CRF を用いたテレビ番組クローズドキャプションからの質問-解答対自動抽出

河野 将弘* 奥村 学† 徳永 健伸‡ 三浦 菊佳§ 山田 一郎§ 住吉 英樹§ 八木 伸行§

1 はじめに

近年，情報大容量化の時代を迎え，世間に発信される情報量は爆発的に増えつつある．その中で，多くの分野において，情報の簡略化や要約の技術に対する要請がますます高まっている．

日本では 1997 年の放送法の改正によって，努力規定として「聴覚障害者に対する説明」が盛り込まれ，クローズドキャプション付きのテレビ番組が増えている．放送局では多種多様な番組が日々大量に制作されているが，現段階ではそれらが十分活用されているとは言えない状況にある．

我々は，この膨大なテレビ番組のデータを利用し，既存の番組情報を新たな視点から提供するシステムを作ることを目的としている．その一環として本研究では，クローズドキャプションを素材として，番組内に存在する質問文とそれに対応する解答文を抽出する．質問文と解答文には，その導入・補足となる情報も重要であることから，これらも合わせて抽出する．

2 クローズドキャプションに対するアノテーション

機械学習の正解データとなり，また評価の際の指針ともなる，クローズドキャプションに対するアノテーションについて述べる．用いるコーパスは，NHK 番組「地球！ふしぎ大自然」19 番組分[1]である．アノテーションのスキーマとしては，図 1 を用いる．

長方形で示されたものがセグメント，破線矢印で示されたものがセグメント間のリンクである．以下に，各セグメントの詳細を示す．

- 質問文の中心
  疑問詞と疑問符両方を含む一文を質問文の中心とする．

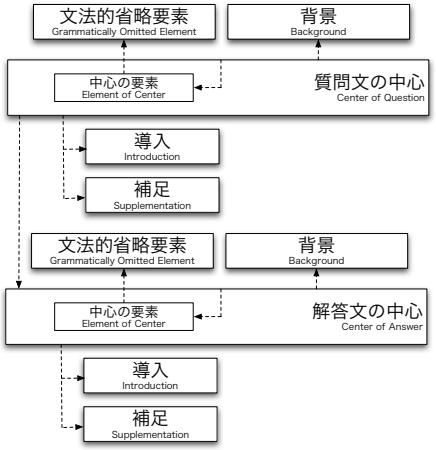

図 1: アノテーションスキーマ

- 解答文の中心
  質問文の中心の述語に注目する．同じ述語が解答となる部分に存在する場合，それを含む一文を解答文の中心とする．このような文が複数ある場合は，より解答の内容を明確に表現している方を選択する．同じ述語が解答となる部分に存在しない場合は，解答要素を解答文の中心とする．解答要素とは，質問文の解答が句となる場合は，その句を指す．解答が文となる場合は，その一連の文を指すが，解答を表現する最短の文とする．
- 導入および補足
  質問文または解答文の中心の前にある，中心に関連する一連の文を導入とする．質問文または解答文の中心の後にある，中心に関連する一連の文を補足とする．導入の始点と補足の終点は以下の点と定める．
  - 場面や話題が完全に切り替わる点．場面の切り替わりは以下の表現を手がかりとするが，これらが明示的に表れないこともある．
    * 体言止め
    * 時間の表現
    * 場所・場面の表現
    * 接続詞
    * 新たな動物の登場

* 東京工業大学 工学部 情報工学科
kawano.m.ab@m.titech.ac.jp
† 東京工業大学 精密工学研究所
‡ 東京工業大学 大学院情報理工学研究科
§ NHK 放送技術研究所

[1]NHK. 2004. 5/10, 2005. 4/11, 4/18, 5/2, 5/9, 5/23, 6/6, 7/11, 7/25, 10/10, 10/17, 10/24, 11/14, 11/28, 12/5, 12/12, 2006. 1/16, 1/30, 2/6 放送分

    * 視点の大きな変化
    * 話題を変えるための発問
  - 新たな質問文の中心，またはその導入が現れた点
- 背景
  導入の中で，質問文または解答文の中心に直接関係する文を背景とする．
  1. 質問文または解答文の中心に直接関係するため，背景とする文
     - 質問文または解答文の中心の，内容に関する特殊性を説明している文
     - 質問文または解答文の中心の根拠，もしくは，それに至る動物の性質，状況，経緯等を説明している文
     - 質問文または解答文の中心の内容を限定，もしくは，程度を説明している文
     - 背景に直接関係する文
  2. 質問文または解答文の中心に直接関係しないため，背景から除外する文
     - 質問文または解答文の内容を具体化しただけの文．ただし，質問文又は解答文の中心の内容が抽象的すぎる場合は，具体化したところを背景とする．
     - 地名などの名詞，実際の数値，動作の描写は特に背景とはしない．ただし，その名詞や数字の解釈は重要なため，背景とする．
     - 動物などの登場場面
     - 単純な映像説明で背景情報が含まれない文
     - 質問文または解答文の中心と内容が同じ文
     - 視聴者を引きつけるためだけの表現 (比喩等)
     - 情報を含まない文
- 中心の要素および文法的省略要素
  質問文の中心または解答文の中心において，質問を行う上で必須の単語が欠けていることがある．そのうち今回は，ガ格，ニ格，ヲ格，所有格を扱う．それらが導入または補足の中に存在する場合，これを文法的省略要素とする．また，文法的省略要素の係り先の単語を，中心の要素とする．

3 質問-解答対抽出手法

質問-解答対を自動抽出する手法は，以下の 3 段階からなる．

1. 質問文の中心から，CRF[5] を用いてその導入部を同定する．
2. 解答文の中心 (一文) を推定する．
3. 推定された解答文の中心から，CRF を用いてその導入・補足部を同定する．

以下では，各段階ごとにその詳細を述べる．

3.1 質問文 (導入) の同定

まず，番組内の一つの質問文に注目し特徴抽出を行う．注目した質問文に対し，それを含む番組の全ての行に対して以下の特徴を抽出し，CRF の素性とする．

- 語彙的連鎖のリンク数
  望月らによれば [2]，意味的に同じ，あるいは，関連する二つの語によって示される文と文の表層的な関係を，語彙的結束性という．また，文書内での語彙的結束性を持つ語の連続のことを語彙的連鎖 (lexical chain) と呼ぶ．語彙的連鎖は以下の三つに分類される．
  1. 同一の語の繰り返しに基づく語彙的連鎖
  2. シソーラスに基づく語彙的連鎖
  3. 語の共起関係に基づく語彙的連鎖

  今回は，語彙的連鎖として 1 と 2 を扱う．
  まず，語彙的連鎖計算プログラム Lexical Chainers[3] を用い，語彙的連鎖を抽出する．シソーラスは分類語彙表 [4] を用いた．ここで抽出された語彙的連鎖を構成する語同士の結びつきを，リンクと呼ぶ．注目した質問文の主語がある文から，質問文の中心までの文より発生するリンクのみを考える．各行に対し，その行から出るリンクの数を計算する．
  なお，質問文の主語に関しては，赤岩ら [1] の定義に従う．すなわち，質問文の中心となる文中の，ガ格またはハ格の一般名詞，固有名詞，サ変名詞，未知語とする．該当がない場合，その文のトピックを主語とする．トピックには，質問文の主語で指定した格に加えて，モ格のものも採用する．さらに，後続助詞が断定の助動詞であるか，体言止めとなっている一般名詞，固有名詞，サ変名詞，未知語についても採用する．それでも該当がない場合は，その文のトピックは前の文と同じとする．
- 注目する行から 2 行先までの，語彙的連鎖のリンク数の和
- 質問文に関する属性

注目した質問文か，それ以外の質問文か，そうでないか．

- サブコーナー[2]の句切り
- 月の表現の位置
- 体言止めの位置
- 注目した質問文からの行数

学習させる正解データには，先のアノテーション済みコーパスのデータを用い，BIO タグを割り当てる．語彙的連鎖について考察したところ，導入部と背景部の区別が難しいことがわかったため，今回はそれらを同一視し，B-Q，I-Q，もしくは O を各文につける．それぞれ，質問文の開始行，途中行，それ以外の行を意味する．

3.2 解答文の中心の推定

解答文の中心の推定には，赤岩ら [1] の手法を用いる．三つ紹介されているアルゴリズムのうち，精度が最大のアルゴリズムを用いる．これは，主に文の類似度を用いて解答位置を検索する手法である．まず，質問文から後の各文に対し，語彙的連鎖のリンク出現回数，主語の一致性，特定の類似表現，指示語の出現で重みづけをする．加えて，質問文の後二文以内に場面が遷移しない場合は，その後に続く文が解答文となっている場合が多い．そのため，該当する一連の文には大きな類似度を与えている．次に，三文を窓として類似度を合計し，合計類似度が最大となった一場面を抽出する．ただし，今回は解答文の中心を一文のみ抽出したいので，類似度が最大の「一文」を抽出する．また，番組固有の情報の使用をできる限り排除するため，以下のルールに関しては実装を見合わせた．

- 同一と判断された質問文に対しては，同じ解答文を与えるルール
- 冒頭および総括を解答位置から除外するルール

3.3 解答文 (導入・補足) の同定

この段階に関しては，3.1 節とほぼ同一の手法を用いる．ただし，学習させる正解データとして，今回は B-I(A)，I-I(A)，B-S(A)，I-S(A)，O を用いる．I(A) は解答文の導入，S(A) は解答文の補足を表す．また，用いる素性も変更する．変更のあるものだけ以下に記す．

- 質問文と解答文に関する属性
  3.1 節で採用した素性のうち，「質問文に関する属性」を変更する．注目した質問文か，それ以外の質問文か，そうでないかに加えて，訓練データとテストデータの場合それぞれに，以下の情報を含める．
  - 訓練データの場合
    アノテーション済みコーパスに基づき，解答文の中心の位置を含める．
  - テストデータの場合
    推定した解答文の中心位置を含める．
- 解答文の中心からの行数
  3.1 節で採用した素性のうち，「注目した質問文からの行数」を変更する．前であれば負，後であれば正とする．

4 評価実験

4.1 評価

まず，解答文の中心の推定精度に対しては，以下の正解率で評価する．

$$\text{正解率} = \frac{\text{正しく解答文に含まれていた数}}{\text{推定された解答文の中心の数}}$$

3.2 節の手法における解答文の中心の正解率は，67.4% であった．

3.1 節と 3.3 節の，質問文の導入，解答文の導入と補足に対する評価は，精度，再現率，F 値を用いる．解答文については，正しく解答文の中心が推定できたもののみ評価対象とする．それぞれの定義は以下の通りである．表 1 に，結果を百分率で示す．

$$\text{精度} = \frac{\text{正しく質問文/解答文と判断された文の数}}{\text{質問文/解答文と判断された文の総数}}$$

$$\text{再現率} = \frac{\text{正しく質問文/解答文と判断された文の数}}{\text{実際の質問文/解答文の数}}$$

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

表 1: 質問文抽出後に解答文を抽出する手法

	精度 (%)	再現率 (%)	F 値
質問文	70.1	71.5	70.7
解答文	80.8	69.0	74.4

[2] 番組内において 2 回程度にわたって挿入される特集

4.2 考察

まず，解答文の中心の推定精度に関する考察を行う．赤岩ら [1] によれば，解答文位置の同定精度は 75.3%となっている．今回数値が 8%ほど低いのは，番組特有の条件を外したことが最大の要因であると考えられる．また，以前の手法では評価方法が明確ではなく，「どのようなとき解答が正しく抽出できたと言えるのか」について詳しく言及されていない．評価方法のあいまいさによる結果のゆれがある程度想定される．

続いて，質問文の導入と，解答文の導入・補足の推定に関する考察を行う．解答文の中心位置を先行研究の手法で特定できたため，それを用いることで精度，再現率ともに高い結果が得られた．

誤りとしては，自明であるが以下の２つに分けられる．

- 区切れるべきところで区切れていない
  導入や補足が長く抽出され，精度低下の原因となる誤りである．
- 区切れるべきところでないところで区切れている
  導入や補足が短く抽出され，再現率低下の原因となる誤りである．

いずれの場合も，体言止めや月の表現など，比較的区切れやすい素性が出てきたところで途切れてしまうことが原因である．これを改善するためには，ある程度コーパスに合わせ，区切れやすいキーワード等の素性を加えることが一応の解決策になる．

また，精度や再現率を下げる要因を特定し分類することは，機械学習を用いた手法のため，一般には難しい．その難しさの要因の一部は，コーパスにアノテーションをする際の方針決定の難しさと関係している．人手でアノテーションを行った際，かなり細かく方針を決定しようとしたが，最終的には文脈から判断せざるを得ない場面が多く見られた．もちろん，体言止めや月の表現など，場面が区切れやすい素性はいくつか見つかったが，それが出現すれば必ず区切れるわけではない．文脈判断に相当する部分に語彙的連鎖を用いたが，それで全ての文脈を判断することは難しかったと言わざるをえない．

本手法を改善するにあたっては，類似度に加えて，構文木の構造類似度 [8] 等を用いた素性も文脈を判断する上で大きな手がかりとなるだろう．また，統語的な情報の他にも，意味論的な解析を用いてより高度な素性を取り入れることも考慮に入れることも考えられる．また，今回の手法では，解答文の中心は一文と仮定したが，これが複数の文になっても，この手法を用いることができる．そのため，解答文の中心の精度向上とともに，複数の文によって中心を推定することで精度向上につながると考えられる．

5 まとめ

NHK「地球！ふしぎ大自然」のクローズドキャプションに対して，質問文及びその質問に対する解答位置の抽出を行った．特に，質問文の導入と解答文の導入・補足を場面変更点によらず同定し，評価方法を明確にした．ただし，解答文の中心位置の同定法や，CRF の文脈を学習させるための素性にはまだ改善の余地があり，例えば統語的・意味的な素性による精度の向上が課題として挙げられる．

参考文献

[1] 赤岩 怜, 徳永 健伸. クローズドキャプションからの QA 抽出. 東京工業大学, 学士論文, 2006.

[2] 望月 源, 奥村 学, 岩山 真. 語彙的結束性に基づく語彙的連鎖の計算. JAIST Technical Memorandum, IS-TM-2000-002.

[3] 望月 源. 語彙的連鎖計算プログラム Lexical Chainers version 1.50.2 マニュアル. JAIST Technical Memorandum, IS-TM-2002-001.

[4] 国立国語研究所編. 分類語彙表 増補改訂版.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of ICML-01, 282-289, 2001.

[6] CRF++ 0.49. http://crfpp.sourceforge.net/.

[7] 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. 日本語形態素解析プログラム茶筌 version 2.3.3. 奈良先端科学技術大学院大学情報科学 研究科自然言語処理学講座 松本研究室. 2003.

[8] 箱田 慶太, 市川 宙, 橋本 泰一, 徳永 健伸. 構文的類似度を用いた文の検索. 言語処理学会第 12 回年次大会, pp.1131 – 1134, 2006, Mar.