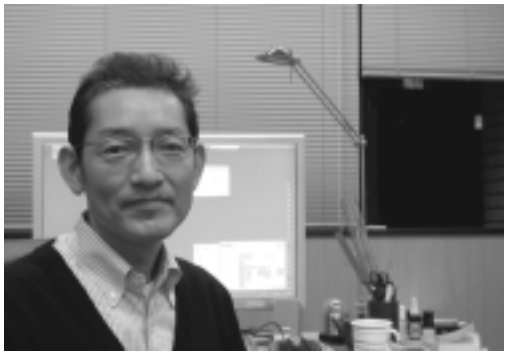
In Laboratory Now

研究室訪問 1

よりよい情報社会へ

徳永 健伸 研究室～計算工学専攻



徳永 健伸 教授

私たちが普段用いる日本語や英語などの言語。これらの言語は人間の営みの中で自然に形成されてきた言語で、自然言語と呼ばれている。この自然言語をコンピュータに扱わせる研究、すなわち自然言語処理に関する研究は世界初のコンピュータが作られた頃から行われてきており、現在も様々な方面の研究が国内外を問わず至るところで盛んに行われている。

今回取材に伺った徳永研究室もこの研究の一端を担っている。ここではその研究のうちテレビ番組の自動要約を目指す研究について紹介しよう。

便利な情報アクセスを

現代の社会では日々膨大な量の情報が生み出されている。新聞、テレビ、雑誌等のメディアは毎日のように新たな情報を提供しており、また個人レベルでもブログなどを用いて情報発信を行う人が増加してきている。

このように社会に氾濫する情報全てに我々が目を通すことは到底不可能である。そこで今の社会で重要になってくるのは、いかに必要な情報の重要な部分だけに速やかにアクセスするかということだ。今ではコンピュータの検索機能が発達してきており、ある程度速やかに必要な情報を探すことが可能になっている。しかし、検索機能で辿り着けるのは、例えばあるキーワードを含むWebページであったり、文書中の自分が探している単語が現れる箇所程度である。これでは不必要な情報にも目を通さなければならず、また本当に自分の知りたい情報が含まれているのかどうかの判断にも手間がかかってしまう。

この問題を解決する一つのアプローチが自動要約である。自動要約とはコンピュータに様々な類の要約を自動で行わせる手法である。自動要約は、一つの文書から要約を生成する単一文書要約をはじめ、複数の文書から一つの要約文を生成する複数文書要約、さらに最近では画像・音声・ビデオなどテキスト以外の入力から要約を生成するマルチメディア要約など、非常に幅広い分野を対象に研究がなされている。

また、別な方法として情報抽出というものもある。自動要約は主に文章から文章を生成するが、情報抽出は主に文書内に含まれる特定の単語や数値を抜きだし、表などを用いて簡潔にまとめるものである。例えば、ある企業の人事異動に関する新聞記事から、どの企業で、誰が、いつ、どこからどこへ、どういう理由で異動するのかというような情報を抽出したり、台風情報の記事から、その台風の位置や速度、進行方向、中心気圧、最大風速というような情報を抽出することなどが考えられている。

このような自動要約や情報抽出を実現するためにはコンピュータに自然言語を扱わせる必要があり、現在徳永研究室で行われている研究の一つもこれらに類するものである。次の章では先生が行っている、テレビ番組の要約を生成する研究について紹介していこう。

クローズドキャプションからのQA抽出

徳永研究室ではクローズドキャプションを用いてテレビ番組の要約を生成する試みを行っている。クローズドキャプションというのは聴覚障がい者も健常者と同等の情報をテレビから得られるようにするためにつけられる字幕のことである。この字幕は我々が普段ニュースのテロップ等で目にする字幕（オープンキャプション）とは違い、番組に出演している人物の発言やナレーション等の音声全てを字幕にしていることが特徴的である。そのためクローズドキャプションはテレビ番組をテキスト化したものといえるのだ。そこで、このテキストデータを利用することで要約を行えるのではないかと先生は考えたのだ。

徳永研究室ではまず手始めに、ＮＨＫの「地球！ふしぎ大自然」という番組のクローズドキャプションを用いてQ&A形式で番組をまとめようとしている。この番組は大自然の映像を流しながら、ナレーターが動物の生態を紹介していく形態である。進行の際にはナレーターが「この動物は何をしているのでしょうか？」のような問いかけをし、「実は～をしていたのです」のようにそれに答えていく形をとることが多い。このことに着目し、質問の部分とそれに対応する解答の部分を抜き出す作業、即ちQA抽出を行い、それをまとめ上げれば動物のデータベースを作ることができるのではないかと先生は考えたのだ。

このQA抽出をコンピュータに行わせるには、コンピュータに言語を扱わせなければならない。しかし、コンピュータは人間のようには言語の意味を理解することはできない。そこで、現段階では文の表層的な特徴をうまく利用し、このQA抽出を実現しようとしている。例えば、疑問詞や体言止めなどの特定の単語や表現に注目すれば、その文が文章中でどのような役割を果たしているのかを推測できると考えられている。また、文の位置関係や文同士の類似度のような文と文の関係もQA抽出の手がかりになると考えられている。

文同士の類似度の計算には語彙的連鎖というものを利用している。**図1**は第9行の文の「光る」という動詞に注目した様子である。このように、同一テキスト内に現れる注目している語と同じ語、あるいは類義語との結びつきを語彙的連鎖という。そして、各文に含まれる語彙的連鎖を形成する語の数をその文のチェーン数と呼んでいる。チェーン数は一文ごとの値や、連続する複数の文のチェーン数の和などが利用され、文同士の類似度の計算に用いられている。

このようなことに着目していくことでコンピュータに言語を扱わせることができると考えられているが、実際にこれらの情報が本当に役に立つのかを判断したり、これらを組み合わせてどのような式を用いて計算すればよいのかを人間が考えるのは困難である。そこで徳永研究室ではCRF（Conditional Random Fields）という機械学習の一手法を用いてその問題を解決しようとしている。ここでのCRFの利用法は、注目するとうまくいきそうな情報（素性）を指定し、人手で様々なタグ付け（アノテーション:**図2**）されたデータを

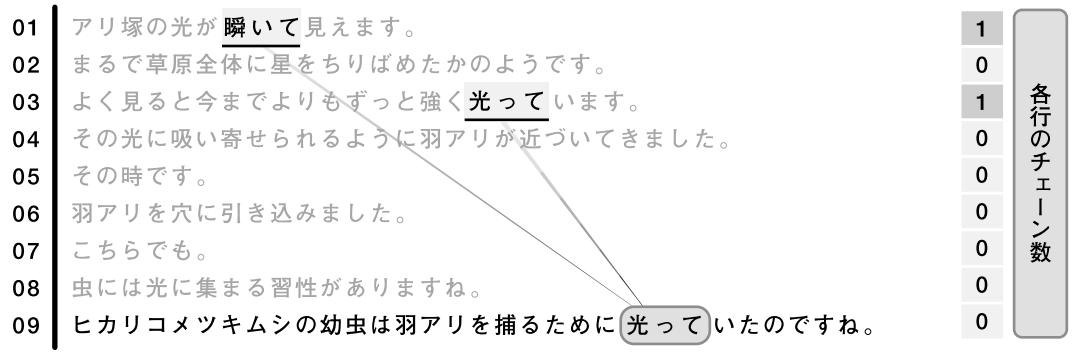


図1　語彙的連鎖

用いてコンピュータに学習させ、どの素性が有効かそうでないかを選別させ、計算式を作り出させるということだ。これによって素性の複雑な組み合わせを用いた計算が可能になる。ここでいう素性とは上述したような文の様々な特徴であり、アノテーションが行われるのは質問や解答の中心となる部分、その背景となる部分、中心の文法的省略要素を補う部分などである。この機械学習に20本ある番組データのうち19本を利用し、最後に学習に使わなかった１本に対して実際にコンピュータにQA抽出を行わせ、どの程度の抽出の精度がでるかを調べる。それが終わったら学習に用いないデータを違うデータにし、また19本新たに学習させ、学習に用いなかったデータで抽出を行わせる。この繰り返しによって研究を進め、どのような素性が有効なのかを模索しているのだ。

それでは、具体的にどのような手順でQA抽出が行われているのか順を追って見ていこう。

１．質問の中心となる一文を見つける

クローズドキャプションを冒頭から調べていき、質問の中心となる一文を見つける。それは疑問詞と疑問符の両方を含む文と定めている。**図３**では「何を」という疑問詞に加え文末に疑問符の振られている①の文が質問の中心に該当する。

２．質問の導入部分を同定する

質問の中心となる一文を見つけたら、次にその質問文に注目して質問の導入となる部分を同定する。質問の中心の一文だけでは質問の状況が把握できないことが多いので、それを補うために質問の中心の文の前後からその質問がなされた状況を説明する部分を抜き出す必要があるのだ。これにはCRFを用いており、以下のような特徴を全文から抽出し素性としている。

・語彙的連鎖

語彙的連鎖を発生させる文の範囲は、質問の中心の文から質問の中心の主語を補う文までとする。この範囲に含まれる各文から語彙的連鎖を発生させ、クローズドキャプションの全ての文のチェーン数を算出する。また各文に対し、その文の二文先までのチェーン数の和、即ち三文のチェーン数の和を計算する。これらの値を用いて質問文と類似度の高い部分を求める。

・質問文に関する属性

今注目している質問文か、それ以外の質問文か、または質問文ではないかを区別させ、他の質問文との位置関係を学習させる。

・サブコーナーの句切り

サブコーナーというのは番組の途中で数回挿入される特集のことである。これは番組の本編の筋とは少し離れた内容であるため、場面が大きく切り換わる。場面の転換というのは導入部分の区切りの目安の一つとして考えられている。

・月の表現および体言止めの位置

「11月。」のような月の表現や体言止めは場面の転換の手がかりとなると考えられている。

・質問の中心の文からの行数

質問の中心からどの程度まで離れた部分が導入となるのかを学習させる。

今の研究段階ではこのような素性に注目して質問の導入部分を同定している。これらの素性を与え学習させることによって**図３**の②の部分が質問の導入部分として抜き出された。

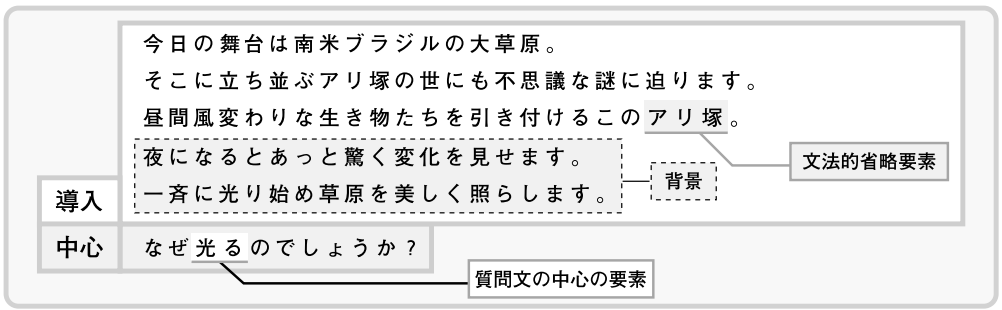


図2　アノテーションの例

３．解答の中心となる一文を見つける

解答の中心は主に質問文との文の類似度を計算することにより同定する。質問文の後の文に対し、語彙的連鎖のチェーン数、主語の一致性、特定の類似表現、指示語の出現に注目して計算を行う。その際、質問の中心の文の直後二文以内に場面の転換がない場合、その部分に解答の中心の文が現れる可能性が高いのでその点を考慮する。これらの結果最も類似度の高い一文を解答の中心とする。この条件に沿って計算を行うと**図３**の③の一文が解答の中心として抽出される。

４．解答の導入・補足部分を同定する

質問同様、解答も中心の一文だけでは不十分である。そのため解答の導入部や補足説明部を同定する必要がある。これにもCRFを用いる。その際に使われる素性は質問の導入の同定に使われたものに解答文の位置関係などを加えたものだ。それらの素性をもとにし、抽出を行うと**図３**の④の部分が解答の導入・補足部分として抽出される。

以上のような手順でQA抽出は行われている。現状では解答の中心の抽出の精度は7割程度、導入・補足部分の抽出の精度は7割から８割程度であるという。精度が低下してしまう原因としては文脈によって同じ表現でも違う性質をもってしまうことが考えられる。例えば、場面の転換を表す手がかりとして月の表現や体言止めがあるが、場合によってはそれらが場面の転換を示さないことがある。その違いはやはり文脈に依存してしまうのだ。そのため今後は言葉の意味にまで関わる方法などを検討し、より精度の高いQA抽出を目指している。

この研究が最終的に目指すものはテレビ番組の情報をQ&A形式に直して再構築し、文字と映像を組み合わせた新しいデータベースを作ることである。冒頭でも述べたように、これからの情報社会においては必要な情報に素早くアクセスすることが不可欠となるであろう。よりよい情報社会を実現するため、徳永研究室では日々研究がなされている。

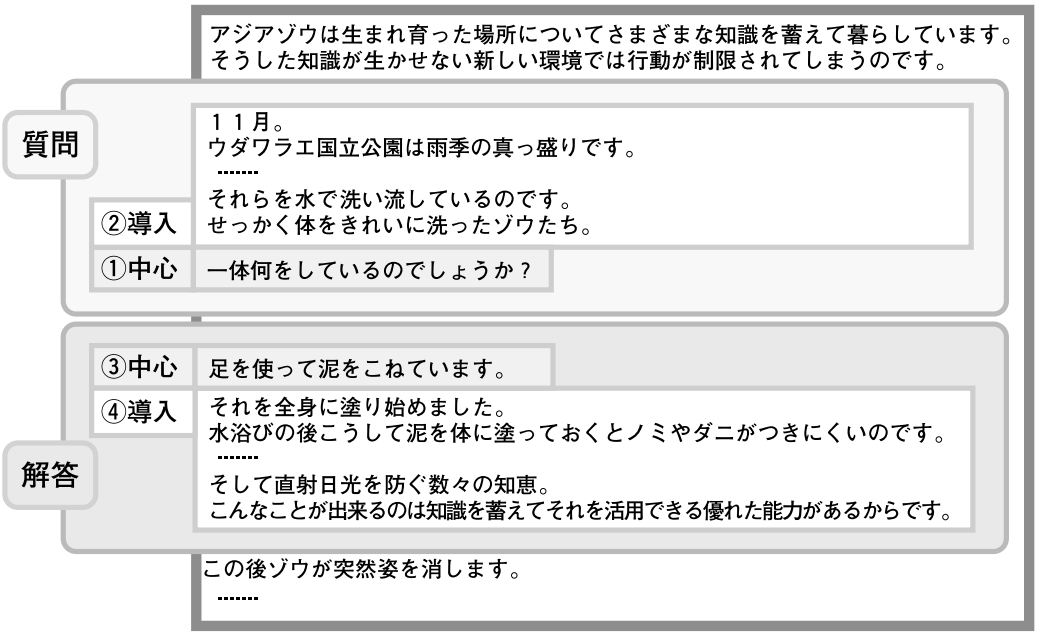


図３　QA抽出の例

今回の取材で伺った話はとても興味深く、勉強になりました。最後になりましたが、快く取材に応じてくださった徳永先生はじめ徳永研究室の皆様に心より御礼申し上げます。　（宇田川　拓郎）