常識判断システムのための WEB を用いたスポーツ知識の獲得

藤田晴樹　渡部広一　河岡司

同志社大学工学部知識工学科

1　はじめに

現在，社会の色々な場面でロボットは普及しつつある．ロボットと人間が上手く付き合うためには，人間のような常識判断ができるコンピュータが必要である．

コンピュータ上で人間のような常識的な判断をするための要素として，あらかじめ判断材料となる知識文を集めた知識ベースと呼ばれるデータベースがある．しかし，人手による知識ベースの作成は，情報量があまりにも膨大であるために，時間が掛かる．そのため，知識を自動的に獲得し知識ベースを作成する技術が求められている．

本稿では，スポーツに関する話題に限定して，スポーツ知識を Web から獲得する手法を提案する．提案手法を用いることで，ロボットが人間との会話においてスポーツに関する様々な常識的な判断を行うことのできる常識判断システムの構築が可能となる．

2　基本事項

2.1　常識判断システム

常識判断システム[1]とは，コンピュータ上で人間のような常識的な判断するシステムのことである．本稿では，常識判断システムの一部として，スポーツ常識判断システムを作成し使用している．

2.2　知識ベース

知識ベースとは，知識を格納したデータベースのことであり，見出し語と知識で構成されている（表 1）．

表 1：知識ベース

見出し語	知識
野球	一チーム九人ずつの二チームが守備側と攻撃側に分かれ、守備側の投手が本塁上へ投げる球を攻撃側の打者がバットで打ち得点を争う競技
野球	アメリカで発達し、日本へは明治初期に伝わった
球技	ボールを扱うスポーツのこと

2.3　概念ベース

概念ベース[1]とは，複数の国語辞書や新聞等から機械的に構築した，語（概念）とその意味特徴を表す単語（属性）の集合からなるデータベースである．概念と属性のセットにはその重要性を表す重みが付与されている．概念ベースには，現在約 9 万語の概念が格納されており，1 つの概念あたり平均 30 個の属性が存在する．

概念 A と属性 a，重み w，属性数 n の関係を式（1）で示し，表 2 で具体的な例を示す．

$$A=\{(a_1,w_1),(a_2,w_2),\cdots,(a_n,w_n)\} \quad (1)$$

表 2：概念ベース

概念	属性/重み
雪	（雪/0.61），（白い/0.30），（下る/0.27），…
白い	（雪/0.16），（白地/0.14），（色/0.14），…
下る	（低い/0.23），（雪/0.21），（雨-0.20），…

2.4　Web からの属性獲得手法

属性候補獲得手法[2]は，Webの情報からある語の属性を獲得する手法である．その手法として，まずロボット型検索エンジンで取得したWebページの文書を茶筌[3]で形態素解析を行い自立語を抽出する．抽出した自立語を属性候補とし，取得した属性候補にそれぞれTF・IDF[4]による重み付けを行うことにより精錬し属性を獲得する．

3　スポーツ概念ベース

3.1　スポーツ概念ベースの必要性

Web から取得してきた文書の中から，概念ベースの属性と重みを用いてスポーツ知識の選別を行う．しかし，現在の概念ベースはスポーツに関する語について見てみると，綺麗に精錬されているとは言えない．そこで，スポーツに関する語のみを格納しているスポーツ概念ベースを作成した（表 3）．作成方法としては手作業で作成した語をもとに Web から属性を獲得し，共起情報を元に属性の重み付けを行った．このような手法で作成した，このスポーツ概念ベースには，約 300 語の概念と 1 つの概念あたり平均 20 語の属性が格納されている．

表 3：スポーツ概念ベース

概念	属性/重み
野球	（野球/0.61），（ボール/0.30），…
打者	（打者/0.16），（野球/0.14），…
サッカー	（サッカー/0.23），（ボール/0.21），…

3.2　重み付け

作成したスポーツ概念ベースに対し汎用概念ベースと同様に概念と属性の関係の深さを示す重みを付加する．重み付けには Web の検索サイトの HIT 件数を元にした概念とその属性の共起情報を用いた．

概念 X における属性 a の重み $Weight(X,a)$は，概念 X の Google による HIT 件数 $Hit(X)$，属性 a の Google による HIT 件数 $Hit(a)$，概念 X と属性 a の Google による HIT 件数 $Hit(X,a)$ とすると，式（2）より求める．

$$Weight(X,a)=\frac{1}{2}\left(\frac{Hit(X,a)}{Hit(X)}+\frac{Hit(X,a)}{Hit(a)}\right) \qquad (2)$$

4　スポーツ知識獲得手法

4.1　獲得するスポーツ知識

本稿では，スポーツに関する普遍的な知識の獲得を提案する．表 4 にスポーツ知識の例を示す．

表 4：スポーツ知識の例

見出し語	知識
野球	一チーム九人ずつの二チームが守備側と攻撃側に分かれ、守備側の投手が本塁上へ投げる球を攻撃側の打者がバットで打ち得点を争う競技
捕手	捕手とは、野球やソフトボールで投手が投げるボールを受ける役割の選手である

4.2　Web からスポーツ知識獲得の流れ

Web には，スポーツに関する知識が数多く存在する．本稿では，Web からスポーツに関する知識のみをスポーツ概念ベースを用いて自動的に獲得する手法を提案する．図 1 に Web からのスポーツ知識獲得手法の流れを示す．

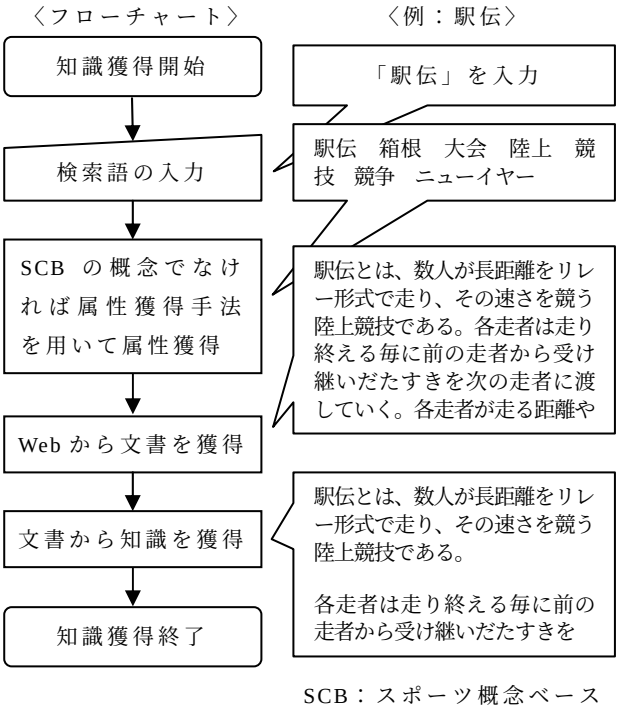

図　1：Web を用いたスポーツ知識獲得の流れ

スポーツ知識の獲得を行う語を検索語とし，検索語がスポーツ概念ベースのなかの概念になければ，Web からの属性獲得手法を用いて属性を獲得する．そして，Web から文書を獲得し，獲得した文書の中からスポーツ知識となる文をスポーツ概念ベース，あるいは先ほど獲得した属性群を用いて選別し獲得する．

4.3　Web から文書の獲得

Webから関連するスポーツ知識を含む文書を獲得する方法は，まず検索対象文書空間として，知識などの情報が整理されている国語辞典goo辞書[5]や百科事典Wikipedia[6]といったWebページや，Google[7]を用いて「検索語＋とは」で検索を行って得られる検索結果のWebページ上位 10 ページの文書を獲得する．そして獲得した文書から不要なタグを除去することで，Webからスポーツ知識を含む文書を獲得する．

4.4　文書からスポーツ知識の獲得

Web から獲得してきた文書を一文ごとに分割する．そして，分割した一文に重み付けを行い，重みの高い順に一文を最大 20 件までをスポーツ知識として獲得する．以下に，一文の重み付け手法を述べる．

一文の重み付け手法は，茶筌で形態素解析を行い自立語を抽出し，抽出した自立語に対して，その自立語の重みの総和を一文の重みとした．

一文 x の重み $SenW(x)$は，文に属する自立語の数を N，式（2）で計算した自立語 i のスポーツ概念ベースにおける自立語の重みを Wi とすると式（3）で求める．

$$SenW(x)=\sum_{i=0}^{N}Wi \qquad (3)$$

また，具体的に一文の重みを計算した例を図 2 に示す．

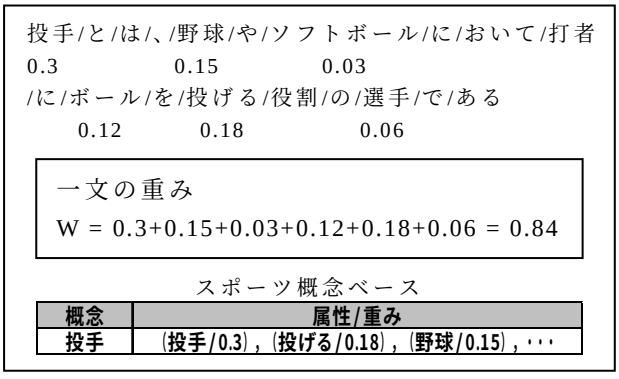

図　2：一文の重みのつけ方（例：投手）

5　スポーツ常識判断システム

スポーツ常識判断システムには，スポーツに関する語を入力すると，それに関する知識をスポーツ知識ベース内から検索し出力するタイプと，スポーツに関する質問文を入力すると，質問文とスポーツ知識ベース内の知識とを単語による表記一致により判断し，見出し語を出力するタイプの 2 つのタイプがある．本稿では，後者の質問文を入力し見出し語を出

力するタイプを用いて評価を行っている．後者を用いて具体的に判別した例を図 3 に示す．

質問文：四年に一度開催されるスポーツの祭典は？

→ 「オリンピック」

スポーツ概念ベース

見出し語	知識
オリンピック	スポーツの祭典と呼ばれ，夏･冬それぞれ四年に一度開催されている国際総合競技大会がオリンピックである
…	…

質問文は，スポーツ概念ベース内の「オリンピック」の知識と，「スポーツ」，「祭典」，「四年」，「一度」，「開催」が一致し、表記が一番近いので，見出し語「オリンピック」を出力する．

図 3：スポーツ常識判断システム

6 評価

6.1 目視での評価

6.1.1 評価方法

テストデータとして人手で作成した 221 個の見出し語（表 5）に対して，本稿で提案した手法を用いてスポーツ知識の獲得を行った．

評価は，提案手法によりスポーツ知識として判断し獲得した情報文がどれだけ正しいスポーツ知識かを割合で示した精度，スポーツ知識を獲得する対象とした全ての Web ページ中に含まれている全ての正しいスポーツ知識の内，本手法により獲得した正しいスポーツ知識の割合を示した再現率，用意した 221 個の見出し語のうち一つ以上スポーツ知識を獲得した見出し語の割合を示した精度の評価を行った．

表 5：人手で作成した見出し語の一部（全 221 個）

スポーツ	野球	選手	球技	チーム
硬式	内野手	投手	捕手	ホームラン
グローブ	ボール	変化球	サッカー	ボール

6.1.2 評価結果

提案手法により，スポーツ知識を獲得する対象とした全ての Web ページ（約 40 万文）の中から，3514 文の情報文をスポーツ知識として獲得した．その中に，スポーツ知識として正しい情報文は 318 文であった（図 4）．また，スポーツ知識を獲得する対象とした全ての Web ページ（約 40 万文）の中には，人手で判別すると 1324 文のスポーツ知識があった．その中の，318 文のスポーツ知識を獲得した（図 5）．

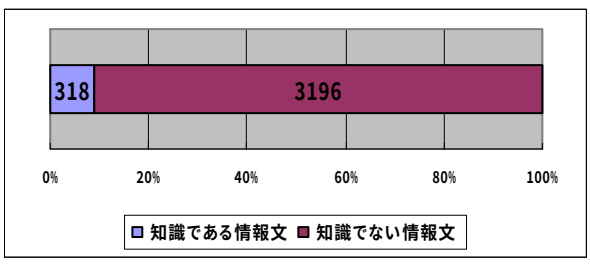

図 4：精度

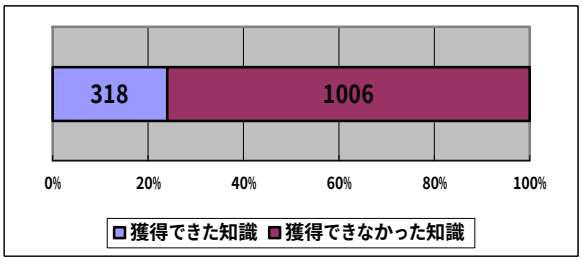

図 5：再現率

221 個の見出し語テストデータの内，122 個の見出し語は獲得した情報文の中に一つ以上スポーツ知識があった（図 6）．

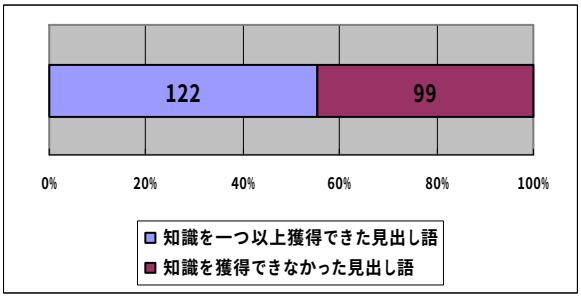

図 6：見出し語による評価

6.2 スポーツ常識判断システムでの評価

6.2.1 評価方法

スポーツ知識を一つ以上獲得した 122 個の各見出し語（表 6）に対して人手により質問文（表 7）を作成し，その質問文をスポーツ常識判断システムにかけて見出し語を出力し，どれだけの割合で出力した見出し語が正解と一致するかにより評価を行った．

表 6：スポーツ知識を一つ以上獲得した見出し語の一部（全 122 個）

スポーツ	野球	内野手	投手	捕手
一塁手	外野手	四球	三振	サッカー
ラグビー	自由形	剣道	砲丸投げ	駅伝

表 7：質問文の一部

質問文	正解
野球、サッカーなどの運動競技を総称して何と言うか？	スポーツ
投手の投げたボールを打者がバットを使って打つ、1 チーム 9 人で行う球技を何と言うか？	野球
四年に一度開催されるスポーツの祭典は？	オリンピック

6.2.2 評価結果

用意した 122 個の質問文を常識判断システムにかけると，出力した見出し語が 48 個正解と一致し，約 39%の精度で質問を答えることができた（図 7）．

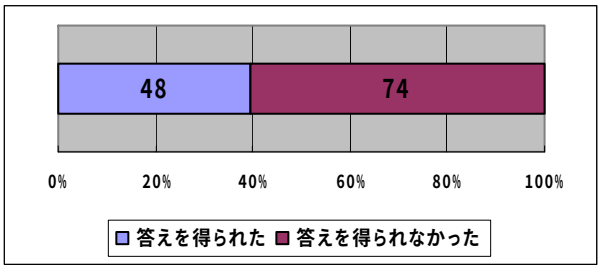

図 7：常識判断システムによる評価

7 考察

提案手法により，スポーツ知識を獲得する対象とした全ての Web ページ（総文数：約 40 万文，スポーツ知識：1324 文）の中から 3514 文の情報文をスポーツ知識として獲得し，その中に正しいスポーツ知識を 318 文獲得することができた．

図 3 左の精度の結果より，獲得した情報文のうち約 9%がスポーツ知識であり，残りの約 91%には，

・ アトリエサードとは［会社案内］
・ [サード] の使用例をウェブから探す

といった，全くスポーツと関係ない情報やノイズが含まれていた．この 2 つ情報文は「サード」のスポーツ知識として獲得したものであり，情報文の中を見ると，「サード」という単語がある．スポーツ概念ベース中の概念「サード」の属性「サード」の重みは高く，それが原因でスポーツ知識として獲得してきたと考えられる．これは，情報文の中に検索語と同じ語しか含まれていない時を考慮した一文の重み付けを行うことで解決できると考えられる．

図 3 右の再現率の結果より，全てのスポーツ知識のうち約 24%のスポーツ知識を獲得し，下のような残りの約 76%のスポーツ知識は獲得することができなかった．

・ メドレー-リレーの略
　→「メドレー」の知識として獲得できなかった
・ ドライバー - ゴルフのクラブの一つ
　→「ドライバー」の知識として獲得できなかった

「メドレー」，「ドライバー」は， Web からの属性獲得手法を用いて属性を獲得すると，その属性が「メドレー」なら音楽関係の属性，「ドライバー」なら工具関係の属性を中心に獲得したために，一文の重みを計算した時に，スポーツ知識として正しい情報文の重みを高く付けることができなかった．そのため，正しいスポーツの知識を獲得できなかったと考えられる．これは，Web からの属性獲得手法を用いて属性を獲得する時，多義性のある語を考慮した獲得を行うことで解決できると考えられる．

図 4 の全見出し語の中のスポーツ知識を獲得した見出し語の割合で評価を行うと，約 55%の見出し語は，スポーツ知識を一つ以上獲得することができた．見出し語で見た場合，この手法は，それなりに有効であると考えられる．

図 5 の常識判断システムを用いて評価を行うと，約 39%の割合で常識を判断することができ，下のような残りの約 61%の質問は判断することができなかった．

・ 雪の上をボードで滑る競技は？
　→ スキー　　（正解：スノーボード）

質問が「スノーボード」のスポーツ知識より「スキー」のスポーツ知識と表記が近いために「スキー」と出力したと考えられる．これは，常識判断システム内で，表記一致だけで判断するのではなく関連度計算なども用いることで解決できると考えられる．

8 おわりに

本稿では，Web からスポーツ概念ベースを用いて自動的にスポーツ知識を獲得する手法を提案した．その手法として，スポーツ概念ベースの自動作成とスポーツ概念ベースを用いた一文のスポーツ知識の判別を提案した．

提案手法により， スポーツ知識を獲得することはできたが，精度，再現率が低かった．そのため，精度を向上させる必要がある．また，スポーツ知識を獲得できた見出し語の割合は高かったが，常識判断システムを用いて評価を取ると，質問文に対して有効なスポーツ知識は少なかった．そのため，いろいろな種類のスポーツ知識を獲得する必要がある．

本稿では，スポーツ知識を獲得することのみを目的として研究を行ったが，スポーツ以外にも常識判断システムに必要な知識は多く存在する．そのため今後はスポーツ以外の知識獲得を可能とする汎用のシステムが必要であると考える．

謝辞

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクトにおける研究の一環として行った．

参考文献

［1］ 渡部広一，河岡司："常識的判断のための概念間の関連度評価モデル"，自然言語処理，Vol.8，No.2，pp.39-54：2001 年
［2］ 辻泰希，渡部広一，河岡司："www を用いた概念ベースにない新概念およびその属性獲得手法"，第 18 回人工知能学会全国大会，2D1-01：2004 年
［3］ 松本裕治,北内啓,山下達雄,平野善隆,松田寛,高岡一馬,浅原 正幸:”日本語形態素解析システム『茶筌』使用説明書”.奈良先端科学技術大学院大学　松本研究室：2002 年
［4］ 徳永健伸：言語と計算 5 情報検索と言語処理，東京大学出版会：1999 年
［5］ goo 国語辞典：http://dictionary.goo.ne.jp
［6］ Wikipedia：http://ja.wikipedia.org
［7］ Google：http://www.google.co.jp