Vol.2010-SLP-82 No.20
2010/7/24

Webからの情報抽出を用いた音声対話システム

吉野　幸一郎[†1]　河原　達也[†1]

本研究では、日々更新される Web テキストに対して、述語項構造に着目した情報抽出を行い、その内容を扱う対話システムについて述べる。本対話システムは、述語項構造により抽出された情報に基づいて、対話の履歴とトピックモデルを利用しながら、ユーザの質問に対応する情報を答える質問応答と、ユーザの要求・興味に沿った情報をプロアクティブに提示する情報推薦を行う。本稿ではこのようなシステムの実現可能性を示すため、プロ野球のニュース記事というドメインに限定してシステムを構築し、その評価を行った。また、こうした対話の枠組みを実現するために、対話のドメインごとに有用な述語項構造パターンをコーパスから教師なしで自動抽出する手法を提案し、その評価を行った。

Spoken Dialogue System based on Information Extraction from Web Text

Koichiro Yoshino[†1] and Tatsuya Kawahara[†1]

In this paper, we present a novel scheme of spoken dialogue systems which uses the up-to-date information on the web. The scheme is based on information extraction which is defined by the predicate-argument (P-A) structure and realized by shallow parsing. Based on the information structure, the dialogue system can perform question answering and also proactive information presentation using the dialogue context and a topic model. Feasibility of this scheme is demonstrated with experiments using a domain of baseball news. In order to automatically select useful domain-dependent P-A structures, information-theoretic criteria are introduced, resulting to a completely unsupervised learning of the information structure given a corpus.

†1 京都大学 情報学研究科
School of Informatics, Kyoto University

1. はじめに

近年、Web 上に集積する情報は爆発的に増加しており、Web 上の情報にアクセスし、活用する機会が増大している。こうした情報の検索は、現在はキーワード型検索が主であるが、ユーザからの情報要求すべてがキーワード型に適しているわけではなく、漠然とした要求も多い。そこで、ユーザの意図・嗜好を対話的に顕在化しながら情報を提示するシステム（情報コンシェルジェ）の研究[1] が行われている。

これまで研究・実用化されてきた音声対話システムはおおむね２種類に分類される。[4] フライト情報案内[5),6] やバスの運行案内[7] などの明確なタスクを定義し、関係データベース (RDB) をバックエンドとした枠組みは、タスク達成に必要な意味表現の定義や対話のフローの記述が容易であった。その反面、Web などの大規模なテキスト情報に対して適用することが困難であった。それに対して、一般的な文書検索を用いた対話システムの研究[8),9] も行われてきたが、表層的なキーワードや係り受け関係、質問タイプなどのみに着目し、深い言語的解析や対話処理は扱われていない。その結果、対話の文脈やユーザの要求とは無関係な、不自然な応答が生成されることがあった。また、情報コンシェルジェにおける重要な機能として、ユーザに対するプロアクティブな情報提示があるが、先行のシステム[10] においては、文書の中から特徴的な文章を提示するにとどまり、文脈やユーザの意図に沿った応答を必ずしも生成しているわけではなかった。

これに対して本研究では、述語項構造に着目した情報抽出を行うことで、RDB のような構造を持たない Web 文書を扱いながら、その意味表現を扱うシステムを構築する。対話を行う上で有用な情報構造はドメインに依存しており、そのような情報構造のテンプレートを作成する必要性が指摘されている[2)]。しかし、ドメインごとに人手でテンプレートを作成する方法論では、Web に存在する様々なドメインに対して適用できない。そこで、本研究ではパーザを用いた述語項構造解析の結果から、自動で特定のドメインにおける重要な情報構造を抽出する方法を検討する。

2. システムの概要

本システムは、Web に存在するニュースサイトや Wikipedia などのテキスト情報を用いて、ユーザの質問に答えながら対話を行う。今回は、扱うドメインをプロ野球に限定し、毎

Vol.2010-SLP-82 No.20
2010/7/24

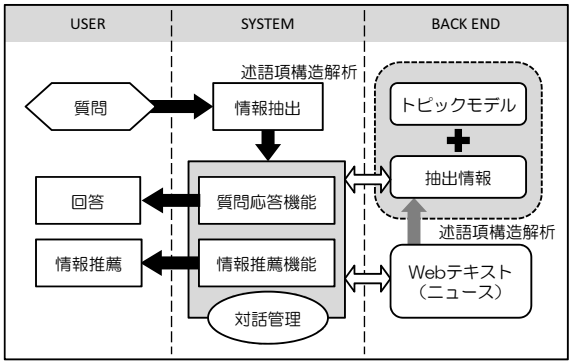

図 1　対話システムの概要

日新聞社のニュースサイト*1を利用している。また、今後他のドメインに容易に拡張できるようにシステムの設計を行っている。

2.1　対話システムの構成

本システムの構成を図 1 に示す。まず、システムは事前に Web から得たテキストに対して、述語項構造解析などを用いた情報抽出を行う。対話中に、ユーザの発話に対しても同様の枠組みで情報抽出を行い、抽出した情報どうしのマッチングを取ることで、ユーザの要求に最も関連の深いテキストを検索し、応答を生成する。

2.2　対 話 戦 略

本システムの対話戦略を図 2 に示す。ユーザの発話と抽出した情報の間でマッチングを行い、完全に一致する情報があれば、その情報を利用した応答を生成する。完全に一致する情報がなければ、ユーザの発話に対して最も近い内容を検索し、その内容を利用した応答を生成する。図 3 にこのような応答の生成例を示す。ユーザが「今日は金本は打ちましたか？」というような発話を行った場合、これを本システムでは、「金本 - が　打つ」という述語項構造の形で情報抽出する。この場合、システムは持っているテキストから抽出した情報として、ユーザが意図した金本の打撃成績の情報を持っておらず、従来の検索型のシステムでは、「情報がありません」と答えるしかなかった。しかし、例にある「＊ - が　打つ」のように、述語項構造の格と用言のペアなど、部分的に一致する情報を検索することで、より自然で頑健な形でユーザとの対話を継続できる。

*1 http://mainichi.jp

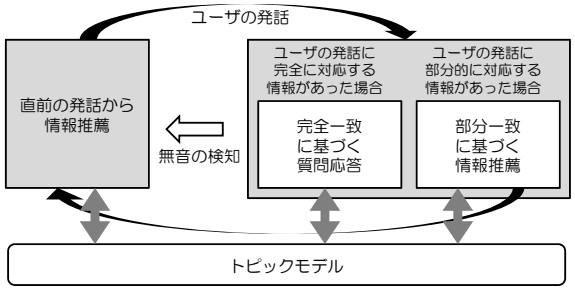

図 2　対話戦略

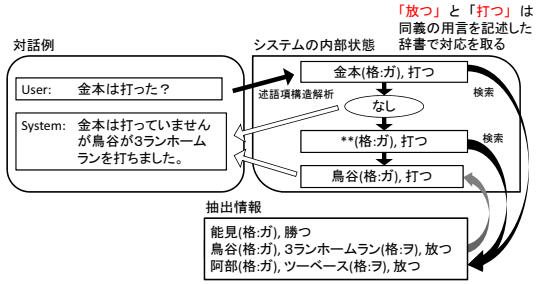

図 3　対話の生成例

対話中にユーザからの音声入力が一定時間以上なくなった場合、システムは直前のユーザとの対話履歴を利用したトピックモデルによってユーザが興味のある情報を推測し、近い内容を持ったテキストからプロアクティブに対話を生成する。

これらの枠組みによって、従来システムよりも雑談に近い対話を行いながら、ユーザの興味に沿った情報提供を行うことができるシステムを構築できることが期待される。

3.　述語項構造解析に基づく情報抽出

3.1　述語項構造解析

テキストから情報抽出を行い、その意味表現を捉えた応答を生成するためには、意味表現の単位と、それを表すことができる抽出形式を定義しなければならない。そこで本研究で

ニュース記事

延長十回、2死三塁から鳥谷が右中間へ決勝適時二塁打を放った。
先発・能見は9回を2安打無失点の好投。

抽出情報

([延長十回 <修飾> 放つ][2死三塁 <から> 放つ]
[鳥谷 <が> 放つ][決勝適時二塁打 <を> 放つ])
([先発・能見 <が> 好投][9回 <を> 好投][2安打無失点 <の> 好投])

図 4　述語項構造解析

は、意味表現の単位として述語項構造を用いる。述語項構造は、「要素 - 格　用言」の関係性に基づいて意味表現を表すものである。述語項構造は、古くから自然言語処理の分野で利用されてきた形式であり、古典的な対話システムにおいても、必要な述語項構造を人手で定義したものが利用されていた。近年統計的手法による大規模で一般的な述語項構造解析の研究[11] が進んでおり、いくつかの自然言語処理タスクで利用されている[12)–14)]。本研究では、述語項構造解析のパーザとして JUMAN/KNP[⋆1] を用いる。

述語項構造解析に基づく情報抽出の例を図 4 に示す。要素とその係り先である用言、その関係性として格を抜き出す。この際、一つの用言に対して複数の要素と格が存在するが、その組み合わせもあわせて保持する。日本語の場合は自明な場合のガ格が省略される傾向にあるが、そうしたゼロ代名詞推定の問題は今回は扱わず、対話生成の際には後に述べるトピックモデルを用いて対応する。また、図 3 の例にあるような「放つ」と「打つ」という用言のペアは、野球という文脈においては同義の用言として扱う必要があるが、これは別途検討する。

大規模テキストから述語項構造のパターンを収集すると、非常に多くのパターンが抽出されるが、情報検索・推薦の対話で有用なものはドメインに依存して限定される。例えば、野球ドメインでは「A 選手 - が　打つ」や、「B チーム - が　連勝」など、経済ドメインでは「A 社株 - が　ストップ高」や、「B 社と C 社 - が　提携する」などの表現が典型例となる。パーザが出力する大量のパターンをすべて利用することは効率が悪いだけでなく、不要なパターンの増加は音声認識誤りや解析誤りなどに対する頑健性の点でも好ましくない。そこで、自動で抽出された述語項構造の中から、当該ドメイン内で対話を行う上で重要なパターンを統計的な尺度によって選別する。

⋆1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/

4. 重要な述語項構造のパターン抽出

コーパスから当該ドメインで特徴的な述語項構造のパターンを教師なしで抽出するために、以下の統計的尺度を考える。

4.1 TF-IDF 法

まず、ドメイン中の単語の重要度を測る手法として一般的な TF-IDF 法を利用する。TF-IDF 法は、単語を w_i 、ドメインもしくはトピックを t としたとき、以下の式で定義される。

$$tfidf(w_i, t) = P(w_i|t) \log \frac{C(d)}{C(d : w_i \in d)} \tag{1}$$

この式の中で、TF 項 $P(w_i|t)$ は、単語 w_i の出現確率

$$P(w_i|t) \approx \frac{C(w_i, t) + \alpha}{\sum_j \big(C(w_j, t) + \alpha\big)} \tag{2}$$

として求められる。$C(w, t)$ は、ある単語 w_i がドメイン t において出現する回数であり、α はディリクレ過程を用いて与えた事前分布である。α の値は、TF 項から導かれる尤度関数に対してニュートン法を用いて推定する。IDF 項 $\frac{C(d)}{C(d:w_i \in d)}$ は、単語 w_i の文書出現確率の逆数に対数を取ったものとして定義され、

$$\frac{C(d)}{C(d : w_i \in d)} \approx \frac{C(d) + \beta}{C(d : w_i \in d) + \beta} \tag{3}$$

により求めることができる。ここで、$C(d)$ はあるドメインの文書数であり、文書出現頻度 $C(d : w_i \in d)$ はその単語がいくつの文書で出現したかの数である。 β は α と同様に、ディリクレ過程を用いた事前分布から推定する。

述語項構造パターンに対する評価値は、用言と要素に対して求めた上記の値の相乗平均によって求める。

4.2 Naive Bayes 法

次に、Naive Bayes を用いた手法を考える。この手法では、単語 w_i が現れたときドメイン t である確率を推定する。

$$P(t|w_i) = \frac{C(w_i, t) + x_t\gamma}{C(w_i) + \gamma} \tag{4}$$

$C(w_i, t)$ はドメイン t の中で単語 w_i が出現した個数である。また、 x_t はドメインごとのコーパスサイズを考慮するための正規化係数で、以下の式で与える。

$$x_t = \frac{\sum_j C(w_j, t)}{\sum_k C(w_k)}. \quad (5)$$

γ は、前節と同様に、ディリクレ過程を用いた事前分布から推定したものを利用する。

述語項構造パターンに対する評価値は、TF-IDF 法と同様に、要素と用言に対して求めた上記の値の相乗平均から求める。

これとは別に、述語項構造パターンの３つ組（「（選手名） - が　完投」など）を１つの単位 PA_i として、そのまま $P(t|PA_i)$ を求める場合も考えた。

4.3 固有表現のクラス化

上記の２つの手法は、α,β,γ のスムージングにより未知語の存在を考慮しているものの、コーパスを用いた教師なし学習に依存するため、未知の固有名詞に対して脆弱である。ニュースにおける固有名詞は、時期によって頻度が異なり、コーパスの偏りの影響を大きく受けてしまう。また、Web テキストや新聞記事は、日々新しい単語や固有名詞が出現するため、それを考慮した頑健なパターン抽出を行う必要がある。そこで、固有名詞の中でも、比較的精度が高く抽出でき、出現頻度が高い人名、組織名、地名においてクラス化を行う。

特定のドメインで頻出する固有名詞は、相乗平均を取る前に単語ごとにクラス化してしまうと、当該ドメインでその固有名詞が頻出するという情報の特徴が失われてしまう。また、あるドメインで意味的に重要な述語項構造パターンは、単語同様に学習テキスト中に頻出する。そこで、単純に単語をクラス化するのではなく、述語項構造パターンに対して固有表現のクラス化を行う。つまり、「金本 - が　打つ」「阿部 - が　打つ」という２つの述語項構造パターンは、「（人名） - が　打つ」の形にクラス化され、この述語項構造パターンの評価値は、この形で表せる全ての述語項構造パターンの評価値の和とする。

ただし、述語項構造のパターンを用いるモデルは、データスパースになりやすい。そこで、ヒットしない場合は、単純に単語のクラス化を行い相乗平均を取るモデルへとバックオフすることも考える。

4.4 抽出手法の評価

提案した手法を、日本プロ野球のニュース記事に対して適用し、評価を行った。学習セットを毎日新聞記事データベース（CD-毎日新聞 2008 データ集）とし、4.1〜4.3 節で述べたモデルの学習を行った。また、毎日新聞の Web サイトにおける 2010 年 4 月 21 日-23 日の記事に対して、対話に利用できる典型的な述語項構造パターンに対して人手でアノテーションを行い、評価セットとした。ただし、動詞「する」「なる」が含まれるものについては、重要度をアノテーションすることが難しいので、評価セットから除いてある。今回用いたテストセットからは述語項構造パターンが 559 個が抽出され、人手によるアノテーションによって 248 個を正解とした。これらの適合率・再現率・F 値を求め、選別に用いる評価値の閾値は、テストセットの 10% を利用したヘルドアウトセットの F 値が最も高くなる点に設定した。この結果を表 1 に示す。ベースラインは選別を一切行わず、述語項構造をすべて用いた場合であり、再現率は 100% であるが、適合率は低い。

表 1　述語項構造パターンの抽出手法の評価

モデル	学習	適合率	再現率	F 値
ベースライン	-	0.444	1	0.615
TF-IDF	用言	0.587	0.840	0.691
	要素	0.658	0.730	0.692
	要素＋用言	0.513	0.843	0.638
NB	用言	0.601	0.879	0.714
	要素	0.661	0.794	0.722
	要素＋用言	0.878	0.726	0.795
	要素＋用言＋バックオフ	0.851	0.782	0.815
NB	３つ組	0.871	0.681	0.765
	３つ組＋バックオフ	0.828	0.738	0.780

この結果から、いずれの尺度でも再現率の低下を抑えながら、適合率を大きく向上させることができることがわかった。このようなドメイン固有の重要なパターンを抽出することは、適切な応答生成を行うことに必要であると同時に、語彙を限定して頑健な音声認識を上でも有効であると考えられる。また、TF-IDF 法と比べて Naive Bayes 法の方が有効である。さらに、学習データがスパースになる問題から、バックオフを行うことが有効であることも示された。

テストセットのすべての述語項構造パターン 559 個のうち、抽出されたもののカバー率は 44.7% に過ぎなかったが、ドメインに特徴的な正解パターン 248 個については、78.6% がカバーできていた。すなわち、頻度が少ない多くのパターンはカバーすることが難しいものの、重要なパターンは学習データに頻出し、１年分程度の新聞記事コーパスから得ることが可能である。また、学習できていなかったパターンのうち、今回用いたもの以外のクラス化（イニング数・成績など）を行うことで抽出できるものが 22 個あり、ドメインに適応した固有表現抽出を行うことで、この精度はさらに向上すると予想される。

抽出されたパターンの例を表 2、表 3 に示す。この例に見られるように、「＊ - が　打つ」、「＊ - が　登板する」など、重要な表現が抽出できた。また、人名などのクラス化を行うことによって、表 3 に示すような、さらに詳細なドメイン固有の表現を抽出することができた。

表 2　抽出された述語項構造パターンの例

（意味役割）　用言
（ガ格）（ヲ格）（ニ格）　打つ
（ガ格）　連敗
（ヲ格）　打ち取る
（ガ格）（時間）　登板する
（ガ格）　逃げ切る

表 3　クラス化を用いて抽出された述語項構造パターンの例

（意味役割）　用言
（人名）- が　サイクル安打
（組織名）- が　先発全員安打
（組織名）- が　完封リレー

逆に抽出できなかったものの大半は、学習コーパスにおいて頻度の低いものであり、誤って抽出されたものの多くは、当該のドメインで頻出するが、ユーザとの対話には必要でないものであった。

また、この評価値は、部分一致に基づく発話生成の際に情報選択に利用することができる。ユーザの発話に対して複数の部分一致する情報が現れた場合、提案したような評価値の最も高いスコアを持つものを選択することができる。

5. 部分一致に基づく情報推薦

ユーザの質問から得られる述語項構造パターンで検索し情報を見つけられなかった場合や、ユーザが一定時間発話しないことを検出した場合、システムはそれぞれ部分一致に基づく情報推薦か、直前のユーザの発話状態を元にした情報推薦を行う。このような協調的な応答生成は従来から研究されているが[15]、本研究ではさらにプロアクティブな情報推薦を含む応答生成を目指す。

5.1 部分一致に基づく応答生成と情報選択

述語項構造パターンが完全に一致するものがなかった場合、要素や用言の一致条件を緩め、応答生成に利用する情報を検索する。この際、複数の候補がマッチし、応答生成のためにいずれかを選択しなければならない場合や、ユーザの質問に複数の述語項構造パターンが含まれ、優先度を付けなければいけない場合が考えられる。そうした場合、4 節で提案した評価値を利用したり、述語項構造パターンどうしの類似度を定義することが考えられる。

User:　今日は楽天はどうだったの？
Sys:　楽天が田中の好投で4－1とソフトバンクを破り、2連勝で第2ステージ進出しました。
User:　鉄平は打ったの？
Sys:　鉄平は打っていませんが、四回、1死一塁からのランエンドヒットで好機を広げ、2死一、三塁から中村真の左前適時打で先制しました。
User:　へー。他には？
Sys:　五回には2四球の走者を置いて山崎武が2試合連続の本塁打を放ちました。
...

図 5　本システムによる対話の例

表 4　システムの応答の評価

システムの挙動	正解	不正解	該当情報なし
完全一致	21	3	-
部分一致	10	5	-
合計	31	8	6

5.2 トピックモデル

話題の管理は、自然な対話を行う上で有用であることが示されている。例えば、野球のドメインにおいては、どのチームを応援しているか、どの選手について興味があるかなどである。本システムでは、日本のプロ野球について、各球団の投手・野手の別について 24 クラスを用意した。この際、選手名などのクラスを構成する単語は Wikipedia と選手名鑑から抽出した。将来的には、このトピックモデルの構築も自動化することを目指す。

6. システムの評価

本システムから生成される対話の例を図 5 に示す。システムが抽出情報から完全一致する情報を見つけられる場合はその情報を利用した質問応答を行い、完全一致する情報を見つけられない場合や、長いポーズを検出した場合はプロアクティブな情報推薦を行っている。

今回、システムの応答生成の評価を行うため、4 節で用いたテストセットから、45 個の簡単な質問文を作成し、テキスト入力でシステムの挙動を確認した。その結果を表 4 に示す。質問文は、テストセットの文を読んだ事前知識のない人間が答えられるようにした。その結果、完全一致の場合で生成された応答は 24 個中 21 個が正解で、残る 3 個は構文解析誤りによる検索誤りであった。また、部分一致の場合で生成された応答については、15 個のうち 10 個は的確なものであった。今回はテキスト入力で実験を行ったが、この枠組みは音声

入力で対話を行う際に、より頑健なマッチングを行えることを期待できる。

7. ま と め

本稿では Web のニュース記事から情報抽出を行い、質問応答や情報推薦を行う対話システムを構築した。情報抽出の手法を用いることにより、ユーザの要求に関係の深い応答をプロアクティブに生成できる。また、特定ドメイン内での情報抽出のパターンにおいて、重要なものを選別するための尺度と手法を提案した。これらは、情報抽出に基づく対話生成において効果的であり、音声認識や言語解析を頑健に行うことにも寄与すると考えられる。今後は、ドメインに特化した言語モデルによる音声認識と統合し、音声対話における提案手法の評価を行う予定である。

参 考 文 献

1) 河原達也，川島宏彰，平山高嗣，松山隆司：対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ，情報処理， Vol.49, No.8, pp.912–918 (2008).
2) R.Grishman: Discovery Methods for Information Extraction, *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp.243–247 (2003).
3) L.Ramshaw and R.M.Weischedel: Information Extraction, Vol. 5, pp. 969–972 (2005).
4) Kawahara, T.: New perspectives on spoken language understanding: Does machine need to fully understand speech?, *Proc. IEEE-ASRU*, pp.46–50 (2009).
5) D.A.Dahl: Expanding the Scope of the ATIS Task: The ATIS-3 Corpus, *Proc. ARPA Human Language Technology Workshop*, pp.43–48 (1994).
6) R.Pieraccini, E.Tzoukermann, Z.Gorelov, J-L.Gauvain, E.Levin, Lee, C.-H. and J.G.Wilpon: A Speech Understanding System Based on Statistical Representation of Semantics, *Proc. IEEE-ICASSP*, Vol.1, pp.193–196 (1992).
7) 安達史博，河原達也，奥乃 博，岡本隆志，中嶋 宏：VoiceXML の動的生成に基づく自然言語音声対話システム，情処研報，SLP-40-23, pp.133–138 (2002).
8) Misu, T. and Kawahara, T.: Dialogue strategy to clarify user's queries for document retrieval system with speech interface, *Speech Communication*, Vol.48, No.9 (2006).
9) Misu, T. and Kawahara, T.: Bayes Risk-based Dialogue Management for Document Retrieval System with Speech Interface, *Speech Communication*, Vol.52, No.1, pp.61–71 (2010).
10) 翠 輝久，河原達也，正司哲朗，美濃導彦：質問応答・情報推薦機能を備えた音声による情報案内システム，情処論， Vol.48, No.12, pp.3602–3611 (2007).
11) 河原大輔，黒橋禎夫：自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル，自然言語処理， Vol.14, No.4, pp.67–81 (2007).
12) Shen, D. and Lapata, M.: Using Semantic Roles to Improve Question Answering, pp.12–21 (2007).
13) Wang, R. and Zhang, Y.: Recognizing Textual Relatedness with Predicate-Argument Structure, *Proc. EMNLP-2009* (2009).
14) Wu, D. and Fung, P.: Can Semantic Role Labeling Improve SMT?, *EAMT-2009* (2009).
15) D.Sadek: Design Consideration on Dialogue Systems: From Theory to Technology - The Case of Artimis -, *Proc. ESCA workshop on Interactive Dialogue in Multi-Modal Systems*, pp.173–187 (1999).