テキスト分析による言語処理学会年次大会 29 年分の研究動向の調査

本間夏樹　飯村翔馬
株式会社 NTT データ数理システム
{homma,iimura}@msi.co.jp

概要

本調査では自然言語処理分野の研究動向の定量的な把握のため、言語処理学会年次大会の 29 年分の論文に対してテキスト分析を行った。先行事例では単語の集計が主だったが、本調査では係り受け表現の抽出やカテゴリ情報の付与を行うことでより実践的な情報の獲得を試みた。機関毎の発表傾向や共起分析による共著関係の可視化、研究分野やテキスト種別毎の論文数等についての分析を行い、定量的に知見を得ることを目指した。

1　はじめに

テキスト分析は、テキストデータから洞察を引き出す重要な手段である。新しい知見を得るだけではなく、定量的な結果を持って自分の感覚と比較したり、他者に共有できる情報として切り出す方法として有効である。テキスト分析の対象のデータのドメインの 1 つに「学術文献」がある。テキスト分析を行うことで、研究分野の傾向を定量的に把握したり、重要な文献の絞り込みを効率的に行ったりすることが可能になる。近年、自然言語処理分野の発展は著しく、その流れを捉えるために言語処理学会の論文を分析することは有用である。言語処理学会の論文を分析した事例として [1][2] がある。[1] では、論文誌と年次大会のタイトルと書誌情報の 10 年分が分析対象である。テキスト分析として、タイトルを Chasen で形態素解析した結果を集計している。また、双対尺度法で所属機関と分野の関係図を作成している。[2] では、年次大会のタイトル・イントロダクション・本文と書誌情報の 19 年分が分析対象である。IBM Content Analytics with Enterprise Search を使用しており、タイトルとイントロダクションから年毎の頻度上位語の抽出を行っている。また、受賞論文に相関の高い表現の抽出、増加傾向、数量の年別傾向等の分析を行っている。言語処理学会の論文を活用した事例として、NLP Corpus Search Engine (NaCSE)[3] が存在し、指定したキーワードを含む論文を検索でき、年度毎の頻度を確認することができる。他の学会の論文を分析した例として [4] がある。日本建築学会の情報・システム・利用・技術シンポジウムの論文集のタイトルを対象として、KH Coder を用いて頻度上位単語の抽出や、期間毎の共起ネットワーク図の可視化を行っている。関連事例として、自然言語処理のトップカンファレンスの論文を集計しているウェブサイト [5] では、所属機関ごとの論文数が確認できる。

本調査では、言語処理学会 29 年分の年次大会のタイトルと書誌情報 (出版年・機関) からテキスト分析を行う。先行事例では書誌情報や単語レベルの分析に留まっていた。本調査では係り受け情報を活用し、より文脈を考慮した実践的な情報を得たり、ルールベースを用いてカテゴリ情報を付与し研究分野やテキスト種別等の解釈しやすい単位で分析を実施する。また、共起を用いて共同研究の概観を把握することを目指す。

2　データ処理

2.1　データ収集

発表論文の書誌情報は、言語処理学会の各年度の年次大会のウェブサイトにアクセスし、そのプログラムからタイトルと著者の所属機関を抽出した。所属機関は第一著者のものを使用し、複数の機関が記載されている場合は左端の機関に絞った。分析データとして、抽出タイトルをテキスト、出版年と機関を属性情報として、Text Mining Studio (TMS) [6] に読み込ませた。ただし、共同研究調査の分析においてのみ各著者の全所属機関の情報を活用している。

なお、論文のタイトルは日本語が大部分であり、

英文タイトルは全体の一割未満である。英文タイトルの論文の情報も分析対象として統一して処理を行うために、OpenAI 社の GPT-4[7] を用いて日本語に翻訳を行った。翻訳結果は目視で確認し、より適切な訳が考えられる場合や後続のテキスト処理の負担を軽減できる場合は適宜修正を行った。和文・英文の両方が含まれている場合[1)]は和文のみ使用した。また、論文内で和文タイトルを発見できた場合[2)]はそれを採用し、プログラムとの差異を発見できた場合[3)]は論文ファイルのタイトルを優先した。

2.2 属性の前処理

属性情報を加工することで目的に沿った粒度や視点で分析が可能になる。出版年を次のように 5 年毎の期間に分割した。「1 期 (1995-1999 年)」「2 期 (2000-2004 年)」「3 期 (2005-2009 年)」「4 期 (2010-2014 年)」「5 期 (2015-2019 年)」「6 期 (2020-2023 年)」とした。

所属機関について、表記揺れや組織改編を考慮して把握している範囲でまとめ上げた。例えば、NICT は通信総合研究所と通信・放送機構を含むようにまとめ上げた。なお、NTT アドバンストテクノロジ、NTT データ、NTT ドコモ等は NTT に集約せずに、それぞれ別の機関として集計した。更に機関を 4 つの種別に分けた。高等専門学校・大学等は「教育機関」、公立の研究所や外郭団体等を「公的機関」、民間企業等を「企業等」、フリーや所属なしを「個人」とした。

2.3 テキストの前処理

テキストの前処理には TMS に搭載されている機能を利用した。TMS は形態素解析・構文解析結果の情報を付与している。英名の技術用語等については適宜辞書を作成した。

更に、分析のトピックに応じて、単語群をまとめ上げるカテゴリルールを作成した。例えば、『文脈解析』カテゴリは、「述語項構造」「共参照」「照応」「談話構造」等の文字列を含むタイトルはそのカテゴリであるというルールを持つ。このような研究トピック等に関連するカテゴリルールを適用することでテキストデータの分類を実施し、その集計等から研究動向を把握するのに適した粒度の情報を獲得できる。ただし、ルール作成の際に見落としがある可能性があるため、TMS に実装されている大規模テキストデータで事前学習したモデルを利用して類義語の抽出を行い、取りこぼしの低減を試みた。

1) The Relationship between Sound and Meaning in Japanese Back-Channel Grunts (あいづちの音響的部品とそれぞれの意味) (1998), An Experimental Classification of English Noun Phrases Used in Metaphorical Expressions (日本語表題: メタファに着目した英語の名詞分類の試み) (2002) の 2 件

2) Detecting Alternation Instances in a Valency Dictionary, 構文意味辞書における類似構文の融合方法 (2002) の 1 件

3) プログラム: 対話データベースの自動プロファイリング: 効率的話題タグ付与をめざして, 論文: 日本語教科書の会話に見られる言いよどみ (1998) の 1 件

3 データ分析と観察

3.1 属性の集計

年毎の発表件数の推移を表したのが図 1 である。前年と比べて減少している年もあるが、全体的な傾向として増加していることが分かる。特に 2023 年の発表件数の伸びは突出している。内訳は第一著者の所属機関の種別となっており、教育機関が過半数を占める。初期は企業の割合が多かったことが確認できる。個人は 0 人の年もあり、多い年も 10 名に満たず少ない。

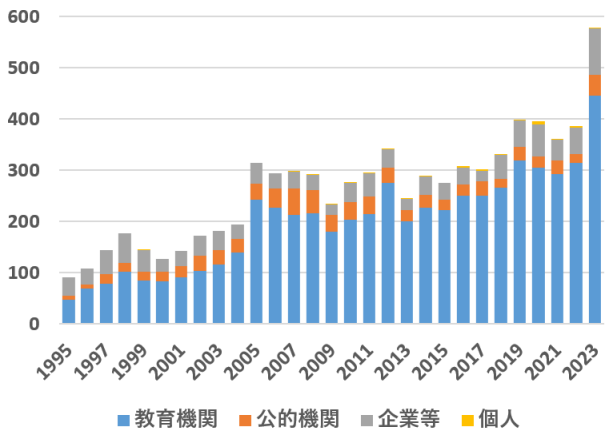

図 1 言語処理学会年次大会の発表件数の推移 (内訳は第一著者の所属機関の種別)

第一著者の所属機関の論文数を集計して、全期間の合計上位 10 機関を抽出した結果が図 2 である。教育機関の割合が高く、最大数の東大に、NAIST、京大が続く。公的機関では NICT、企業では NTT の件数が多いことが分かる。名大、筑波大は 3 期以降に件数が伸びている。

共同研究の動向を把握するために、共著者の所属機関を共起ネットワークで可視化した。分析対象とする論文は著者が二人以上で、異なる機関に所属しているものに絞った。全ての共著関係を抽出すると煩雑な図になるため、合計 8 本以上の共著論文があ

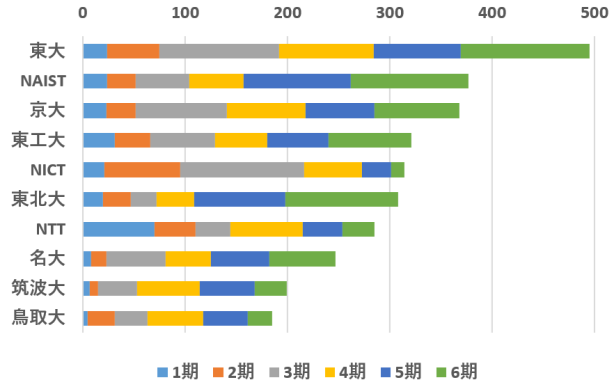

図 2　第一著者の発表数上位 10 機関

る機関で、信頼度[4]が 60 以上の関係のみに絞って抽出した結果が図 3 である。ノードの大きさが論文数を、矢印が信頼度 60 以上の共著関係を表す。東大や理研等様々な機関と共同研究しているノードには、複数の矢印が入り込む形となっている。宇都宮共和大のように複数の矢印が出ているノードは、一つの論文で複数の機関と共著関係になっていることが考えられる。

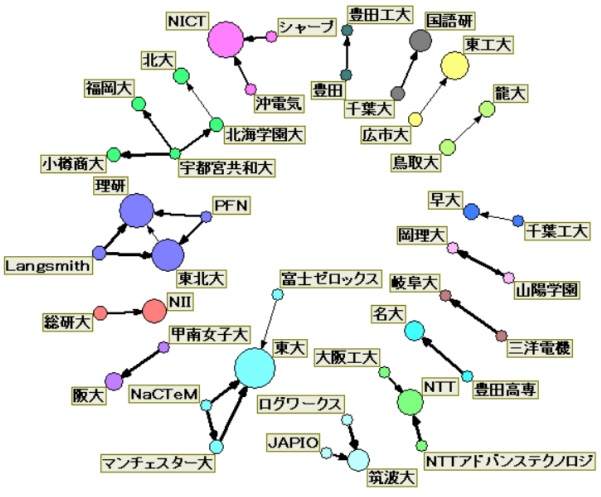

図 3　共起ネットワークによる共著関係の可視化

3.2　単語・係り受け表現の集計

タイトルを形態素解析・構文解析したものを集計した。品詞が「名詞・形容詞・形容動詞・動詞」の単語を期間毎に頻度上位 10 件抽出した結果が表 1 である。「用いる」「基づく」等論文で使用される基本的な語彙を確認できる。なお、1 つのタイトル内で同じ単語が複数回使用されている場合、その単語の頻度は 1 回として集計されている。

単一の単語では文脈から切り離されて、その意味の解釈が難しいことがある。それに対処するために、単語間の修飾関係を抽出した係り受け表現を利用して、テキストの内容をより深く把握することを試みる。係り元単語が「名詞」、係り先単語が「形容詞・形容動詞・動詞・サ変名詞」の係り受け表現を抽出した結果が表 2 である。10 年毎に頻度上位 10 件の係り受け表現が抽出されている。単語単体と比べて頻度は小さくなっている。手法に着目して係り受け表現を確認すると、1-2 期では N-gram や SVM、3-4 期では SVM、5-6 期では BERT や分散表現が使用されていることが読み取れる。

表 1　10 年毎の頻度上位 10 件の単語

1995-2004 年 (1・2 期)		2005-2014 年 (3・4 期)		2015-2023 年 (5・6 期)	
単語	頻度	単語	頻度	単語	頻度
用いる	224	用いる	443	用いる	626
基づく	117	基づく	262	基づく	340
利用	86	利用	203	構築	184
分析	51	分析	132	利用	166
日本語	44	構築	123	考慮	153
抽出	41	抽出	110	分析	134
評価	40	検討	81	検討	121
構築	36	日本語	77	向ける	106
コーパス	33	考慮	69	提案	76
作成	32	対象	66	日本語	67

表 2　10 年毎の頻度上位 10 件の係り受け表現

1995-2004 年 (1・2 期)		2005-2014 年 (3・4 期)		2015-2023 年 (5・6 期)	
係り受け	頻度	係り受け	頻度	係り受け	頻度
N-gram-用いる	5	機械学習-用いる	9	BERT-用いる	19
教師なし-学習	5	Web-利用	7	分散表現-用いる	18
検索速度-影響	5	SVM-用いる	6	文脈-考慮	13
コーパス-基づく	4	Web-用いる	6	構築-向ける	10
機械学習-用いる	4	コーパス-基づく	5	言語モデル-用いる	9
有効性-評価	4	収集-分析	5	構築-分析	9
SVM-用いる	3	開発-評価	4	分散表現-基づく	8
コーパス-利用	3	機械翻訳-利用	4	クラウドソーシング-用いる	6
印象-基づく	3	句-基づく	4	データセット-構築	6
確率モデル-基づく	3	構築-応用	4	強化学習-用いる	6

3.3　カテゴリ情報の集計

特定の視点からテキストを分析するには、カテゴリルールを作成して集計するのが有効な手段の一つである。研究動向を把握するために、研究分野のカテゴリを作成し、全期間で集計した上位 10 カテゴリの結果が図 4 である。『生成』や『対話』は、大規模言語モデルや生成モデルへの実用化が進む産業界への応用が拡大されている段階で、研究活動が活発になっている可能性がある。『生成』の次に多い『形態素・構文解析・固有表現』の論文数は期間 1-6 で他分野と比べて増減が少ない傾向から、安定した研究関心が得られていると考えられる。『埋め込み』は 5 期に急激に増加し 6 期で減少しているが、普及度が高まるにつれてタイトルに明示的に記載される頻度が減少している可能性が示唆される。

4)　{A から B への信頼度 } = {A と B の共著論文数 }/{A の共著論文数 } × 100

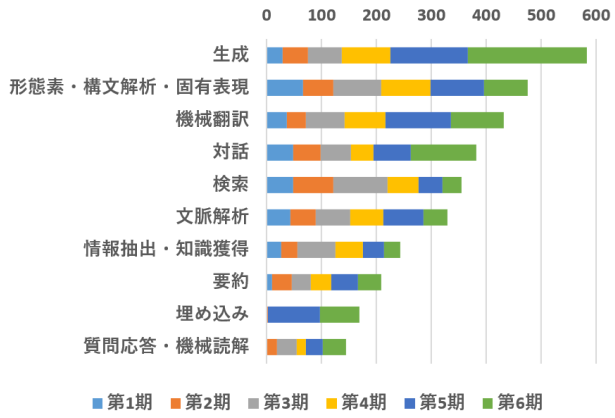

図 4　分野カテゴリの集計

研究分野のカテゴリを上位 10 機関でクロス集計した結果が図 5 である。グラフの格子上の丸の大きさは論文数を示す。横方向に確認すると該当分野でどの機関が多く発表しているか把握できる。例えば、『機械翻訳』では NAIST と NICT が多いことが見受けられる。縦方向は該当機関でどの分野で研究が活発か確認できる。東工大は『生成』に関わる論文が多い。東大は特定の分野に偏らずに幅広く研究していることが読み取れる。

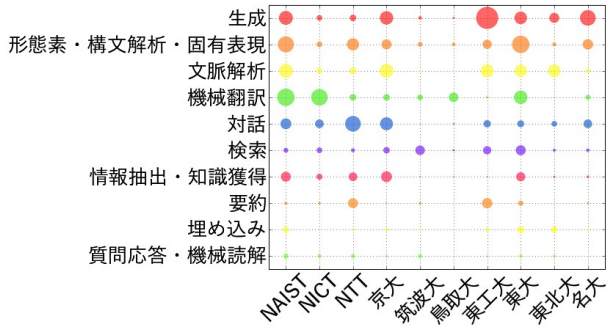

図 5　分野カテゴリと 10 機関の集計

対象のテキストの傾向把握のためにテキスト種別のカテゴリを集計した結果が図 6 である。日本の学会であるため『日本語』カテゴリが多い。なお、このカテゴリは単語「日本語」に加えて「日本語文」「日本語フレームネット」等を含むため表 1 の日本語の合計よりも多い。『外国語』は、頻度が多い順に英語・中国語・ウイグル語・韓国語・モンゴル語・スペイン語・インドネシア語と続き、30 言語以上あり多様性を確認できた。

言語処理学会ではテキスト以外のデータを対象とする場合もあるため、その傾向を確認した結果を図 7 に示す。6 期では 3 つのカテゴリで論文数が増加しており、近年マルチモーダルなモデルやタスクに注目が集まっている影響に起因すると考えられる。

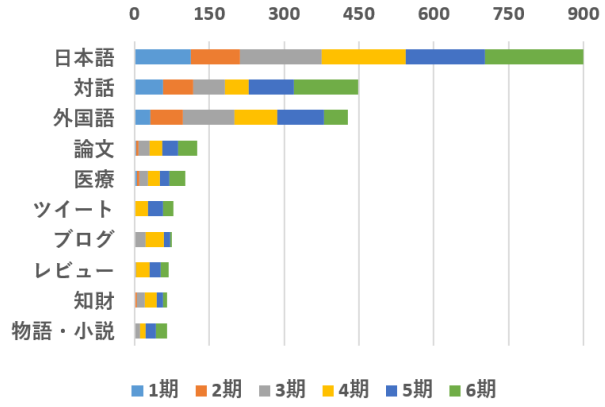

図 6　テキスト種別論文数

『画像系』の伸びが顕著なのは Text-to-Image モデルの普及が影響している可能性がある。

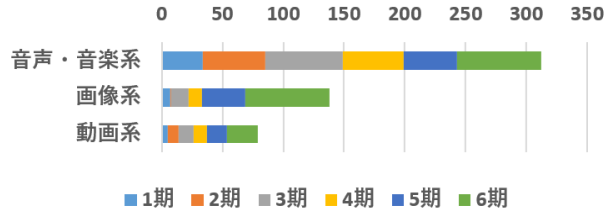

図 7　データ種別論文数

4　おわりに

本稿では、第 1 回から第 29 回までの言語処理学会年次大会のタイトルと書誌情報から研究動向を把握することを試みた。テキストは形態素解析・構文解析を行い、単語と係り受け表現の集計を行い、期間毎の傾向を確認した。またカテゴリルールを作成し、研究分野やテキスト種別といった特定の観点からデータを分析した。属性情報を加工し、共著関係の可視化を行い、機関間の関係を調べた。本調査の分析を深める方法として、異なる粒度で集計する、単語や係り受けの特徴度合い [8] を算出して比較する、特定の単語を含む係り受け表現を抽出する等が考えられる。

今後の展望として、作成した辞書やカテゴリルール等を改良し、テキスト分析対象を論文誌にも広げた文献データベースの構築を検討している。テキスト分析と大規模言語モデルを併用して効率的に調査およびデータの構造化を進め、構造化したデータベースを参照する対話システムの構築等を行うことで、これまでに到達できなかった知見を誰もが活用でき、本大会スローガン「30 年のプロンプトから未来を創造する」方向性を目指したいと考えている。

参考文献

[1] 村田真樹, 一井康二, 馬青, 白土保, 井佐原均. 過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査. 言語処理学会 第 11 回年次大会 発表論文集, 2005.

[2] 那須川哲哉, 西山莉紗, 吉田一星. 学術文献のテキストマイニング 言語処理学会年次大会 19 年分の予稿データの知的資産としての活用可能性の検討. 言語処理学会 第 20 回年次大会 発表論文集, pp. 800–803, 2014.

[3] 田淵龍二. 論文閲覧を支援する試み ― 文脈検索可能な NLP 予稿集コーパス構築. 言語処理学会 第 24 回年次大会 発表論文集, pp. 686–689, 2018.

[4] 福田知弘. テキストマイニングを用いた建築情報分野の論文タイトル分析. 日本建築学会・情報システム技術委員会第 41 回情報・システム・利用・技術シンポジウム 2018, pp. 150–153.

[5] 村脇有吾. 日本所属の言語処理トップカンファレンス論文 (2022 年). https://murawaki.org/misc/japan-nlp-2022.html.

[6] NTT データ数理システム. Text Mining Studio, 2024. https://www.msi.co.jp/solution/tmstudio/index.html.

[7] Open AI. GPT-4 technical report, 2023. https://cdn.openai.com/papers/gpt-4.pdf.

[8] 内山将夫, 中條清美, 山本英子, 井佐原均. 英語教育のための分野特徴単語の選定尺度の比較. 自然言語処理, 2004.

A　付録

参考情報としてカテゴリルールを記載する。『二重かぎ括弧』で示されるカテゴリと、「かぎ括弧」内の分類ルールから構成される。テキストは一つ以上の分類ルールが該当したカテゴリに分類される。ルールの＊記号は任意の文字列、&記号はその前後にある文字列の両方が含まれることを示す。なお、紙面の都合上一部のカテゴリルールは割愛している。

A.1　研究分野のカテゴリルール

『生成』「＊生成＊」「Ｇｅｎｅｒａｔ＊」「ｇｅｎｅｒａｔ＊」

『テキスト分析』「＊テキストマイニング＊」「＊テキスト分析＊」「＊テキストアナリティクス＊」「＊Ｔｅｘｔ　Ｍｉｎｉｎｇ＊」「＊Ｔｅｘｔ　Ａｎａｌｙｔｉｃｓ＊」「Ｔｅｘｔ&Ｍｉｎｉｎｇ」「Ｔｅｘｔ&Ａｎａｌｙｔｉｃｓ」

『形態素・構文解析・固有表現』「＊単語分割＊」「＊分かち書き＊」「＊わかち書き＊」「＊形態素解析＊」「＊サブワード分割＊」「Ｔｏｋｅｎｉ＊」「＊Ｍｏｒｐｈｏｌｏｇｉｃａｌ＊」「＊ｍｏｒｐｈｏｌｏｇｉｃａｌ＊」「Ｗｏｒｄ&Ｄｉｖｉｄ＊」「ｗｏｒｄ&ｄｉｖｉｄ＊」「品詞推定」「＊構文解析＊」「＊係り受け＊」「＊係受＊」「＊依存構造＊」「Ｐａｒｓｅ＊」「Ｄｅｐｅｎｄｅｎｃｙ」「＊句構造＊」「＊統語＊」「ｓｙｎｔａｃｔｉｃ&ａｎａｌｙｓｉｓ」「Ｓｙｎｔａｃｔｉｃ&Ａｎａｌｙｓｉｓ」「Ｓｙｎｔａｘ」「＊固有表現＊」

『音声認識』「＊音声認識＊」「＊Ｓｐｅｅｃｈ－ｔｏ－Ｔｅｘｔ＊」「ＳＴＴ」「ＡＳＲ」

『埋め込み』「＊埋め込＊」「＊分散表現＊」「＊埋込＊」「＊Ｅｍｂｅｄｄｉｎｇ＊」「＊埋め込み＊」「＊Ｗｏｒｄ２Ｖｅｃ＊」「＊ｗｏｒｄ２ｖｅｃ＊」

『可視化』「＊可視化＊」「＊Ｖｉｓｕａｌｉｚ＊」「＊ｖｉｓｕａｌｉｚ＊」

『音声合成』「＊Ｔｅｘｔ－ｔｏ－Ｓｐｅｅｃｈ＊」「ＴＴＳ」「＊音声合成＊」

『質問応答・機械読解』「＊質問応答＊」「Ｑｕｅｓｔｉｏｎ&Ａｎｓｗｅｒｉｎｇ」「＊質疑応答＊」「＊ＶＱＡ＊」「＊機械読解＊」

『機械翻訳』「＊機械翻訳＊」「＊Ｍａｃｈｉｎｅ　Ｔｒａｎｓｌａｔｉｏｎ＊」「Ｍａｃｈｉｎｅ&Ｔｒａｎｓｌａｔｉｏｎ」

『誤り認識・訂正』「＊誤り認識＊」「＊誤り訂正＊」「＊誤り検知＊」「＊校正＊」「＊スペルチェッ＊」

『文脈解析』「＊共参照＊」「＊照応＊」「＊ａｐｈｏｒａ＊」「＊Ｃｏｒｅｆｅｒｅｎｃｅ＊」「＊ｃｏｒｅｆｅｒｅｎｃｅ＊」「＊直示＊」「＊ダイクシス＊」「＊ｄｅｉｘｉｓ＊」「＊Ｄｅｉｘｉｓ＊」「＊代名詞＊」「＊指示語＊」「＊省略＊」「＊橋渡し＊」「ｂｒｉｄｇｉｎｅ&ｒｅｆｅｒｅｎｃｅ」「＊格解析＊」「＊ガ格＊」「＊述語項構造＊」「＊談話構造＊」「＊論述構造＊」「＊請求項構造＊」

『対話』「＊対話＊」「＊Ｄｉａｌｏｇｕｅ＊」

『検索』「＊検索＊」

『情報抽出・知識獲得』「＊獲得＊」「＊情報抽出＊」

『言語判定』「＊言語判定＊」

『文書分類』「＊文書分類＊」「＊文章分類＊」「＊テキスト分類＊」「＊Ｄｏｃｕｍｅｎｔ　Ｃｌａｓｓｉｆｉｃａｔｉｏｎ＊」

『要約』「＊要約＊」「＊Ｓｕｍｍａｒｉｚａｔｉｏｎ＊」

A.2　テキスト種別のカテゴリルール

『外国語』「＊Ｃｈｉｎｅｓｅ＊」「＊Ｅｎｇｌｉｓｈ＊」「＊アラビア語＊」「＊アルタイ諸語＊」「＊アジア言語＊」「＊イタリア語＊」「＊インドネシア語＊」「＊ウイグル語＊」「＊ウズベク語＊」「＊ギリシア語＊」「＊ギリシャ語＊」「＊クメール語＊」「＊シンハラ語＊」「＊スペイン語＊」「＊ソマリ語＊」「＊タイ語＊」「＊タタール語＊」「＊チェコ語＊」「＊テトゥン語＊」「＊ドイツ語＊」「＊ハンガリー語＊」「＊ハングル＊」「＊ビルマ語＊」「＊フランス語＊」「＊ブルガリア語＊」「＊ペルシア語＊」「＊ペルシャ語＊」「＊ポーランド語＊」「＊ポルトガル語＊」「＊マラーティー語＊」「＊ミャンマー語＊」「＊モンゴル語＊」「＊ルーマニア語＊」「＊ロシア語＊」「＊ロマンス語＊」「＊ロマンス諸語＊」「＊英語＊」「＊外国語＊」「＊韓国語＊」「＊中国語＊」「＊朝鮮語＊」

『ツイート』「＊Ｔｗｉｔｔｅｒ＊」「＊ツイート＊」「＊つぶやき＊」

『レビュー』「＊レビュー＊」

『物語・小説』「＊Ｆｉｃｔｉｏｎ＊」「＊Ｎｏｖｅｌ＊」「＊Ｓｔｏｒｙ＊」「＊ストーリー＊」「＊フィクション＊」「＊小説＊」「＊物語＊」

『金融』「＊Ｆｉｎａｎｃｅ＊」「＊ファイナンス＊」「＊為替＊」「＊金融＊」「＊銀行＊」「＊証券＊」「＊保険＊」

『対話』「＊対話＊」

『論文』「＊論文＊」

『SNS』「＊ＳＮＳ＊」「＊ソーシャルネット＊」「＊ソーシャルメディア＊」

『レシピ・食べ物』「＊レシピ＊」「＊食べ物＊」「＊飲み物＊」「＊飲食＊」「＊食感＊」

『日本語』「＊日本語＊」「＊Ｊａｐａｎｅｓｅ＊」

『法律』「＊法律＊」「＊刑事＊」「＊民事＊」「＊司法＊」「＊訴訟＊」「＊判決＊」「＊勝訴＊」「＊敗訴＊」「＊裁判＊」「＊条例＊」「＊調書＊」「＊原告＊」「＊被告＊」「＊被疑者＊」

『試験』「＊試験＊」「＊入試＊」

『ブログ』「＊ブログ＊」「＊ｂｌｏｇ＊」

『行政』「＊行政＊」「＊官公庁＊」「＊役所＊」「＊公文書＊」「＊公務員＊」「＊官僚＊」「＊ｅ－Ｇｏｖ＊」

『医療』「＊医＊」「＊療＊」「＊カルテ＊」「＊看護＊」「＊診察＊」「＊診断＊」「＊病＊」「＊症＊」「＊リハビリ＊」「＊患者＊」「＊Ｍｅｄｉｃａｌ＊」「＊ＭＲＩ＊」「＊健診＊」「＊Ｍｅｄｉｃｉｎｅ＊」「＊Ｈｏｓｐｉｔａｌ＊」「＊Ｋａｒｔｅ＊」「＊Ｐａｔｉｅｎｔ＊」「＊ｐａｔｉｅｎｔ＊」「＊Ｒｅｈａｂｉｌｉｔａｔｉｏｎ＊」「＊健康＊」「＊助産師＊」「＊管理栄養士＊」「＊聴覚障害者＊」「＊癒し効果＊」「＊疾患＊」

『知財』「＊知的財産＊」「＊知財＊」「＊特許＊」「＊商標＊」「＊著作権＊」