常識的場所判断システムの構築－複合語の場所判断－

手原信太朗　渡部広一　河岡司
同志社大学工学部

１．はじめに

人間とコンピュータがより円滑にコミュニケーションを取れる手法が必要とされている．そのためには，コンピュータに人間と同様の常識的判断のできる能力を持たせる必要がある．その常識的判断の一つとして，場所に関しての常識的判断を行う「場所判断システム」が提案されている．本稿では，これまで提案されている「場所判断システム」[1]において，精度の低かった複合語・カタカナ・略語に対しての精度向上を図る手法を提案した．具体的には，概念ベースを用いた未知語処理の導入と新たな場所語の獲得を「場所語学習支援ツール」により行う手法である．

２．場所判断システム

場所判断システムとは，ユーザから入力された語が場所かどうかを判定し，指定された場所に関する知識を連想するシステムである．場所に関する知識とは，何が存在するかを示した「主体語」と，人間にとって何をする場所なのかを示した語を「目的語」である．本システムでは，人手で作成した「場所語知識ベース」という場所に関する知識ベースを用いる（**表１**）．しかし，全ての場所語と主体語・目的語を知識として持たせることは困難であり，効率が悪い．そこで，代表的な語のみを登録し，シソーラス[2]や概念ベース[3]により構築した連想システムを用いて知識の連想を行い，場所語知識ベースに登録されていない語に対しても対応できるようにする．

シソーラスとは，一般名詞の意味的用法を表す約2700 語の意味属性の上位下位関係，全体部分の関係を木構造で示したものであり，約 13 万語が登録されている．例を示すと，“ ビール ”の上位は“ 酒 ”である．

概念ベース（以下「CB」と呼ぶ）とは語（概念）と意味(属性)のセットを約 9 万語蓄積されている，国語辞書等から自動構築された汎用データである．**図１**に概念ベースの一部を示す．また，概念と概念の関連の深さを定量的に表す手法を関連度計算[4]といい，その値を関連度という．関連度は 0 から 1 までの連続値で，例えば，ビールとウイスキーの関連度は 0.351 と表される．

２．１．場所語知識ベース

使用頻度の高い場所語を「代表語」として場所語知識ベース（以下「場所語ＫＢ」と記す）に 443 語持たせておく．また，場所語ＫＢにはシソーラスのノードから選ばれた代表語を分類した「分類語」が 120 語存在している．それらにも「主体語」と「目的語」を与える（**表１**）．分類語と代表語は親子関係にあり，主体語と目的語の継承が可能となっている．

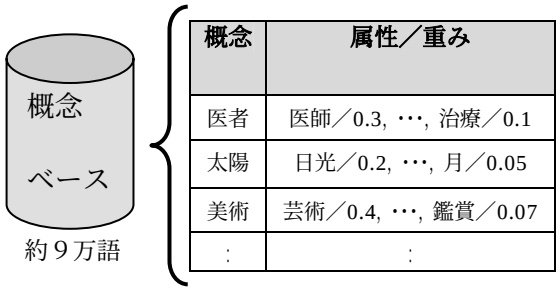

図１　概念ベース

表１　知識ベースの一部

代表語	主体語	目的語	親分類語
検察庁	検察官 etc	捜査 etc	司法官庁

分類語	主体語	目的語
司法官庁	公務員，役人 etc.	司法 etc.

２．２．従来のシステムの流れ

2.2.1 場所かどうかの判断

まず，あらゆる一般名詞（文章で入力した場合，名詞を抽出）が入力された時に，それが場所であるかどうかの判断について述べる．その処理の流れを**図２**に示す．

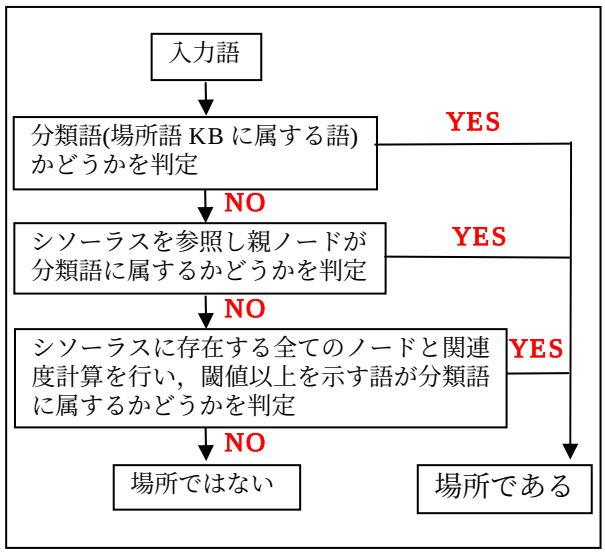

図2　場所判定のアルゴリズム

2.2.2 主体語・目的語の連想

入力語が場所語であると判断した場合，その場所語の主体語・目的語の取得を行う．そのために，入力語が場所語 KB に属するかどうか，概念ベースに属するかどうか，シソーラスに属するかどうかといった種類を特定する．その後，**図３**に示したそれぞれの手法で，主体語・目的語を取得する．

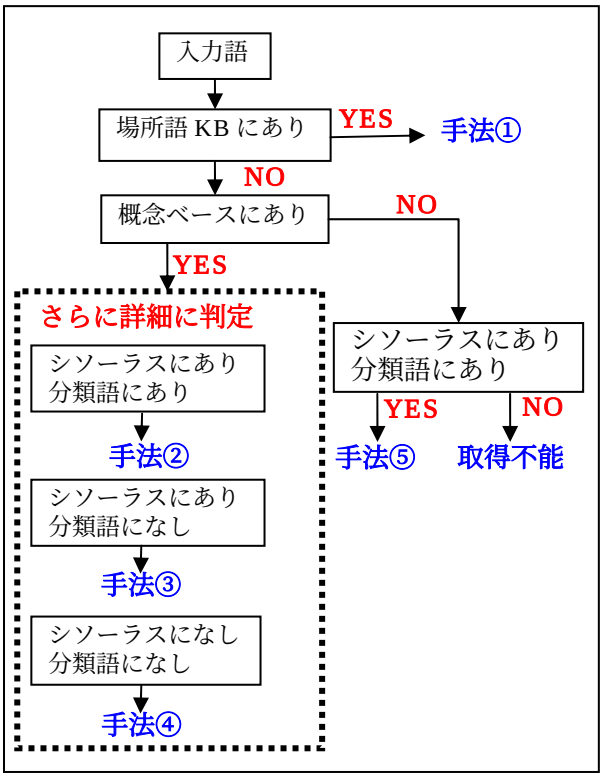

図3　入力語の種類の特定

手法①：「知識ベースに存在する語」

Ⅰ．属する分類語の主体語・目的語を取得する．
Ⅱ．その語自身の主体語・目的語を取得する．
Ⅲ．属する分類語の中に存在する全ての代表語と入力語を関連度計算し，関連度が閾値以上の語を代替語候補とする．
Ⅳ．代替語候補の全ての主体語・目的語と入力語とで関連度計算し，関連度がそれぞれ閾値以上の語を新たな主体語・目的語として追加する．
Ⅴ．これまで取得されている全ての主体語・目的語についてシソーラスによってノードを特定する．次に，入力語の属性を参照し，それらと同じノードに属する属性があるかどうかを探す．そして，同じノードに属する属性があれば閾値以上を示した語を新たな主体語・目的語として追加する．
Ⅵ．入力語の属性を展開し，ⅠとⅡで取得された主体語・目的語一つ一つとその属性を関連度計算し，関連度が閾値以上の語を新たな主体語・目的語として追加する．

以下，知識ベースに存在しない場合

手法②：「CB に存在し，シソーラスにも分類語にも存在する語」

Ⅰ．属する分類語の主体語・目的語を取得する．
Ⅱ．手法①のⅢ～Ⅵと同じ処理．

手法③：「CB に存在し，シソーラスにも存在するが，分類語には存在しない語」

Ⅰ．シソーラスを見て，シソーラス内に分類語があるかどうかを判定する．分類語があればその中で関連度計算し，最高関連度の示す語を代替語とする．
Ⅱ．手法①のⅣ～Ⅴと同じ処理．

手法④：「CB に存在するが，シソーラスにも分類語にも存在しない語」

Ⅰ．シソーラスの全最下位ノードと関連度計算をとり，関連度が閾値以上を示す語の中で，分類語が存在するかどうか判定し，存在すれば，その分類語の中に存在する全ての代表語と入力語を関連度計算し，関連度が閾値以上の語を代替語候補とする．
Ⅱ．属する分類語の主体語・目的語を取得する．
Ⅲ．手法①のⅣ～Ⅴと同じ処理．

手法⑤：「CB に存在しないが，シソーラスにも分類語にも存在する語」

Ⅰ．属する分類語の主体語・目的語を取得する．

３．従来のシステム精度

システムの精度向上にあたって，現在の精度を調べる必要がある．そこで，精度を調べるための評価データを作成した．客観的なデータを得るためにアンケートを行った．アンケートの内容は，「場所語」「場所ではない語」各 20 個ずつ考えてもらうというものである．その結果，23 人から 計 460 個のデータが集まった．その中から重複した語を削除し，「場所語」233 語，「場所ではない語」307 語に対して評価を行った．入力語に対して「場所語」を「場所である」，「場所ではない語」を「場所ではありません」と返答することを「正解」として，まず場所であるかどうかの判断が正確にできているかどうかの評価をとった．その結果を**図４**にそれぞれ示す．

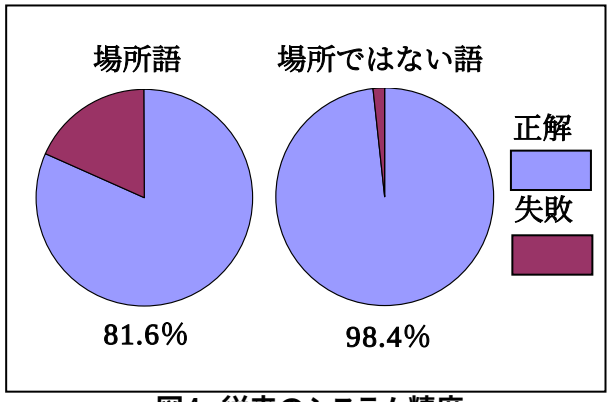

図4　従来のシステム精度

精度は場所語 81.6%，場所ではない語 98.4%となった．それぞれの成功例，失敗例を**表２**に示す．

表２ 成功例･失敗例

場所語	**成功例**	**学校，プール，駐車場**
	失敗例	**剣道場，コンビニ，ゲーセン**
場所でない語	**成功例**	**村長，パソコン，山登り**
	失敗例	**水道，携帯電話，磁場**

表２の失敗例より，複合語・カタカナ語・略語（以下，まとめて「拡張語」と呼ぶ）に対する精度が低いと考えられる．そこで，評価データの場所語の中から拡張語のみ（140 語）を抽出して再び評価をとった．その結果を**図５**に示す．成功例・失敗例を**表３**に示す．

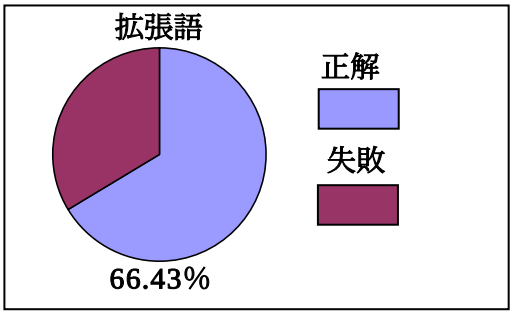

図５ 「拡張語」の精度

表３ 拡張語の成功例･失敗例

拡張語	**成功例**	**学生食堂，原発，スタジアム**
	失敗例	**剣道場，コンビニ，立体駐車場**

拡張語の精度は 66.43%であり，**図４**の精度と比べ低くなっている．拡張語に対しての精度を向上させることで，システムをより良いものにすることができる．

４．場所判断システムの拡張

拡張語に対しての精度向上として,「未知語処理の改良」と「場所語学習支援ツール」の二通りの手法を行う．

４．１．未知語処理の改良

場所語 KB に存在しない語（以下,「未知語」と呼ぶ）は，２．２節で示したように CB と関連度計算を用いて，その語に最も関連の強い代替語に置換し場所判断を行っている．この処理を「未知語処理」と呼んでいる．

拡張語は場所語 KB に存在しないため，未知語処理を行っている．この未知語処理を改良することで，拡張語に対する精度向上を図る．

拡張語に対する未知語処理（以下「未知語処理２」と呼ぶ）は拡張語を判断可能な語に置き換えるという処理を行っている．従来のシステムで「場所ではない」と判断された場合，未知語処理２で語を置き換える．そして，置き換えた語で場所判断を行う．

従来のシステムでは「屋」「前」がつく語は不正解であることが多かった．そのため，これら二つの語については特別処理を行う．「屋」に対しては「屋」を「店」に置き換えるという処理を行う．「前」に対しては「前」の直前に付いている語が場所かどうかを判断するという処理を行う．それ以外の拡張語に対しては，語を形態素解析し前部と後部に分け，後部で場所であるかどうかを判断し，前部から主体語・目的語を連想するという処理を行う．例えば，ボウリング場の場合は前部「ボウリング」と後部「場」に分けられる．そして，後部「場」より場所であると判断され，前部「ボウリング」から「ボウラー・ピン・倒す・競う etc」を連想する．

未知語処理を改良したシステムの精度を測った結果,「場所語」88.0%,「場所でない語」97.7%,「拡張語」79.3%となった．拡張前と拡張語の精度を**図６**に示す．また，成功例・失敗例を**表４**に示す．

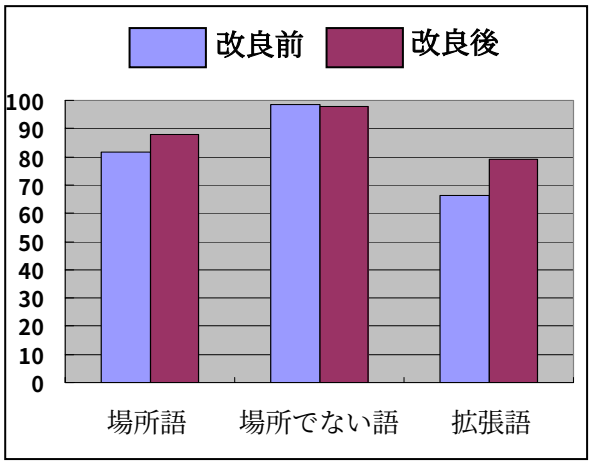

図６ 未知語処理の改良による精度変化

表４ 成功例･失敗例

正解例	失敗例
原子力発電所	コンサートホール
靴屋	学食
スポーツジム	ゲーセン

従来のシステムと比較して，精度が「場所語」6.5%向上,「場所ではない語」0.7%低下,「拡張語」12.9%向上している．「場所語でない語」の精度が低下した原因として，未知語処理２によって分けられた語が場所に関係する語と高関連を取りやすくなったからと考えられる．例えば,「時刻表」の場合「時刻」「表」に形態素解析する．そして，後部「表」で場所判断を行う際に,「道路」と高い関連を取ってしまい,「場所である」という間違った判断をする．「場所でない語」の精度は低下しているが，その他における精度は向上しており，システム全体の精度は上がっている．よって，今回の未知語処理の改良によって拡張語に対しての精度向上は行えたといえる．

表４より，拡張語中の複合語に対しては，正しい判断が多くできた．しかし，カタカナ語・略語に対しての精度は上がってない．そこで，カタカナ語・略語に対しての精度向上手法として「場所語学習支援ツール」を提案する．

４．２．場所語学習支援ツール

場所語学習支援ツール（以下,「支援ツール」と呼ぶ）とは，少ない知識を入力することでその語を拡張し，多くの知識をコンピュータに学習させるというシステムである．例えば,「バス停」を学習させたい場合，学習させたい語「バス停」とバス停の主体語・目的語「バス・乗る」をユーザが入力すると，「バス停」と関係のある場所語「停留所」を取得し，「停留所」の主体語「時刻表・ベンチ・旅客」と，目的語「降りる・乗車・下車」をバス停の主体語・目的語としてコンピュータに学習させる．

4.2.1 場所連想システム

上記の例「バス停」から「停留所」を取得する際に，そこに存在するものや目的から場所を連想する「場所連想システム」[5]を用いている．場所連想システムとは，そこに存在するものや目的から場所を連想するシステムである．例えば「おみくじ・引く」から「神社」を連想する．

4.2.2 場所語学習支援ツールの流れ

支援ツールの流れを以下に示す．

1）学習させたい語（以下,「登録語」）とその主体語・目的語を入力する．
2）登録語と関連のある場所語（以下,「場所類似語」）を取得する．場所類似語が複数出力された際には最適な語を選択する．
3）場所類似語から主体語・目的語を取得する．不適切な語が得られた場合や，必要な語が出力されなかった際にはユーザが編集可能である．
4）知識が正しいと判断できれば場所語 KB に登録する．

作成した支援ツールの有用性を調べるために，評価を行う．評価データは，３章で用いた「拡張語」の評価データの中から不正解であった 47 語を用いる．評価方法としては,「正しい場所類似語を得ることができたかどうか」を正答条件として行う．場所類似語が複数出力された際は，その複数の中に正しいと判断できる場所類似語が１つでも含まれていたら，正解とする．つまり,「テーマパーク」を学習さる時,「テーマパーク，乗り物，遊ぶ」という入力から「カジノ・歓楽街・公園・遊園地」という複数の場所類似語が得られる．「テーマパーク」の場所類似語は「遊園地」であると考えられるので，この場合は正確に出力が行われていると判断できる．このような条件の下，支援ツールの評価を行った結果，支援ツールの精度は 91.5%となった（**図７**）．

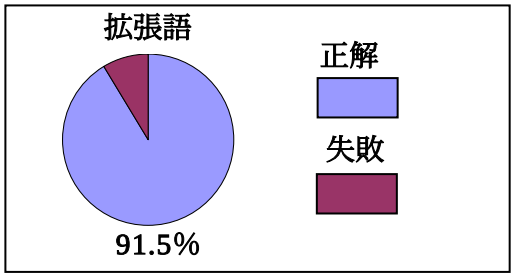

図7　場所語学習支援ツールの評価

支援ツールを用いて，91.5%という高い精度を得ることができた．よって，支援ツールとしての有用性はあると考えられる．成功例として,「カフェ」と「コーヒー・飲む」から「喫茶店」を得ることができた．失敗例として，入力「コンサートホール，歌手，聞く」から場所類似語として「消音室」が取得された．「消音室」は正しい場所類似語とは考えられない．しかし,「消音室」の主体語「音楽家・歌手」，目的語「演奏・歌う」はコンサートホールの主体語・目的語として正しいといえる．よって，場所類似語が不正解でも語を拡張できると考えられる．

この支援ツールを用いることで，未知語処理では困難であったカタカナ語・略語に関する知識をコンピュータに学習させ，システムの精度を上げることができると考えられる．

５．おわりに

本稿では，コンピュータに場所に関する常識をもたせる場所判断システムの精度向上を行った．

従来のシステムでは判断が困難であった複合語・カタカナ語･略語に対しての精度向上を目的とした．その手法として，複合語に対しての「未知語処理２」を用いた内部的アプローチと，カタカナ語・略語に対しての「場所語学習支援ツール」を用いた外部的アプローチの二通りを考案した．これらの手法を用いることによって，拡張語に対しての判断の正確さが増し，システム全体の精度向上を行うことができた．以上より，場所判断システムをより知的で柔軟なものにすることができたといえる．

今後の課題として，複合語だけではなくカタカナ語・略語における未知語処理の考案と，場所語学習支援ツールにおける場所類似語取得法や，主体語・目的語取得法を改良することを挙げることができる．それらを改良していくことで，ユーザにとってより使いやすいシステムにする必要があると思われる．

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「人間と生物の賢さの解明とその応用」における研究の一環として行った．

参考文献

[1] 杉本 二郎，渡部 広一，河岡 司,“常識判断システムにおける場所理解に関する研究”,情報処理学会自然言語処理研究会資料，2003-NL-153，pp.81-88，2003.
[2] NTT コミュニケーション科学研究所監修,「日本語語彙体系」，岩波書店，東京，1997.
[3] 小島 一秀，渡部 広一，河岡 司,“連想システムのための概念ベース構成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol. 8，No. 5，pp. 93-110，2002.
[4] 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.
[5] 荒井 亮太,“概念ベースを用いた場所連想システムの構築”，卒業論文，2003.