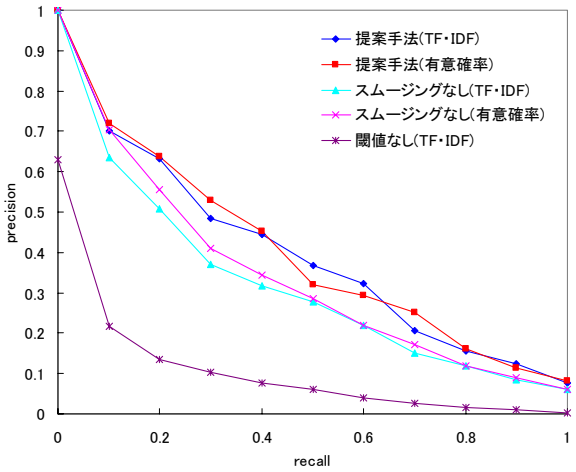


図３ 用語一致率

なお，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９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う場合が原因であった．

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．本稿では，専門用語の定義として「専門分野の概念を説明することば」用いて，特徴用語を設定した．提案手法は，情報検索技術と語の重み付け技術を組み合わせて語を順位付けする．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．

今回の実験では，正解データは１人の作業で作成したが，今後は複数の作業者によって正解を作成し，評価したい．

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] 国立国語研究所，専門語の諸問題，国立国語研究所報告 68，秀英出版，1981.
[2] S.E. Robertson and S.Walker, "Okapi/Keenbow at TREC-8," TREC-8, 1999.
[3] 久光徹，丹羽芳樹，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] 後藤功雄，加藤直人，同時通訳支援に有用な用語集についての検討，言語処理学会第 11 回年次大会，2005.
[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	TF・IDF利用	有意確率利用
1	消費者	消費者
2	消費者契約法	消費者契約法
3	金融サービス法	金融サービス法
4	消費者金融	消費者金融
5	国民生活センター	国民生活センター
6	消費者金融会社	消費者金融会社
7	事業者	消費者行政
8	利息制限法	利息制限法
9	解約料	解約料
10	悪徳商法	事業者
11	消費者行政	マルチ商法
12	英会話教室	不当条項
13	マルチ商法	悪徳商法
14	商工ローン	英会話教室
15	訪問販売法	訪問販売法
16	不当条項	国民生活審議会
17	上限金利	消費者側
18	消費者側	消費者契約法案
19	国民生活審議会	み講
20	契約内容	契約内容
21	消費者団体	商工ローン
22	重要事項	重要事項
23	最終報告	ＰＬセンター
24	マルチまがい商法	マルチまがい商法
25	み講	過剰融資
26	規制緩和	中途解約
27	同法	消費者団体
28	少年団	上限金利
29	中途解約	最終報告
30	過剰融資	契約締結
31	紛争処理機関	事業者側
32	消費者契約法案	少年団
33	ＰＬセンター	消費者契約
34	金融商品	トラブル防止
35	連帯保証人	クレジット会社
36	貸付金利	指導料
37	消費者契約	同法
38	指導料	免責決定
39	免責決定	紛争処理機関
40	クレジット会社	早期制定
41	トラブル防止	貸付金利
42	事業者側	連帯保証人
43	多重債務者	消費者保護
44	新農業基本法	多重債務者
45	早期制定	規制緩和
46	貸金業規制法	全国消費者大会
47	契約締結	金融商品
48	ローン支払い	外国語会話教室
49	消費者保護	新農業基本法
50	国民生活センター理事長	ローン支払い