言語横断質問応答システム

関根聡

ニューヨーク大学
sekine@cs.nyu.edu

1. はじめに

本稿では、英語で質問をしてヒンディ語の新聞記事からその解答を探し、英語に直して解答を提示する言語横断質問応答システムを報告する。現在、「インドの首相は誰ですか」というような自然文で書かれた質問に対し、直接、「バジパイ」とその答えを返す質問応答という研究が盛んに行われており、実用的にも期待されている [Voorhees 2000] [NTCIR QAC 2002/2003] [Moldovan et al. 2003]。しかし、これまでは質問と答えのある情報ソースはほとんどの場合同一言語であった。だが、世界の情報は様々な言語で書かれており、特に即時性、ローカル性を追及すると、単一言語だけでは十分ではないという実情がある。日本では特に、日本語で質問して英語の中から解答を捜すという需要は高いものと思われる。このような状況に鑑み、CLEF（Cross Language Evaluation Forum）2003 [Magnini 2003a]でもひとつの課題としてヨーロッパ語間の言語横断質問応答の評価が行われるなど、その実現が試されている。今回、我々は次に述べるサプライズ・ランゲージのプロジェクトの一環として英語とヒンディ語の間の言語横断質問応答システムを開発した。

2. サプライズ・ランゲージ

サプライズ・ランゲージ（Surprise Language Exercise = SLE）のプロジェクトは米国の国防省系のDARPAが行っているTIDES (Translingual Information Detection, Extraction and Summarization) [TIDES HP] というプロジェクトの中で行われた。DARPAは基本的に、米国の国防に関わる技術の研究開発を取り仕切る団体である。米国の国防上、現在のような世界情勢の下では、いつどこの国が戦略的に重要になるか分からない。もし、ある国で何かが起こったら、その国の情報を得ることは非常に重要である。もし、その国の言語が英語やよく知られている言語でなかったらどうするか。時間は限られている。自然言語処理の技術は使えないか。例えば、新規の言語に対して機械翻訳システムを短時間で実現するのは可能なのか。これらを実証するために、TIDES に参加している米国の１５程度の大学、研究所に課せられたプロジェクトが SLE である。参加者はある日、なんらかの言語名を知らされ、それに関するシステムを 1 ヶ月で作成するというものである。実際には2003年6月2日に対象の言語として「ヒンディ語」が指定され、6月30日までに４つの分野、すなわち、１）機械翻訳、２）外国語のテキストから情報検索をしようという言語横断検索(Cross Lingual Information Retrieval)、３）人名、組織名などの名前を抽出する固有表現抽出、そして、４）要約のそれぞれの分野においてシステムを作成し、評価しよう、というものであった。本来、NYU のグループは固有表現抽出だけをすればよかったのだが、それを利用した応用としてここで報告する言語横断質問応答システムを開発した。

SLE のプロジェクトは非常に協力的に行われた。大量に流れたメイリングリストでも、どこかで見つけたり、独自に作成されたデータやツールの情報が多く流れた。日本では IREX、NTCIR など評価型のプロジェクトが行われているが、1 つの目的に向かって協力して行う形のプロジェクト（日本でもICOT、EDR などがあったが）も目的によっては興味深いものであるという印象を受けた。

3. システム

システムは大きく、データにタグ付けする部分と、実際の質問応答をする部分の2つに分かれる。図1にシステム概略図を示すが、それぞれこの図における右上の部分と残りの部分に該当する。

タグ付けでは、ヒンディ語の新聞記事に対して固有表現と数値表現をタグ付けした。固有表現タガーは、人名、地名、組織名に対して、60万単語分のトレーニングデータを利用して作成された HMM の

タガーである。このタガーの精度は同じ種類の新聞記事で評価したところ、適合率 82%、再現率 74%であった。数値表現タガーは数値に続けて書かれる数助詞や名詞をキーにしたパターンベースのもので、年齢、金額、割合、期間、長さ、広さ、体積、速度、温度、時刻、重量、数量、国数、場所数、人数の 15 種類を対称にしているである。限られた時間で開発したために精度は高くなく、適合率 52%、再現率 35%であった。このようにして新聞記事中で固有表現または数値表現がタグ付けされた表現は質問のタイプに従った答えの候補となる。例えば、質問が「長さ」に対するものである場合には「長さ」のタグ付けされているものが答えの候補となる。

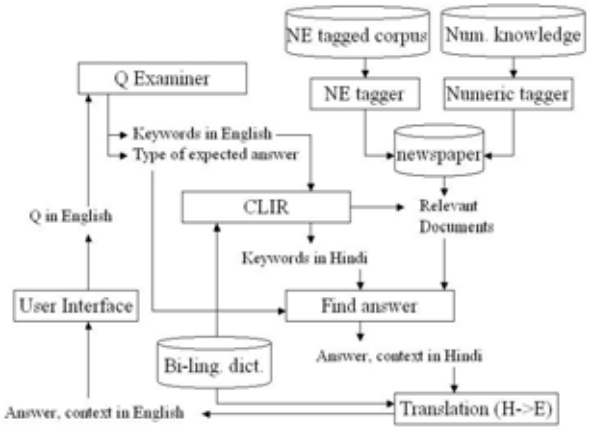

図 1. システム概要図

質問応答システムは基本的には翻訳部分を除くと、単言語でよく使用されている方法と同じものである。[Voorhees and Tice 2000]。つまり、言語横断質問応答システムでは、質問解析(QE)、 言語横断情報検索システム(CLIR)、解答抽出(AF)、および結果を英語で表示するための翻訳(MT)の４つのコンポーネントからなる。

QE では、システムに入力された英語の質問文を POS タガーとチャンカーを使って解析し、期待される解答タイプとキーワードを抽出する。解答タイプは固有表現の 3 種類か数値表現の 15 種類のどれかである。タイプが認定できない場合には全てのタイプを可能な解答タイプとする。キーワードはストップワードが除かれ、品詞の種類や DF などによって重み付けられ、以下の処理で使用される。基本的にこの部分は前に開発していた英語の質問応答システムの質問解析部分が流用できた。

CLIR では質問から抽出されたキーワードを使用して、ヒンディ語の新聞記事から関連する記事を検索する。基本的には、ヒンディ語の新聞記事を使った DF に基づき、TFIDF をスコアリングの方法として使用している。キーワードは英語－ヒンディ語辞書を使って翻訳される。翻訳曖昧性については全く対処せず全ての翻訳語を使用する。ここで使用された翻訳辞書は、ウエッブで入手可能なものや他の SLE 参加者が作成したものを利用した。特記すべきなのは、IBM がパラレルコーパスから自動的に抽出した翻訳辞書である。IBM が持つパラレルコーパスは公開されなかったが、そこから抽出したこの辞書は公開された。この辞書の性能は誤った翻訳語を多く含み、明らかに悪いものであったが、実験ではこの辞書を使用した方が CLIR の成績は良かった。複数のキーワードによる補正効果があったものと考えられる。

AF では CIRL で検索された記事のうち高いスコアの 20 記事の中から解答を抽出する。QE で判定された解答タイプにマッチする表現のうち、式１のスコアが高いものを解答候補とする。

$$score(e) = \sum_{k \in keywords} \frac{C_1}{dist(k,e) + C_2} * keyword\_weight(k) * article\_score \quad (式1)$$

ここでC_1 とC_2 は定数、*dist(k,e)* は表現とキーワードの単語距離、*keyword_weight* はQEで求められたキーワードの重み、*article_score*はCLIRで求められた記事のスコアである。より正確率を高めるためには、言語的、意味的な知識を導入する必要があるが、対象言語がヒンディ語であり複雑な処理を組み込む時間もなかったために簡単な処理を利用している。結果的には、単言語でも利用されている通り、このような簡単な処理でもある程度の精度が得られた。

MT では、得られた解答と共に、解答があった文章を翻訳し、ユーザーがその解答の信頼性を測れるようにしている。翻訳は ISI の開発した文単位のヒンディ語－英語の翻訳システムと、単語単位の翻訳の 2 つが実現されている。ISI の翻訳システムも完全ではないし、きちんとした英語が作成されない場合もあるため、単語単位の翻訳も利用価値がある。

システムはWEB 上で利用できるようになっており、ユーザーが質問を入力し、３－５秒後に解答が表示される（図２，３）。ISI の翻訳システムは現在動いていないが、システムを利用してみたい場合には連絡をいただければ、パスワードをお送りする。

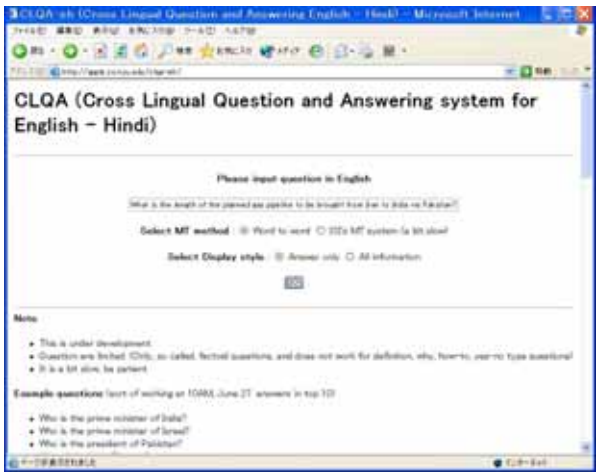

図 2.　質問入力画面

図３．解答表示画面（ISI 翻訳システム利用）

４．　評価

我々は、ヒンディ語を母国語とする人に作成してもらった質問を基にシステムの評価を行った。基本的には 1 ヶ月で作成したシステムであるが、ある程度の質問に答えられ、評価者にも直感的にはある程度利用価値のあるものとの判断を受けた。質問は、新聞記事を見て解答となるものを探してもらい、それから質問を英語で作ってもらった。したがって、TREC の QAC トラックの初回と同様、質問が簡単になっているという指摘は免れない。全部で 56 個の質問が作成され、タイプとしては、人名(19)、組織名(6)、地名(4)、年齢(2)、時間(7)、重さ(4)、長さ(2)、割合(2)、人数(2)、金額(2)、質量(1)、スピード(1)、数量(4)となっている。図４に質問のサンプルを正解の解答を得た順位と共に示す。-は 10 位以内に正解が得られなかったという意味である。

1: Who is the prime minister of India?
1: Which company created a video game on Iraq War?
1: When was modern Jewish nation establishd in Tel-Aviv?
-: When was Shimla Agreement signed between Idea and Pakistan?
3: How much money was involved in Bali bomb attach?
-: What was the age of Harivansh Rai Bachhan when he died?

図 4.　サンプル質問

評価結果は、56 質問中、10 位以内の解答候補に正解が含まれていたのは 25 個であった。この中には３つのトランスリタレーション（ヒンディ語の英語表記。英単語を日本語のカタカナで表現するようなもの）の正解も含まれている。上位 5 位までででの MRR(Mean Reciprocal Rank)は 0.25 であり、短期間で全く新しい言語を対象とした言語横断質問応答システムを作成した背景を考えると、悪くない結果だと考えている。

５．　結果分析

各段階における結果の分析を行う。表２は QE での解答タイプの推定誤りの結果を示す。英語の質問解析を利用している点と、質問作成者がやや優しい問題を作ったためか、高い精度が認められる。

正しい解答タイプ	51
誤った解答タイプ	3
解答タイプが見つからない	2

表２．　QE での解答タイプ推定正解率

解答タイプが見つからない２つの質問は“ Which”で開始される質問文で、主辞名詞がタイプ判定辞書に存在しなかったためにタイプが見つからなかったものである。誤ったものは”How many/much”で始まる質問文が数量と体積などの曖昧性があるために誤ったものである。

次にCLIRの精度を見た。ここで使用した辞書は短期間で集められたものでありごみが非常に多く混ざっている。例えば、”India”の翻訳では、11個のヒンディ語の翻訳中3個だけが正しいものである。この誤りは対訳データから統計ベースで自動作成されたIBMの辞書によるものが多い。56質問において268の英語のキーワードが抽出され、そのうちの182個については辞書項目が存在したが、そのうちの105には誤った翻訳後が含まれていた。1つの英語に対しては平均3.4個のヒンディ語の翻訳が存在していた。翻訳のなかった86単語はほとんど固有名詞であった。翻訳の存在していない単語を含む質問文と正解を導けたかどうかという関係を見ると、高い相関が見られ、言語横断質問応答システムにおいて翻訳辞書の重要性が認められる。

最後のヒンディ語から英語への翻訳では、４つの間違えが認められた。そのうち２つはトランスリタレーションのときに生じた英語のスペルミスであり、残りの２つは数値表現の数助詞の翻訳ミスであった。後者は辞書に適切な翻訳が存在しなかったという問題である。

５．　結論

本論分では、英語で質問を入力し、ヒンディ語の新聞記事から解答を抽出し、結果を英語に直して表示する言語横断質問応答システムを紹介した。１ヶ月という短期間で開発されたにもかかわらず、ある程度の精度が単言語の質問応答システムと同じような仕組みで実現できることを証明した。結果を分析してみると、NE タガーの精度と翻訳の精度が全体の精度の劣化に影響を与えており、今後の課題となった。また、このシステムはTIDESのSLEのプロジェクト内で実現されたものであり、このプロジェクトにおいて１つの目標に向かって複数の研究機関が協力的に進めた様子はいろいろな意味で価値のあるものであった。

６．　謝辞

本研究は、TIDES の多くの研究者との協力があって成し遂げられたものである。SLE のリーダー及び、データ収集、ツール作成、ノウハウの蓄積とそれらの共用利用を推進した多くの参加者に感謝したい。特に、言語データ(MITRE)、翻訳システム(ISI)、辞書(LDC, IBM, CMU, SPAWAR, U. of Sheffield, ISI, BBN, NYU)、ツール(UMD, Alias-I)、 CLIR の重要なノウハウ(CMU)、NE データ(BBN, LDC, NYU)に感謝する。また、本システムで利用したヒンディ語 NE タガーの開発を行った NYU のグリッシュマン教授と、開発を手伝ってくれた 2 人のインド人にも感謝する。

参考文献

HARABAGIU, S., PASCA, M AND MAIORANO, S. 2000. Experiments with Open-Domain Textual Question Answerin, *Proc. of International Conference on Computational Linguistics (COLING 2000),* pp. 292-298.

MAGNINI, B., ROMAGNOLI, S., VALLIN, A., HERRERA J., PENAS A., REINADO V., VERDEJO F. AND RIJKE M. 2003. The Multiple Language Question Answering Track at CLEF 2003. *CLEF-2003 workshop homepage*, http://clef.iei.pi.cnr.it:2002/

MOLDOVAN, D., CLARK, C. AND HARABAGIU, S. AND MAIORANO, S. 2003. COGEX: A Logic Prover for Question Answering, *HLT-NAACL 2003,* Edmonton, Alberta, Canada,166—172

NTCIR-QAC 2002/2003 http://www.nlp.cs.ritsumei.ac.jp/qac/index-j.html

TIDES HP: http://www.darpa.mil/ipto/programs/tides/

VOORHEES, E. and TICE, D.2000. “Building a Question-Answering Test Collection”, Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 192-199.

VOORHEES, E., 2002. Overview of TREC-9 question answering track. Text Retrieval Conference 9.