同時通訳支援に有用な用語集についての検討

後藤功雄　加藤直人
{isao.goto, naoto.kato}@atr.jp
ATR 音声言語コミュニケーション研究所

1　はじめに

自然言語処理技術を使って同時通訳者への支援を行うことを検討する．ここでは，講演の原稿は事前に入手できず，講演のタイトルのみが事前に得られる場合を想定する．

通常，同時通訳者は，通訳する話の中に専門用語が含まれると想定される場合には，専門用語の対訳用語集を事前に作成し，記憶してから通訳に臨む．同時通訳では，原言語の音声を聞いてから，通訳するまでの時間がきわめて限られているという時間的制約がある．このため，辞書や用語集などの資料は，通訳している最中には基本的に利用することはできない．そのため，記憶している情報が重要となる．

そこで，支援の方法としては，用語集を作成するために有用な情報を網羅的にそして適切に順位付けして提示することが効果的だと思われる．通訳者は，その情報から必要な用語と訳語を認識し，用語の意味を理解し，最終的に自分が記憶しやすい形に再編集して独自の用語集を作成する．そのためには，用語とその訳語だけではなく，用語の簡潔な説明や，関連性の高い語彙とその関係なども同時に提供することが有用だと思われる．

本稿では，同時通訳者が用語集を作成する際に利用する情報のうち，まず基本となる用語を自動的に収集し，順位付けする手法について述べる．

2　収集目的用語の設定

ここでは，同時通訳の対象をテレビ放送の独話など，専門家向けではなく一般向けの講演の場合を考える．一般向け講演の場合，同時通訳者にとって意味が全く分からない専門用語はあまり出現しない．しかし，使用頻度の低い語や定訳のある用語を全ての分野にわたって常に正確に訳せるとは限らない．そこで通訳者は，講演のタイトルが分かっている場合には，講演内容に関連が深いが，すぐには訳せない用語を事前に調べる．

そこで本稿では，講演タイトルが与えられたときに，新聞記事から用語を収集する手法について述べる．ここで，用語収集の情報源として新聞記事を選んだのは，用語の専門性の程度が一般向けの講演と新聞記事では同程度であると考えたからである．

ここで，用語を４つの種類 A,B,C,D に分けた図を図１に示す．A は講演に関連し，通訳者が訳せない用語，B は講演に関連し，通訳者が訳せる用語，C は講演に関連せず，通訳者が訳せる用語，D は，講演に関連せず，通訳者が訳せない用語を示している．本稿で収集目的としているのは，図１で網掛け表示されているＡの部分である．

通訳者が事前に作成する用語集はそのほぼ全てが名詞である[5]ことから，収集する用語は名詞とする．

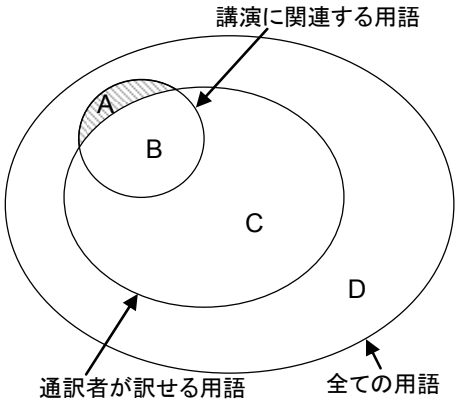

図１　用語の種類

3　自動用語収集手法

手法の概要を図２に示す．本手法では，はじめに，講演タイトルと新聞の各記事との記事関連スコアを計算する．次に，新聞記事から抽出した用語を，記事関連スコアと用語の頻度分布を考慮して，用語の順位付けを行う．以下では，これらの手法の詳細について述べる．

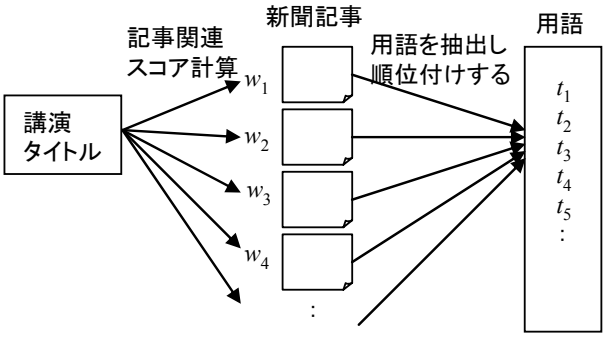

図２　手法の概要

3.1　講演タイトルと記事との関連スコアの計算

大量の新聞記事の中には，講演内容に関連性が高いものもあれば，低いものもある．講演内容に関連性が高い記事から用語を取得すれば，講演内容に関連する用語が得られると考えられる．そのために，講演タイトルと各記事との関連性を表すスコアを計算する．このスコアを記事関連スコアと呼ぶ．

表１　用語抽出の例外規則

番号	規則	例
1	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）
2	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）
3	ひらがなのみの形態素を除く．	や，つまり
4	名詞-数＋名詞-接尾-助数詞を除く．	２２日，４８％
5	名詞-数のみの形態素からなる用語を除く．	２２，１０５
6	記号を除く．ただし，カタカナ表現に囲まれた・は除かない.	◇＝「」、。“ ”～
7	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）
8	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）
9	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）

記事関連スコアの計算方法は，次の通りである．はじめに，タイトルと新聞記事を形態素解析し，機能語を削除する．次に，タイトルの前後に先頭と末尾を示すノードを追加する．さらに，タイトルにおいて任意の長さで連続する形態素列を１つのノードとし，位置関係が交差重複しないように各ノードを接続したラティスを生成する．例えば，講演タイトルが「消費者契約法制定へ」の場合，内容語の形態素は「消費/者/契約/法/制定」となる．これに対応したラティスは図３のようになる．

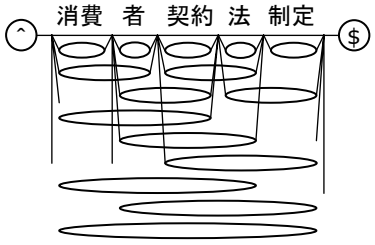

図３　タイトルから生成したラティス

各ノードのスコアを１式で計算し，ラティスの先頭から末尾までのノードのスコアの和の最大値である２式を記事関連スコアとする．w_jは記事番号 j の記事の記事関連スコアである．２式は動的計画法を利用して計算する．

$$m \times \left(\frac{1}{DF + h} \right) \tag{1}$$

$$w_j = \max \left(\sum_k m_k \times \left(\frac{1}{DF_k + h} \right) \right) \tag{2}$$

ここで，m はノード中の形態素数，DF はノードが示す形態素列を含む記事数，h は定数であり，出現頻度の違いによる影響の大きさを調整するパラメータである．k はラティスの先頭から末尾までの経路中のノードを示す．また，DF が 0 のノードは生成しない．1 式の 1/(*DF*+*g*)の部分は，多くの文書に出現する表現に対してスコアを小さくするための項である．この要素を考慮する場合には，IDF = log(*N*/*DF*)がよく知られている[2]．ここで，N は総記事数である．IDF と 1/(*DF*+*h*)の主な違いは，対数をとるかどうかの違いといえる[1]．

3.2　用語の単位設定と抽出

次に，新聞記事から用語を抽出する．その際に，用語の長さの単位を設定する必要がある．例えば，「世界貿易機構」という用語は，「世界/貿易/機構」と３つの形態素からなっているが，形態素単位に分解して，「世界」と「貿易」と「機構」に分けてしまうと，本来の意味が失われてしまう．また，分割しても本来の意味が失われない場合でも，連続する関係を提示することが支援に有用な場合も考えられる．

品詞のパターンにより単位を設定する手法[1]も提案されているが，ここでは，文節の単位を活用することにする．文節は，「意味が分かる範囲で文をできるだけ小さく区切ったときの一区切り」であるため，文節内のうち，機能語を除いてできた単位を用語の単位とする．ただし，対象を名詞に限っているので，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．これは，同一文節内において，名詞よりも前に位置する形態素は，後に出てくる名詞を修飾する関係にあると考えられ，全体で名詞としての１つの意味を示していると思われるからである．以上のように用語の単位を設定することにより，例えば，「立ち退き（動詞）料（名詞）」と解析された結果を１つの用語として抽出することができる．ただし，表１に示すいくつかの例外規則[2]を設ける．

---

[1] １式に N を掛けて $h = 0$ とすれば IDF で対数をとらない場合と同じになる．対数をとらない場合には定数 N の有無は結果に影響しない．

[2] 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．

表２　提案手法による用語

1	消費者契約法	51	消費者相談窓口
2	豆知識	52	消費者対事業者
3	不当条項	53	審議会内部
4	消費者契約	54	JTB熟年
5	消費者契約法案	55	ダイオキシン情報
6	消費者行政	56	悪質商法
7	豊田商事	57	事業者同士
8	全国消費者大会	58	消費者契約適正化事業
9	国民生活審議会	59	普及促進事業
10	威迫	60	28業種
11	菊田医師	61	日本司法書士会連合会
12	新商法	62	加入契約
13	為替証書	63	検定制度
14	家庭経済	64	子どもセンター
15	官報販売所	65	及川昭伍
16	消費者契約法制定	66	全国消費者団体連絡会
17	政府刊行物サービスセンター	67	法文化
18	西新橋中央ビル内	68	ゼロ査定
19	全官報普及サービス部	69	規制緩和時代
20	クルーズセミナー	70	仕上げ作業
21	監督人	71	代理店商法
22	代理権	72	フラオグルッペ
23	民事ルール	73	子ども放送局
24	消費者政策部会	74	業務独占資格
25	西新橋3	75	52団体
26	約160項目	76	全国子どもプラン
27	補助人	77	消費者被害
28	取り消し権	78	国民生活審議会消費者政策部会
29	事前規制	79	消費者保護策
30	利用人口	80	販売手法
31	事業者側	81	郵便切手
32	復活折衝	82	契約締結時
33	ネットショッピング	83	福祉マンション
34	国民生活センター理事長	84	市場機能
35	ソフト型	85	相対方式
36	勧誘行為	86	敏一
37	経企庁案	87	保佐人
38	取りまとめ大詰め	88	シロアリ
39	消費者取引	89	契約関係
40	日和佐信子事務局長	90	悪徳商法
41	被害急増	91	事業者団体
42	不退去	92	セントラルプラザ
43	不実告知	93	規制改革
44	目的隠匿	94	契約無効
45	ペーパー商法	95	全国消費生活相談員協会
46	預託商法	96	契約条項
47	解約問題	97	事後チェック型行政
48	契約締結手続き	98	業法
49	国生審	99	製造物責任法
50	次期国会見送り	100	法律行為

表３　IDF を用いた手法による用語

1	日本	51	住民
2	人	52	明らか
3	必要	53	理由
4	問題	54	導入
5	消費者	55	一つ
6	国	56	利用
7	企業	57	時代
8	政府	58	方針
9	米国	59	会社
10	女性	60	新た
11	国民	61	中国
12	契約	62	社会
13	指摘	63	ケース
14	自分	64	参加
15	子供	65	動き
16	自民党	66	電話
17	逮捕	67	内容
18	情報	68	責任
19	検討	69	評価
20	実施	70	自由党
21	東京都	71	反対
22	対応	72	国会
23	声	73	期待
24	意見	74	中心
25	法律	75	提供
26	事件	76	制度
27	自治体	77	意味
28	全国	78	法案
29	判断	79	厚生省
30	制定	80	銀行
31	調査	81	国旗
32	説明	82	首相
33	対象	83	実現
34	東京	84	景気
35	家族	85	同社
36	男性	86	活動
37	提出	87	君が代
38	話	88	被害者
39	成立	89	環境
40	発表	90	関係
41	主張	91	手
42	可能性	92	行政
43	自由	93	目
44	影響	94	地域
45	疑い	95	協力
46	批判	96	公明党
47	世界	97	課題
48	患者	98	商品
49	仕事	99	状況
50	議論	100	大阪市

3.3 用語スコアの計算

さらに，抽出した全ての用語に対して，用語スコアを計算し，順位付けをする．ここでは，注目している用語を t で表す．

3.3.1 用語関連頻度スコア

図１の A と B の用語は, 講演の話題に関連する記事集合での出現頻度が多いと考えられる．そこで，この要素を考慮したスコア l_t を計算する．このスコアを用語関連頻度スコアと呼び，3 式とする．

$$l_t = \sum_{j=1}^{n} w_j \times tf_j \qquad (3)$$

ここで, j は記事番号，tf_j は記事 j 中の用語の出現頻度，n は総記事数である．

3.3.2 用語特殊性スコア

図１の A と D の用語, つまり同時通訳者が訳せない用語の特徴を考える．出現頻度が少ない用語は知らない可能性が高く，特定の時期や分野だけに出現する用語は専門的で，その話題や分野に詳しい人以外は知らない可能性が高いと考えられる．このような用語は，新聞記事集合では，「頻度が少ない」「分布の偏りが大きい」という特徴を持つと思われる．そこで，これらの特徴を反映するスコア g_t を 4 式で計算する．この g_t を用語特殊性スコアと呼ぶ．4 式で利用する f_i と N_i を表４[3]に示す．ここで，λ_i は各要素の重み付けである．N_i に 1 を足しているのは 0 になるのを避けるためである．

$$g_t = \prod_i \left(1 - \frac{f_i}{N_i + 1}\right)^{\lambda_i} \qquad (4)$$

表４　頻度の種類

i	f_i	N_i
1	t の頻度	総用語数
2	t を含む記事数	総記事数
3	t を含む日数	総日数
4	t を含む月数	総月数
5	t を含むカテゴリ数	総カテゴリ数

3.3.3 用語スコア

用語関連頻度スコアと用語特殊性スコアの積を，用語の順位を決定する用語スコア s_t とする．用語スコアは 5 式となる．

$$s_t = l_t g_t = \left(\sum_{j=1}^{n} w_j \times tf_j\right) \prod_i \left(1 - \frac{f_i}{N_i + 1}\right)^{\lambda_i} \qquad (5)$$

[3] カテゴリとは，新聞の掲載面種別コード（国際，経済，家庭など）を意味する．

3.4 用語の収集例

NHK の番組「あすを読む」の中から「消費者契約法制定へ」について用語を収集する実験を行った．新聞記事には放送日前の１年分の毎日新聞記事を用いた．収集した結果を表２，表３に示す．表２は提案手法，表３は記事関連スコア，用語特殊スコアともに IDF を用いた手法の結果の上位 100 位である．形態素解析と文節認定には茶筌[4]と CaboCha[3]を用いた．パラメータ h の値はとりあえず 10 とし，λ_i は全て１とした．また，新聞記事のデータには，記事見出しと本文の 2 種類が存在するため，この情報を活用し，関連スコアを計算する際に，記事見出しと一致する表現を持つノードのスコアは 1 式の 2 倍とした．ここで，記事関連スコアに IDF を用いた手法とは，1 式の代わりに $m \times \log(N/DF)$を用いた場合である. 用語特殊性スコアに IDF を用いた手法とは, 4 式の代わりに $g_t = \log(N/DF)$を用いた場合であり，この N と DF は表４の N_2 と f_2 である．

表２と表３を比べてみると，表２の提案手法の方が，表３の IDF を用いた手法よりも普段あまり使われず，講演タイトルとの関連も深いと思われる用語が多いように見える．

本手法は，IDF を用いた場合と比較して，対数をとらないことにより，記事関連スコアについて記事数の差が大きく影響するようになり，複数の分布を考慮することで，より幅広い観点で分布の偏りを反映させることができる．

4 おわりに

同時通訳支援のための用語の収集手法について述べた．本稿では，一般向けの講演について，講演のタイトルを利用して新聞のデータから用語を収集する手法を提案した．実際の講演タイトルを用いて，新聞から用語を収集した結果例を示した．収集した用語とその順位付けが通訳支援にどの程度有効であるかの評価については，評価の方法も含めて今後の課題である．

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] Koichi Takeuchi et al., 2004, *Construction of Grammar-Based Term Extraction Model for Japanese*, 3rd International Workshop on Computational Terminology.
[2] 北研二ほか, 2002, 情報検索アルゴリズム, 共立出版.
[3] 工藤拓ほか，2002，チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，43(6)，pp.1834-1842.
[4] 松本裕治ほか，2002，形態素解析システム『茶筌』version 2.2.9 使用説明書.
[5] 米原万里，1998，不実な美女か貞淑な醜女か，新潮文庫.