サッカー辞書の作成手法の研究

Researh on how to create a soccer dictionary

（2021年３月31日受理）

藤　本　宏　美
Hiromi Fujimoto

Key words：形態素解析，用語辞書

抄　　録

自然言語処理における日本語の解析システムは，1990年代にそれまでの研究が解析ツールとして結晶し，現在では，各種の応用システムにおいて，それらの解析ツールが入力文を解析する解析モジュールとなってきている。

そこで，機械学習を用いサッカー場におけるベイジアンネットワークを用いた有事における避難システムの開発（＝解析モデル）である。本稿では，開発システムの中で用いるための機械学習で回答の正否判定をするためのシステムの判定精度を向上するために必要不可欠である形態素解析器MeCab用のサッカー専門用語辞書とその利用について報告し，シソーラス辞書に関する構築についての研究の進捗について報告する。

1．は　じ　め　に

人間が日常で書いたり話したりする日本語や英語といった言語を自然言語という。この自然言語は，単語の集合であり，数値のデータではない。この自然言語で書かれたテキストデータをコンピュータで扱うためには，数値でデータを変換する必要がある。単語を数値で変換することができれば，機械学習や深層学習のアルゴリズムを使って学習し，文書分類やトピック抽出などに適用できる。この一連の作業技術を自然言語処理（Natural Language Processing,NLP）という。日本語の解析システムは，1990年代にそれまでの研究が解析ツールとして賜物となり，現在では，各種の応用システムにおいて，それらの解析ツールが入力文を解析する解析モジュールとなってきている。

そこで，機械学習を用いサッカー場におけるベイジアンネットワークを用いた有事避難システムの開発（＝解析モデル）である。有事の危険に対しての予測し対応する危険予知能力が必要である。しかしながら，その危険予知の予測のためには，ヒューマンエラーを考える必要もある。また，実際に訓練を行うにしても，主催者側は観客を入れた訓練になるため，大人数に協力依頼し，行わなければならない。観客側から見ると，常にスタジアムに行くわけでもなく，滅多に起こるか起こらないかわからない確率の時のため，スタジアムに行き訓練を行うのは適切ではない。そのため，教示者の指示の下で，訓練者がイラストの中に潜む危険を予測し，対策を提案する4R法が適してくるのではないかと考えられる。しかしながら，4R訓練法の問題は，教示者（この場合は主催者側の責任者）・他の訓練者（この場合はスタッフや観客）がいる必要もあり，またコロナ禍の中では，大きな訓練をすることも不可能だと思われる。作業者一人では学習もできず，訓練の機会を大きく制約されてしまう点もある。この問題を解決するためには,教示者を電算機（PC）に代替わりさせることで，教示者が必要である問題を解決し，学習における時間的・空間的制約を緩和できるデ

ジタル4R法［１］が適しているのではないかと考えられる。

本稿では，開発システムの中で用いるための機械学習で回答の正否判定をするためのシステムの判定精度を向上するために必要不可欠である形態素解析器MeCab用のサッカー専門用語辞書の必要性とその利用について進捗状況を報告する。シソーラス辞書に関する構築に関する研究について報告する。

２．自然言語処理

自然言語処理は

１．テキストデータの解析

２．解析したテキストデータの活用

の２段階がある［２］。詳細については図１に記述する。

２．１　形態素解析

STEP2の作業で行う処理の１つに形態素解析学(Morphological Analysis)がある。この形態素解析とは，文法ルールや辞書データに基づいて会話や文章を単語に分割し，それぞれに品詞を付与する処理である。コンピュータに形態素解析を実行させるツールとして，形態素解析エンジンがある。オープンソースエンジンには，ChaSen，JUMAN，MeCab，Janomeなどがある。これらのエンジンで利用することのできる辞書はIPADICが一般的に用いられている。

言葉は，意味が曖昧なもの，読む人によっても解釈が異なってしまう場合もあるので人間がやっても精度が100%にならないものが多く，コンピュータが100％の答えを返すことは難しいのが現状である。形態素解析の結果は,それ以後の処理がすべて依存することとなるので,精度の低さ（エラーなど）が生じたときには，情報の抽出，文書分類などで思わぬ結果を生むこととなる。この解釈を100％の精度に近づけるには，形態素解析エンジンで使用する言語資源・ツールである辞書が詳しくする必要性がある。例えば,竹内ら［３］の動詞項構造シソーラス[1]の登録動詞を拡張し，Lexeed内のほぼ全ての動詞,形容詞,形容動詞について概念と意味役割を付与する『語彙概念構造辞書』や述語に対して，概念(「提供」)を付与し，述語にかかる係り元(項と呼ぶ)に意味的な関係を付与する『意味役割付与システム』などの研究が進められている。

そこで本研究では，サッカーの用語に関してより細かく解釈できるようにサッカー専用辞書の作成とサッカー用語シソーラスの構築を考える。

３．サッカー専用辞書

３．１　MeCab の動作

２．１で挙げたMeCabを使って文章を単語に分割した時に，複数の単語が含まれる複合語は思ったように分割できない。特に日本サッカーについての単語については複雑なものが多い。チーム名で例を挙げると，

①Ｊリーグ・JFLのチーム名についてあげるとチーム名についてはある規則がある。Ｊリーグチームまたは，Ｊリーグ準加盟クラブなど，JFL以下のカテゴリーに所属し，Ｊリーグ加盟を目指すチームは,「名前」,「地域名」,「ＳＣやＦＣなどサッカー団体と分かる語句」の３つのうち１つ以上を組み合わせる。

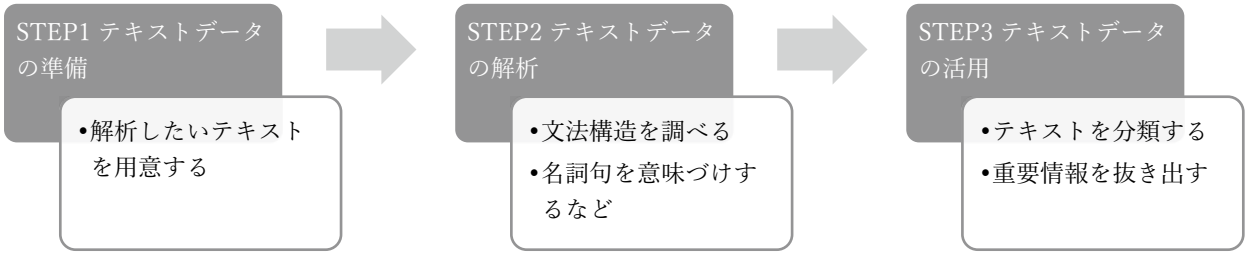

図１　自然言語処理の流れ

[1]同義語・類義語だけでなく用語間の階層関係を取り入れた辞書。シソーラスによって，システム上での活用度が大幅に向上し，用語の相互関連や位置づけが把握しやすくなり，用語の理解や文書作成に威力を発揮する。

例：
Ｊリーグチーム
・ジュビロ磐田（ポルトガル語の歓喜の意味＋所在地（静岡県磐田市））
・愛媛ＦＣ（所在地（愛媛県）＋ＦＣ）
関西サッカーリーグ１部
・おこしやす京都ＡＣ（京都弁＋所在地（京都府京都市）＋ＡＣ）
②JFL以下の企業サッカークラブは企業名が入っていても良い。
例：
・Ｈｏｎｄａ　ＦＣ（Ｈｏｎｄａのサッカー部）
・佐川急便大阪ＳＣ[2]（佐川急便大阪支社のサッカー部）
といったように実際は１つの単語として扱いたいのに，辞書に登録されていない場合は，チーム名と地域といったように分別されることも少なくない。
実際にIPA辞書のみのMeCabの動作を行なってみたところ，
ａ．チーム名
オリジナル10[3]である「サンフレッチェ広島」と途中から加盟の「ＦＣ東京」
・「サンフレッチェ広島」
サンフレッチェ広島　　名詞, 固有名詞, 組織, *, *, *, サンフレッチェ広島, サンフレッチェヒロシマ, サンフレッチェヒロシマ
・「ＦＣ東京」
ＦＣ　　名詞, 固有名詞, 組織, *, *, *, *
東京　　名詞, 固有名詞, 地域, 一般, *, *, 東京, トウキョウ, トーキョー
　オリジナル10であるサンフレッチェ広島は１つの単語（固有名詞かつ組織）として認識されているが，ＦＣ東京は分割されている。上記の結果からオリジナル10は登録されていると判断しようとしたが，確認のため行なったら, 清水エスパルスについては２つの単語に分かれた。また，途中からの加盟である柏レイソルについては１つの単語（固有名詞かつ組織）として登録されていた。
ｂ．文章
「ヴィヴィくんはＶ・ファーレン長崎のマスコットである」を形態素解析すると，表１に示す通り[4]となった。

３．２　MeCab 辞書の概要

　辞書への単語の追加は２つの方法がある。

表１　文章をMecabで解析

表層形	品詞	品詞細分類					原型	読み	発音
ヴィヴィ	名詞	一般	*	*	*	*	*		
くん	名詞	接尾	人名	*	*	*	くん	クン	クン
は	助詞	係助詞	*	*	*	*	は	ハ	ワ
V	名詞	固有名詞	組織	*	*	*	*		
・	記号	一般	*	*	*	*	・	・	・
ファー	名詞	一般	*	*	*	*	ファー	ファー	ファー
レン	名詞	一般	*	*	*	*	レン	レン	レン
長崎	名詞	固有名詞	地域	一般	*	*	長崎	ナガサキ	ナガサキ
の	助詞	連体化	*	*	*	*	の	ノ	ノ
マスコット	名詞	一般	*	*	*	*	マスコット	マスコット	マスコット
で	助動詞	*	*	*	特殊･ダ	連用形	だ	デ	デ
ある	助動詞	*	*	*	五段･ラ行アル	基本形	ある	アル	アル

[2]大阪府大阪市を本拠地としていた佐川急便大阪支社の企業サッカークラブ。同じく佐川急便の支社の佐川急便東京ＳＣと合併しSAGAWA SHIGA ＦＣとして活動をしていたが，2013年１月をもってトップチームの活動を停止．現在はSAGAWA SHIGA FOOTBALL ACADEMYという下部組織は活動している。SAGAWA SHIGA ＦＣの元所属選手やアカデミー出身者がＪリーガーとして活躍している。
[3]1992年の日本プロサッカーリーグ（Ｊリーグ）発足時に加盟した10クラブ（鹿島アントラーズ・ジェフユナイテッド市原・浦和レッドダイアモンズ・ヴェルディ川崎・横浜マリノス・横浜フリューゲルス・清水エスパルス・名古屋グランパスエイト・ガンバ大阪・サンフレッチェ広島）を指す通称。
[4]本来の正しい分割は，「ヴィヴィくん」「は」「Ｖ ・ファーレン長崎」「の」「マスコット」「で」「ある」

① システム辞書への追加

辞書更新が頻繁ではないときや，解析速度を落としたくない時には，直接システム辞書を変更するのがよい。ただし，システム辞書の更新に時間がかかる。

② ユーザ辞書への追加

辞書の更新が頻繁な場合や，システム辞書を変更する権限が無い場合には，ユーザ辞書を作る方がよい。

専門用語作成のためのフォーマット等は科学技術用語形態素解析辞書［４］等の方法を参考にしている。辞書作成については，今回追加する語句は活用語のない名詞，固有名詞である。エントリは，MeCabの解析に必要な

表層形，左文脈ID，右文脈ID，コスト，品詞，品詞細分類1,品詞細分類2,品詞細分類3,活用型,活用形,原形,読み，発音

と略して表記する場合なら略した表記名を与えている。左文脈IDは,その単語を左から見たときの内部状態IDで,右文脈IDは，その単語を右から見たときの内部状態IDである。このIDは空にしておくと，mecab-dict-index が自動的にIDを付与する。また，コストはその単語がどれだけ出現しやすいかを示していて小さいほど，出現しやすいという意味になる。MeCabでのコスト算出は単語の出現しやすさ(生起コスト)と品詞のつながりやすさ(連接コスト）のみから最適解を求めている。そのためコストの算出方法については，CRFによるモデル化とモデルからコストの算出の２通りがあり，既存の単語よりも低い生起コストで追加し，学習コーパスの少ない（または全くない）分野での専門用語辞書に対して妥当なコストを与え,専門用語の過分割を避ける必要がある。例えば，松本山雅ＦＣ（まつもとやまがＦＣ）のコストの確認をすると，表２ならびに表３[5]に示す通りになった。表２では「松本山雅」が「松→人名」,「本山→人名」,「雅（みやび）→人名」と認識となっており，所望の解析結果が出ていないことが分かる。表３のように制約付き解析によって求めたところ少しだけ改善され，「松本→地域」,「山→地域」,「雅→一般」と,ひとまず分割単位が近くなっている。この結果より最適解の累積コストは表２より27724，所望に近い解析結果の累積コストは表３の制約付き解析結果を29368となる。29368－27724＋1 ＝1645だけ「松本」の生起コストを下げることにすると,「松本山雅ＦＣ」が優先的に抽出されるようになるはずである。しかしながら，このように手作業で適切な生起コストを求めることは，登録単語が増えれば増えるだけ大変な作業となってくるので，今回は登録語の品詞を「用語を構成する中で最後に現れる単語（語根）の品詞」とし，コストを(2000－文字長×2)とした。

表２　解析における単語生起コスト，連接コスト（%pC），文頭からの累計コスト

```
松本山雅FC
松     名詞  固有名詞  人名  姓    *  *  松     マツ      マツ      10266  -1655   8611
本山   名詞  固有名詞  人名  姓    *  *  本山   モトヤマ  モトヤマ   7036  -1121  14526
雅     名詞  固有名詞  人名  名    *  *  雅     ミヤビ    ミヤビ     8814  -7009  16331
FC     名詞  固有名詞  組織  *     *  *  *                          13835   -959  29207
```

表３　制約付き解析における単語生起コスト，連接コスト（%pC），文頭からの累計コスト

```
松本   名詞  固有名詞  地域  一般  *  *  松本   マツモト  マツモト   7323   -310   7013
山     名詞  接尾      地域  *     *  *  山     サン      サン      12602  -9617   9998
雅     名詞  一般      *     *     *  *  雅     ミヤビ    ミヤビ     5834    950  16782
FC     名詞  固有名詞  組織  *     *  *  *                          13835    234  30851
```

[5]表２では標準のフォーマットの末尾に単語生起コスト（%pw），連接コスト（%pC），文頭からの累計コストを追加している。

３．３　用語辞書を利用した形態素解析

実際の簡単な文章を解析し，固有名詞とされた語を抽出した。目視による結果，おおむね抽出したい用語が抽出できていたが，同義語として判別したい語句（例えばジェフユナイテッド市原・千葉とジェフユナイテッド市原のように正式名称の変更など）については，分類ができなかった。そのため，仮想シソーラスの構築が必要となるので今後の課題としたい。

４．今後の展望

サッカー用語シソーラスの構築について優先に行う課題ある。３．３で述べたように，日本サッカーおける類義語はたくさん存在している。これらの語句をシソーラスすることによってより精度の高いテキストマイニングが行えられるといえる。その後，いくつかのKYTシートを作成し，チーム関係者や審判ならびにサポータ等に実験を行い，実験による回答より正例・負例を振り分けた成否判定を行い評価を行うことによりサッカー専用用語辞書の正確性の実験を行い精度の改善法の研究に努める。

５．おわりに

サッカー専用語句について，MeCab用のユーザ辞書を作成，サッカー分野の文書を形態素解析し，用語抽出を行った。この結果により，サッカー専用語句に関する辞書が作成できた。しかしながら，新しい語句がSNS等の普及により日々追加されているのも現状である。そのため新しい語句を更新するためにこの辞書をもとに教師あり機械学習をする必要性があること，類義語がたくさん存在していることからシソーラス辞書の作成をすることが今後の課題となる。

参　考　文　献

［１］箕輪弘嗣，竹内孔一，藤本宏美，デジタル4R訓練システム構築のための成否判定システムの最適化手法の研究，FIT2016（第15回情報科学技術フォーラム）論文集，pp. 179-180,2016.

［２］柳井孝介，庄司美沙，Pythonで動かして学ぶ自然言語処理入門，翔泳社，2019

［３］Takeuchi Lab,
http://www.cl.cs.okayama-u.ac.jp/rsc/

［４］科学技術用語形態素解析辞書,
https://dbarchive.biosciencedbc.jp/jp/mecab/data-1.html

［５］足立悠，機械学習のための「前処理」入門，リックテレコム，2019