コンピュータと教育 4－3
（1989. 2. 25）

用例検索による日本語教師支援システム

堤　豊
日本IBM(株)東京基礎研究所

隅田　英一郎
ATR　自動翻訳電話研究所

日本語教師支援のための類似検索による日本語文の用例検索システムについて報告する。本システムでは、検索要求として、日本語の文を許し、一般化規則により構文を抽出し、データベース中の用例文とマッチングを行う。この方式では、ユーザーは検索キーについてほとんど考慮せずに検索ができる。またデータベースの構築は、機械的に行うことができ、特別な知識を必要としない。一般化規則は、ユーザーに開放されており、これを変更することで、類似の判断基準を変えられる。本稿では、この一般化規則を変更した例についても言及し、本システムがいろいろな検索要求に答えることができることを示す。

An Assistant System for Teachers of Japanese by a text retrieval

Yutaka Tsutsumi

Tokyo Research Laboratory, IBM Japan Ltd.
5-19,Sanban-cho,Chiyoda-ku,102

Ei-ichirou Sumita

ATR Interpreting Telephony Research Laboratories
Twin 21 Bldg.,2-1-61,Shiromi,Higashi-ku,Osaka,540

Given an idiom, this system analogically retrieves example sentences of it from Japanese sentence database. The method for retrieving consists of two parts: extracting a syntax information from a Japanese sentence by applying generalization rules, and searching the database for sentences matching the syntax.

This function assists teachers of Japanese language in making materials for instruction. Teachers need not to identify the exact syntax of the sentence but can input it as it is.

The database can be constructed easily without special knowledge, and the generalization rules are open to the user's change. The change of the generalization rules, which meets various requirements, is also exemplified.

従来の用例検索システムは、２節で述べたように、単語単位で指定するもの、あるいは予め決められた検索キーを指定して検索するものであった。前者の方法だけでは、期待する用例を出力することが困難であり、また、後者の方法では簡単に使用するというわけにはいかず、また、データベースの作成が非常に困難であることは前述の通りである。本システムでは、類似検索を導入しているため、検索の指定は簡単であるが、次に示すような複数の検索レベルを設けることで複雑な検索要求にも答えられるようになっている。検索の要求として許される形式は次の通りである。

（１）単語
（２）文節
（３）文

ここで、単語を指定するとシステムの動作としては、ＫＷＩＣと同じになる。しかし、本システムでは、データベース構築時に、検索の速度を上げるための工夫をしている［１］ため、検索に要する時間は非常に短い。文節や文を指定した場合には、次のように検索が行なわれる。

（１）検索入力そのものを含むような用例がデータベース中に存在すればその用例をすべて出力（表示）する。
（２）もしなければ、一般化規則で決められた順序で一般化する。
（３）一般化された入力文と、データベース中の用例を同様に一般化してマッチングし、一致した用例をすべて出力する。もしなければ、（２）にもどり、さらに一般化を続ける。

これにより、最も"類似している"用例をデータベースから検索することができる。ここで問題となるのは「何をもって"類似している"というのか」ということである。われわれは、文と文が類似しているということを次のような３つと考えた。

（１）字面が似ている／音韻が似ている
　　昨日、歯医者に行った。
　　昨日は医者に行った。
（２）構文が似ている
　　彼は決して怒らない人だ。
　　彼女は決して泣かない。
（３）意味的に似ている
　　太郎はおなかが痛かった。
　　太郎は腹痛がした。

ＥＴＯＣでは現在は、このうち（２）を一般化規則により実現している。

検索入力には、ワイルドカードとして"～"を使用することもできる。具体的には、「～は～してしまった」等のように、名詞や動詞の代りに使用する。これにより、具体的な例文を思い付かなくても、構文だけで検索を行うことができる。

３．２　データベースの作成方法

いくら賢い検索が可能であったとしても、データベースを構築することが困難であれば、使用勝手は著しく低下してしまう。それは、ユーザーが自分用にデータベースを作ることができないからである。教師が教材を作ることなどを考えると、やはり最新の話題を盛りこんだり、あるいは教室の話題に沿ったものを用例として確保しておきたいであろう。このためには、誰にでも簡単にデータベースを構築することができなければならない。

ＥＴＯＣでは、用例文をファイルに入れて、プログラムを呼ぶだけでデータベースの構築が可能である。システムの中では、次の３．３で述べる形態素解析作業と同一の作業を行ない、さらにパターンマッチングの速度を上げるために必要なデータ構造に変換するという作業を機械的に行っている。この間、人手がかかるのは、用例文をファイルに入れる作業だけである。

自然言語処理技術はかなり完成されてきつつあるが、まだ完全ではない。日本語の形態素解析も同様で、９５％くらいの文についてはうまく処理することができるが、間違える可能性も５％くらいはある。しかし、ＥＴＯＣでは、データベース構築時と検索時に同じ形態素解析プログラムを使用しているため、たとえ品詞名を間違えたとしても、かなりの場合、検索は成功する。

３．３　システムの実現

図１にＥＴＯＣの概観を示す。図に示すように本システムは大きく分けて、入力文解析部、パターンマッチング部、一般化部の３つから構成される。入力文解析部では、入力された文を構文解析し、単語単位に分割し品詞名を付ける。パターンマッチング部では、データベース中から要求を満たす文を検索する。もしデータベース中になければ、次の一般化部により入力文を予め与えられている知識に基づいて一般化し、パターンマッチング部に戻る。以下、システムの各部について述べる。

３．３．１　入力文解析部

入力文解析部では、与えられた入力文を形態素解析し、品詞名を認定する。この目的は、字面だけがマッチングしても、使われ方が異なるものを区別し、必要とされるものだけを出力するためである。例えば、

（１）酒を酌み交わそう［では］ないか。
（２）天気予報［では］雨だと言っていた。

１．　はじめに

コンピュータを教育に応用する場合に、２つの方法が考えられる。１つは、ＣＡＩのように教師の代りをしてくれるものであり、もう１つは直接学習者に対するものではなく、ＫＷＩＣ（KeyWord In Context）のように教師を支援するものである。ＣＡＩについては非常に多くの研究が報告されており、現在も盛んな分野である。一方、教師を支援するシステムについては、それほど活発な研究がなされていない。しかし、近年の日本語熱の高まりとともに、日本語教師の量的、質的な不足が問題になっている。このために、計算機を使用して日本語教師を支援することが重要になってきた。

著者らは、日英対訳用例文の知的検索システムＥＴＯＣ(Easy TO Consult)を研究開発している［１］。本来、このシステムはＭＡＨＴ(Machine-Aided Human Translation－翻訳支援システム）のために開発されたが、この枠組みはそのままで日本語教師支援システムとしても有効である。

本システムの特長は、類似検索方式をとっており、ある日本語文から、重要語を抽出し、構文的に似ている文をデータベースから検索するということである。また、一般化規則という知識ベース中に「何を以て似ているとみなすか」という情報が入っており、ユーザーは、これを自由に変更することで、「似ている」という判断基準をある程度自分で変えることが可能である。

２節では、用例検索の必要性と従来からの研究の概要と問題点を述べる。３節では、ＥＴＯＣの概要について、４節では、実際の検索結果について検討する。５節では、一般化規則の変更による検索ストラテジーのカストマイズについて述べる。

２．　計算機による用例検索の必要性と従来システム

言語学習者にとって、ある例文と似た例文を目にすることは学習の能率という観点から非常に望ましいことである。用例の収集を人手で行うという作業は、例えば、辞書や文法書を引いて該当する文を探すことであった。この方法では、大量の文が集められない、文が古くさい（最新の話題ではない）、引くのに時間がかかる、何をキーに引けばいいのか判らない場合がある、などの問題点があった。大量の文のデータベースから単語を検索してその前後を表示するものとしてＫＷＩＣがある。これは言語学研究者には非常によく使われているが、教育用の用例検索として使用するには、次のような問題点がある。

（１）検索キーとしては単語を１語だけしか使用できない。複数の単語をキーとして使用したい場合には、複数回の操作を行わねばならない。

（２）単語以外の情報を検索キーとして用いることができない。

（３）膨大な検索結果が出力される場合が多い。このため、本当に必要な文を見付けるためには、さらに他のキーを指定して検索するか、あるいは人手で抽出しなければならない。これは、検索キーとして単語１語しか指定できないことによる。

このようなＫＷＩＣの欠点を解消しようと試みた研究の１つに［２］がある。これは、各文ごとに検索キーを予め付けておき、それによって検索する。また、検索キーとして複数の項目を設定することや、品詞名で検索したり、あるいはシソーラスを使用して高度な検索を行えるようにしている。しかし、これは使用者としてあくまでも研究者に照準を当てており、教育用のツールとして考えると次のような問題点がある。

（１）品詞名やそのほかの検索キーとなるべき情報について使用者は熟知していなければ使用できない。

（２）複数の検索キーの指定方法が複雑である。

（３）ＰＲＯＬＯＧをベースとした推論システム上で実現されているため検索が遅い。

（４）検索するためのキーとなるべき情報をすべて人手で付与しなければならない。これは、膨大なデータベース（例えば１０万文）を作成する場合には相当な作業量となるであろう。また、これにより、個人毎にデータベースを持つということは不可能に近い。

以上のことから、われわれは、教育支援に使用する用例検索システムに対する要件として次の３項目を考えた。

（１）　検索方法が簡単であること。

（２）　データベースが簡単に作成および拡張できること

（３）　検索速度がある程度速いこと

３．用例検索システムＥＴＯＣの概要

本稿で紹介する用例検索システムＥＴＯＣの特長は大きく分けて３つある。検索方法が類似検索であり、誰にでも簡単に使用できること、データベースの構築が人手をほとんどかけずに簡単にできること、検索ストラテジーをある程度自由に変えてカストマイズできることである。本節では、ＥＴＯＣの概要として、検索の方法と、データベース構築の方法および、システムの実現方法について述べる。検索ストラテジーの変更については５節で述べる。

３．１　検索の方法

ここでは、ＥＴＯＣによる用例検索の方法を述べる。

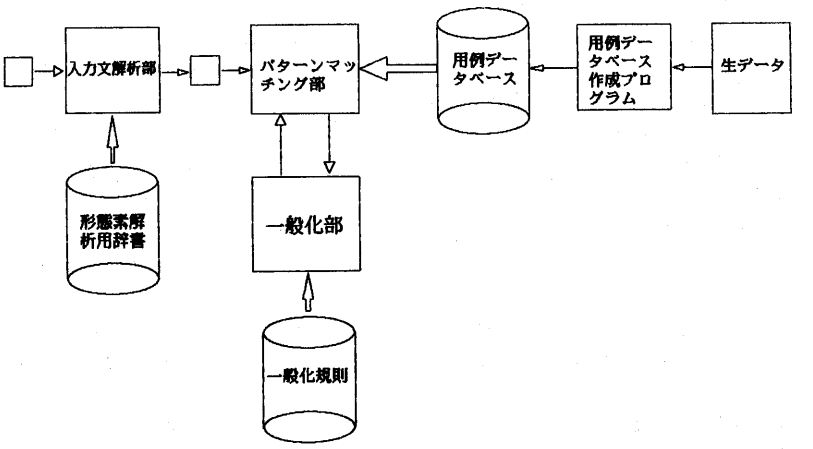

図1. 用例検索システムの概要

====> 働けば働くほど非難される。
6 ～ ば ～ ほど ～
見れ【ば】見る【ほど】不思議な絵だ。
The longer you look at this painting, the more amazing it becomes.
情報は多けれ【ば】多い【ほど】いい。
The more news we have the better.
解決しなけれ【ば】ならない問題が山【ほど】ある。
There is a mass of problems which have to be solved.
出発の時期は早けれ【ば】早い【ほど】良い。
We should leave as soon as possible.
多けれ【ば】多い【ほど】いい。
The more, the better.
彼はつきあえ【ば】つきあう【ほど】味の出る人だ。
The better you know him, the more interesting he seems.

(a)

====> 彼は笑ってしまった。
～ は ～ てしまった
彼女【は】手を滑らせてコップを割っ【てしまった】。
Her hand slipped and she broke the glass.
倒産して彼の会社【は】人手に渡っ【てしまった】。
His company went bankrupt and passed to other hands.
弁解するわけで【は】ありませんが急用ができて遅れ【てしまった】ので
I don't mean to make excuses, but I was late because an emergency arose.
あの夫婦【は】, 性格の不一致で別れ【てしまった】。
That couple split up because of a personality conflict.
私【は】戦争で家族と別れ別れになっ【てしまった】。
I became separated from my family during the war.
彼【は】悪い仲間の誘惑に負けて非行に走っ【てしまった】。
He was led into delinquency by some bad friends.
彼【は】ややこしい事件に巻き込まれ【てしまった】。
He ended up getting himself in a real mess.
空襲で町【は】ほとんど焼け【てしまった】。
The town was almost entirely destroyed by fire in the air raid.
彼【は】火事で家も財産も焼い【てしまった】。
His house and property went up in flames.
彼【は】昨夜泥酔して警察に留められ【てしまった】。
He got dead drunk and had to spend last night at the police station.

(b)

図2. 用例検索結果の例

代名詞 －> ～
名詞 －> ～
『の』の削除
『に』を『へ』に変える
動詞語尾の削除
動詞 －> ～
形容詞・形容動詞語尾の削除
形容詞・形容動詞－> ～
修飾語の削除
格助詞の削除
接続助詞の削除
助動詞の削除

図3. プリミティブ規則

上の２つの文中の［では］は、使われ方が異なっている。これを区別するために、［では］の品詞名を調べる必要がある。

データベース作成時に入力文解析部での形態素解析と同じ処理が全ての用例文に対して行なわれており、データベース中には各用例文とその文中の全ての単語とその品詞名が格納されている。

３．３．２　パターンマッチング部

パターンマッチング部では、入力文とデータベース中にある用例文とを比較し、一致したものを出力する。パターンマッチングの速度を上げるために早見表の作成などいくつかの工夫がなされている［１］。一致するものが見つからなかった場合には次の一般化部へと制御が渡され、要求が満たされるまで繰り返される。

３．３．３　一般化部

入力文とデータベース中の用例文で一致するものが見つからなければ、入力文を一般化規則により、一般化してパターンマッチング部に制御を戻す。一般化規則は、複数のプリミティブ規則から構成されており、１度の一般化要求に対して、１個のプリミティブ規則が適用される。もしも全てのプリミティブ規則を適用してもパターンマッチングが成功しなければ、該当用例文なしとして検索が終了する。

４．用例検索結果の例

図２に用例検索の結果を示す。図２（ａ）は、検索要求が「働けば働くほど非難される」という文であったが、これから、「～ば～ほど」がキーとして抽出され、期待する検索結果が得られた。この例に関して言えば、人手により辞書や文法書から検索するのは大変な作業である。「ば」で検索すればいいのか、あるいは「ほど」で検索すべきなのか悩むところであろう。また、検索キーを予め決めておく方式では、「～ば～ほど」を構文として最初に認定しておく必要がある。

図２（ｂ）は、「～してしまった」という文に対する検索結果である。これを例えば、品詞名のリストで検索するとすれば、かなり複雑になってしまう。

５．一般化規則による検索ストラテジーの変更

本システムの特長の１つとして、検索のストラテジーを簡単に変更できるということがある。これは、ストラテジーの部分が、パターンマッチング部から独立し、しかも一般化規則として、開放されていることによる。一般化規則は、複数のプリミティブ規則から構成されている。このプリミティブ規則は、「動詞の語幹の削除」「"に"を"へ"に変える」などの単純な規則であり、この順序のみによって一般化規則が作られる。プリミティブ規則として現在使用可能なものを図３に示す。

前節までで述べた検索結果は、すべてデフォルトとしてシステムが提供している一般化規則を使用している。この一般化規則は、構文的に類似しているものを検索するように設定されている。本節では、この一般化規則を変えることによって、違った検索をした結果を示し、この機能の有用性について述べる。

図４に、システム提供の一般化規則（ａ）と、今回の実験のために使った規則（ｂ）を示す。図に示されるように、（ａ）が名詞や動詞などの内容語をなるだけ早く一般化し、格助詞などの機能語を残そうとするのに対し、（ｂ）では、逆に内容語を残すようにしている。図５は、同じ検索要求に対して、（ａ）、（ｂ）それぞれでの検索結果を示す。このように、（ａ）は構文的に同じものを出力し、（ｂ）は、同じ単語を持つ文を出力する。

これを少し改良したものが（ｃ）である。これでは、ただ単にキーワードの組合せではなく、助詞を重視し「ＡがＢする」という検索を実現するための一般化規則となっている。（ａ）にかなり似ているが、修飾語や活用語尾を早い段階で一般化している点が異なっている。この検索ストラテジーは、ある名詞と動詞がどのような格関係を取りやすいかを調査するのに向いている。

このように、検索のストラテジーを簡単に変更できるため、日本語教育用のツールとしてはもとより、言語研究者のデータ収集ツールとしても有効であると考えられる。

６．まとめ

本稿では、類似検索による用例検索システムを日本語教育支援システムとして使用することについて述べた。本システムの最大の特長は、用例検索を文の類似によって検索することである。その類似検索を実現するために一般化規則を導入した。この一般化規則はユーザーが必要に応じて簡単に変更することができ、類似の判断基準をカストマイズすることができる。現在は、本システムは、ＩＢＭ３０８１システムの仮想計算機上にＬＩＳＰ言語で実現されたものと、ＰＳ／５５上にＣ言語で実現されたものの２種類が稼働中である。それぞれのシステムの外観を付録に示す。ホスト版のシステムは、本研究のプロトタイプ・システムであり、翻訳支援システムとして実現されている。ＰＳ版のシステムは、教師支援システムとして実現されており、漢字辞書の検索機能や、品詞を直接指定した検索機能などがインプリメントされている。また、ＰＳ版の方が、データベースの大きさ、検索速度ともに勝っている。ちなみに、ＰＳ版のシステムでは、デー

```
  ＥＴＯＣ　検索規則編集画面
====>
 －使用されている規則－
 代名詞　－＞　～
 名詞　－＞　～
 『の』の削除
 『に』　－＞　『へ』
 動詞語尾の削除
 動詞　－＞　～
 形容詞・形容動詞語尾の削除
 形容詞・形容動詞　－＞　～
 修飾語の削除
 格助詞の削除
 －－未使用の規則－－
 接続助詞の削除
 助動詞の削除
```

（ａ）

```
  ＥＴＯＣ　検索規則編集画面
====>
 －使用されている規則－
 『の』の削除
 接続助詞の削除
 格助詞の削除
 修飾語の削除
 形容詞・形容動詞語尾の削除
 動詞語尾の削除
 助動詞の削除
 形容詞・形容動詞　－＞　～
 代名詞　－＞　～
 動詞　－＞　～
 －－未使用の規則－－
 『に』　－＞　『へ』
 名詞　－＞　～
```

（ｂ）

```
  ＥＴＯＣ　検索規則編集画面
====>
 －使用されている規則－
 『の』の削除
 修飾語の削除
 形容詞・形容動詞語尾の削除
 動詞語尾の削除
 助動詞の削除
 接続助詞の削除
 格助詞の削除
 形容詞・形容動詞　－＞　～
 代名詞　－＞　～
 動詞　－＞　～
 －－未使用の規則－－
 『に』　－＞　『へ』
 名詞　－＞　～
```

（ｃ）

図４．　一般化規則の変更例

```
====>　食べるほど太る
 彼は大学生と思えない【ほど】幼稚な人間だ。
He's so immature that it's hard to believe he's a college student.
 彼は酔う【ほど】におしゃべりになる。
The more he drinks, the more talkative he gets.
 １００ドル【ほど】融通お願いできませんか。
Could you please lend me about 100 dollars?
 彼は山【ほど】仕事を抱えている。
He is really piled up with work.
 僕は君に質問が山【ほど】ある。
I have a bunch of questions to ask you.
 海岸では真夏の太陽がじりじりと照りつけて，肌が痛い【ほど】だった。
The midsummer sun was so sizzling hot at the beach that my skin hurt.
 猫の手も借りたい【ほど】忙しい。
I'm so busy it's enough to make my head spin.
 彼の病気はもう手の施しようがない【ほど】ひどくなっている。
His illness has gotten so bad that nothing can be done.
```

```
====> 私は学校に行く
6　　～　は　～　に　行　く
 今で【は】月【に行く】こともできる。
It's possible to go to the moon now.
 両親【は】彼に大学【に行く】ように勧めた。
His parents advised him to go to college.
 以前【は】大学【に行く】高校生は少なかった。
Few high school graduates used to go to college.
 父と私【は】よく魚釣り【に行く】。
My father and I often go fishing.
 彼【は】私がそこ【に行く】べきだと頑張った。
He insisted that I go there.
```

（ａ）

```
====> 食べるほど太る
2　　食べ　～　太　～
 彼女はよく【食べ】るのに，少しも【太】らない。
Even though she eats a lot, she doesn't gain any weight at all.
 彼女はそんなに【食べ】ない。それでもあんなに【太】っている。
She doesn't eat so much. And yet look how fat she is.
```

```
====>　私は学校に行く
　　私　～　学校　～　行　～
 【私】の【学校】は駅から歩いて【行】けます。
We can walk to school from the station.
 【私は】毎日歩いて【学校に】【行】きます。
I walk to school every day.
```

（ｂ）

```
====> 食べるほど太る
2　　食べ　～　太　～
 彼女はよく【食べ】るのに，少しも【太】らない。
Even though she eats a lot, she doesn't gain any weight at all.
 彼女はそんなに【食べ】ない。それでもあんなに【太】っている。
She doesn't eat so much. And yet look how fat she is.
```

```
====>　私は学校に行く
　　私は　～　学校に　～　行　～
 【私は】毎日歩いて【学校に】【行】きます。
I walk to school every day.
```

（ｃ）

図５．　一般化規則を変更した場合の検索例

タベースとして、約５万例が格納されており、検索速度は平均約５秒と十分実用に耐えうるシステムとなっている。また、データベースには、特別な工夫をしなくても、最大約１０万文が格納可能である。

今後の課題としては、実際の使用に際して、どういう一般化規則が必要になるかを調査し、一般化規則を構成しているプリミティブ規則が今のままでいいかどうかについて検討する必要がある。また、意味的な類似については現在全く考慮していないが、名詞および動詞についてシソーラスを品詞と共に一般化規則で使用することにより、ある程度意味的な類似を実現できると思われる。ただし、この場合一般化規則の適用が現状の数倍行なわれることを考えると検索速度がかなり遅くなることが予想される。さらに、このとき一般化規則をどう構成するかなどについても興味深い研究課題である。

実用面からの研究課題としては、形態素解析部で話し言葉を扱えるようにすることが当面の最大の課題である。従来、計算機上では、書き言葉については非常によく研究されているが、話し言葉についての研究はあまりやられていない［３］。音声認識やワード・プロセッサなどでも話し言葉の解析は今後非常に重要となるので、自然言語処理の観点からも積極的に取り組むべき課題である。

現在は、検索入力文は単純であることが要求されるため、例えば、実際の日本語文（小説や新聞記事など）で複雑な文をそのまま入力しても期待された通りのものが検索されるとは限らない。複文や重文に対する処理については、形態素解析のレベルでは難しいため、将来は係受けの解析も導入したいと考えている。

謝辞

本研究遂行にあたり、形態素解析部を開発していただいた丸山宏氏、大深悦子氏、諸橋正幸氏、また、いろいろと活発な議論をしていただいた、国立教育研究所の及川昭文氏ほかの諸氏に深謝いたします。

参考文献

［１］ Sumita E. and Tsutsumi Y., "A translation aid system using flexible text retrieval based on syntax-matching", Proc. of the 2nd international conference on theoretical and methodological issues in machine translation of natural languages, CMU, Pittsburgh, 1988.

［２］ 三吉ほか，"構造化キーワードを用いた用例検索システムの試作",自然言語処理研究会資料NL70-8,1989.

［３］ 大曽、小山，"話しことばとデータベース",文部省科研費・試験研究「パソコンによる外国人のための日本語教育支援システムの開発」研究発表会予稿集，1988.

付録

ホスト版のシステムとＰＳ版のシステムを示す。ホスト版は、翻訳支援システムとして実現されているため、画面が３つに分かれている。検索用の画面（上）、英文作成用画面（左下）および日本語参照画面（右下）である。一方、ＰＳ版は教師支援システムであり、本稿で述べた類似検索は左のウインドウから行ない、右のウインドウは、品詞名による検索を行ったり、辞書を検索するために使用されている。