Generating Instrumental Expressions in a Multilingual Question-Answering System

Monojit Choudhury (1), Elixabete Murguia (2), Sudeshna Sarkar (1), Véronique Moriceau (2), Asanee Kawtrakul (3) and **Patrick Saint-Dizier (2)**
(1) IIT Kharagpur, India, sudeshna@gmail.com
(2) IRIT-CNRS, Toulouse, France, stdizier@irit.fr
(3) Kasetsart University, Bangkok, Thailand, asanee_naist@yahoo.com .

Abstract

Instrumental questions are special type of factoid questions that ask about the instrument that is/has been used to accomplish a particular task. Unlike the other factoid questions (who, what, when, etc.), there has been very little work on the instrumental questions ("with/by what"). In this paper, we investigate the issues involved in answering instrumental questions in a multi-lingual setup, and analyze the morpho-syntactic realization of the instrumental case across six languages (Bengali, English, French, Hindi-Urdu, Spanish and Thai). Based on the analysis, we propose a fine grained classification of the instrumental questions and present the decision graphs for fIVE languages. The decision graphs are helpful in analysis and classification of the query as well as generation of the response.

1 Introduction

The last decade has seen an ever increasing interest in web-based Question-Answering (QA) systems, which, if successful, would replace the current keyword based search-engines to a large extent. Indeed, considerable progress has been made in answering simple factoid questions, where the answer to the question is a single word or phrase that can appropriately replace the question word (e.g. who, where, which, when) in the query sentence. List type and definitional questions have also been dealt with to a good extent. Nevertheless, the current technology is still not suitable to answer more complex questions that require both explicit or implicit reasoning on the query and/or on the documents.

Answering questions about instruments required to realize an action is a rather new area in question-answering (QA). Instrumental questions (a simple case of procedural questions) are the most frequent questions on the Web after factoid questions. They range over technical matters, but also, and primarily over social and everyday life problems. While answers to procedural questions may be a procedure or sequence of steps, when dealing with instrumental questions we typically like to restrict the answer to a phrase (just one instruction). Instrumental question answering tries to find the instrument that allows an agent to accomplish a goal (Kawtrakul et al, 2006). For procedural questions, one is looking for a plan to accomplish the goal. Questions such as

Q1 What is used for cutting glass?

Q2 What do you open a lock with?

Q3 Which algorithm can we use to sort an array of numbers?

Q4 How do you travel from Paris to London?

are examples of typical instrumental questions. The syntactic variations in the questions, answers and documents, as well as the vagueness around the notion of instrumentality make the instrumental questions far more difficult to answer, even though they can be conceptualized as a subclass of factoid questions. While the preposition "with" in Q2 hints at the instrumental nature of the required answer, in Q1 the verb phrase "is used" plays the same role. Same amount of syntactic variation is also observable in the syntactic structure of the possible answers to question Q1, as shown below.

A1 Glass is cut by/with a *diamond cutter*.

A2 *A diamond cutter* is used/employed/applied to cut glass.

As we shall see shortly, there are crosslinguistic differences regarding the notion of instrumentality as well as its syntactic realization. These issues call for a special treatment of instrumental questions, especially in a multilingual question answering setup (e.g. Dorr 1997).

When using the Web, one of the main difficulties is to sort among the large number of responses a prototypical response. Indeed, on the web, we find a very large number of metaphorical uses or unexpected uses for instruments or for actions (e.g. write with a pen/with your heart). To resolve this problem, approaches around data integration (Moriceau 2006) need to be developed. Another difficulty, as shall be seen below, is the conceptual proximity of the notion of instrument with other notions like paths (which can be instruments also) and manners.

In this paper, besides outlining the different conceptual facets of instrumentality, we investigate and compare the notion of instrumentality and its surface realization across six languages, namely Bengali (B), English (E), French (F) Hindi-Urdu (H), Spanish (S) and Thai (T). Based on our analysis, we provide a fine grained classification of instrumental questions and construct the decision graphs for these languages. The decision graph enumerates the morpho-syntactic features used to realize the different subtypes of instrumental case in a language. Thus, these graphs are useful in both analysis of questions of Web or textual documents, and for the generation of the answers.

The end objective of this work is to develop a QA system, where the question (which asks about some instrument) is in language L_1, the documents are in a variety of languages $(L_1, L_2, \ldots, L_n)$ from which the answer(s) is/are extracted and finally the answer is again presented in the language L_1. This requires at least the following four steps:

1. analysis of the question and its classification,
2. data extraction from documents or web pages via a query,
3. data integration, when there are several different responses, and answer consolidation,
4. production of a response in natural language (NL) in the user's language.

We focus on the first and last points in this paper, in particular on the semantic parameters that govern morphological, lexical and grammatical aspects of instrumental expressions, so that queries and documents can be analyzed, and responses can be produced in a variety of languages. In 2004, around 35% of the Web is reported to contain English documents. There are many languages which have large number of speakers but their presence is negligible in the Web. So it may be worth extracting answers to questions in English and produce the output in the language of the user.

The paper is organized as follows. Section 2 discusses the different issues and challenges involved in answering instrumental questions. Section 3 describes a classification of instrumental questions achieved through a cross-linguistic analysis. In section 4, the decision graphs are presented for five languages; these graphs describe the lexical, morphological and syntactic features that are used for the realization of the different subclasses of the instrumental case. In section 5, we provide some preliminary ideas on how the decision graphs can be used to analyze a question and generate the response. Section 6 concludes the paper by summarizing the work and future directions of research.

2 Issues in Answering Instrumental Questions

In order to understand the challenges involved in answering instrumental questions, we describe below a general framework of a QA system for answering instrumental questions and identify the issues related to the different steps/parts of the system. The system handles queries in the form of a task, and aims at finding instruments for executing it. Any available corpora and the web is searched to retrieve documents that include the task and possibly the answer. The steps involved are (for a concrete example, we assume that the query is in Hindi):

1. The query is presented to the system in Hindi (or any source language (SL)). It can be real NL or simplified language, skipping useless words (How dismount graphic card).
2. The query is processed to identify the "task" for which the instrument is queried for and it is categorized into some predefined "query type" (described in Section 3). This step would require a question parser in language SL. To construct this parser it is important to know various marks used in SL to denote the instrumental case. Other elements need also to be taken into account like the verb.
3. The documents in several target languages are searched for the "task" specified in the query. A bilingual dictionary is necessary to translate the task to these other languages. Again, the search process is dependent on the "query type" and the language. We need a language specific parser which identifies the instrumental roles in a given sentence. This is by far the most complex NL aspect.
4. The search results are analyzed and responses are obtained.
5. The responses are clustered (for similar or nearly similar instruments) and categorized according to the prototypicality of the instruments and other factors. At this level, prototypicality must be dealt with in order to rule out the numerous cases which are inappropriate (metaphors, jokes, etc.). This is carried out by means of data integration techniques (Moriceau 2006).
6. Responses are combined to generate the answer, using a template-based approach to simplify the process and to make it less ambiguous.
7. The final answer presented to the user is in Hindi (SL) that is generated through an appropriate natural language generation (NLG) system.

2.1 Notion of Instrumentality

It is far from obvious which queries are to be classified as instrumental types and what objects qualify as valid instruments to that query. In WordNet (Fellbaum 1993) it is defined as 'an artifact, or a set of artifacts, that are instrumental (i.e. behave as instruments) in accomplishing some end', i.e. reaching a certain goal. In this definition, the triple relation agent-instrument-goal (as in: *John cuts the bread with a knife*, where *John* is agent, *knife* is instrument that does the cutting, and 'bread cut' is the goal or target resulting state), is left vague in what concerns the exact involvement of the agent and the instrument in the action, and the control the agent has on the instrument and on the action. In this defintion, instruments are limited to artifacts, which sounds too restrictive, as will be seen below. Similarly, nothing is specified about the 'deep' nature or the prototypicality of the instrument. WordNet lists some quite diverse types of prototypical instruments: systems, means, implements, hardware, furnishings, equipment, device, transport, container, etc.

2.2 Syntactic Challenges

Languages use different markers to denote instrumentality (Talmy 2001, Spark Jones et al. 1985). These markers include prepositions, postpositions, affixed forms, morphological marks, and derived terms like deverbals. A language uses one or more of these different markers. The choice of the marker depends on context, and on the type of the instrument. These will be discussed further in Section 3.1. Some of these markers are not unique to the instrumental role but can be used for other purposes. The language parser must take into account all these issues for identifying the instrumental roles, in particular the verb structure in the sentence (WordNet and FrameNet are of much interest for English, we have unfortunately much less available resources for the languages we are dealing with).

For example, *with* and *by* are two of the common English prepositions, highly polysemic, that are used to denote the instrumental role in English. Consider the following sentences:

1. I go to the University by bus.
2. The town stands by the canal.

In sentence 1, 'bus' is the instrument, whereas in sentence 2, 'the canal' does not qualify for the role. However in both cases the same preposition is used. The disambiguation is done via the semantic type of the verb: stand being of type locational, the preposition 'by' gets, by default, a locative interpretation.

In Bengali, the instrumental case can be denoted by case markers or by postpositions. The case markers used for the instrumental role (e.g., -e, -te, etc.) are also used for other roles. 'diye' is a common Bengali postposition for the instrumental role. However 'diye' is used in other contexts, e.g., it is the non-finite form of the verb 'to give' and can be used as such.

Thus in order to identify the instruments in a sentence corresponsing to a given action, the parser should have knowledge about the different case markers for the instrumental case, and also appropriate syntactic and semantic rules to eliminate false matches.

2.3 Organizing Responses

Advanced QA, since 2004 (Harabagiu et ali. 2004), have investigated techniques to select a single answer among a set of candidates. This is realized via fusion techniques, applied either on numerical data (Moriceau 2006) or on conceptual data. Most alogorithms are based on probability distribution measures that outline semantic approximations. A set of seven fusion operators have been elaborated, which are relevant for our purpose: contradiction (with some degrees), addition, refinement, agreement, generalization, tendency and irrelevant. In (Webber et ali. 2002), a general survey of the relations that may hold between response candidates is given. It includes relations such as: equivalence, inclusion, aggregation and alternative. These four cases are typically encountered in our corpora.

In general, we get a variety of responses to a given query. These different responses may be various forms of a similar prototypical instrument, or may contain a number of different types of instruments which are alternate means of doing the task. This may be because the task can be performed by multiple means. For example, for cutting glass, the instrument to be used depends on the type of glass, of object to produce (e.g. an aquarium) and on the type of breaking desired. Responses found on the Web are extremely diverse (cutter, diamond cutter, CNC cutter, water jet cutter, diamond point, laser, tungsten carbide wheel, diamond steel small wheel, hammer), and some of them are really unexpected (head, hand, heart, words, songs, etc.). Note that some of these instruments are closely related to each other and are different specialized variants. For example, we have different types of cutters, namely, cutter, diamond cutter, CNC cutter, water jet cutter, etc. These are all variants and a cutter is the most prototypical of these instruments.

In fact we may have several cases, some of which are:

1. We may select several acceptable instruments which are almost equivalent.
2. We may select a set of instruments out of which a few are more prototypical than others, via a domain ontology and conceptual metrics (see below).
3. We may organize sets of instruments for complex actions that require several instruments.

Instruments can also be realized as instrumentalized actions. Finally, they are often also associated with warnings (e.g. precautions to take) and prerequisites that need to be included in the response.

The next stage is to organize the response, at a rather conceptual level, before generating a response. We also want the response to be cooperative, i.e. a response that contains explanations or that reflects in some way the diversity of instruments, instead of selecting just one answer, e.g. the most frequently encountered answer, as is done in basic QA systems. This task consists of:

1. Identifying the most prototypical instruments for actions with single instruments: this is realized via a graph, based on conceptual metrics and a general purpose ontology. The graph may characterize conceptual distances between objects and therefore can allow for the identification of one or more prototypical instruments. It is not practically possible to define an ontology of instruments, since almost everything can be an instrument in a given context. However some basic domain ontologies can be used in the metrics.
2. Identifying chains of instruments for complex actions.
3. Keeping track, in text form, of associated warnings and prerequisites associated with the actions or with the instrument (e.g. use a 3 inch key, but clean it carefully before any use).

3 Cross-linguistic Analysis and Classification

3.1 The conceptual parameters

In (Kawtrakul et al 2006), the morphology and the syntax of marks used to realize instrumental expressions are described. This work is based on the observation of 12 languages from 5 families: French, German, Spanish, Italian, Berber, Arabic, Hindi, Kasmiri, Urdu, Bengali, Thai and Malay. The characteristics of these languages allowed the definition of quite accurate conceptual distinctions, essential in language production. These marks include: prepositions, postpositions, affixes, morphological marks, derived terms like verbs, meant to introduce certain types of instruments in specific contexts.

We outline below the different conceptual distinctions relevant for the NLG of instrumental expressions, explaining how the required knowledge can be found. We then show how these expressions are generated for French, Spanish, German, Hindi, Bengali and Thai.

The following parameters turn out to be essential for generating instrumental expressions in a multilingual perspective:

1. Concrete versus abstract instrument e.g. Thai: duai vs. tam, Bengali: diye vs. dwArA. Among concrete instruments, some distinctions are quite generic: the instrument is a recipient or it is part of the body.
2. Manner-instruments: some instruments are close to manners and are realized as such, e.g. Thai: khian duai muek daeng (write with red ink).
3. Causal instrument and agentive instrument: French: à cause de, agentive instrument: Urdu: ke zariye, vs. causal instrument: Urdu: ki vajah se,
4. Paths and sources as instruments, based on spatial metaphors: 'action is motion, goals are paths, actors are travellers', etc. Among paths, distinctions are made on the medium, and on the channel of transmission in the case of communication: Melalui telefon in Malay. In Bengali suffixes -e and -te denote paths where the agent that does the action has no control over it, whereas diye and dhare involve control. In Thai, source need a special mark, chak.
5. Means of transportation are also treated as a special case, In Bengali kare, -te kare or -ya kare are used: Bengali: *Nouko-ya kare phuketa jAo* (boat-ya kare phuket go: go by boat to Phuket).
6. Orientation: indicating a positive (thanks to, Thai: khop khun) or a negative (Spanish: por culpa de, litt. by the fault of) orientation.
7. Language level: some marks are proper to formal language, in some domains, German: Mittels, Kraft, Anhand.

3.2 Generating multilingual instrumental expressions

The starting point of the generation process are (1) the terms used in the question, mainly the verb, and (2) the list of instruments. Information is extracted from these elements as follows:

1. Ontological data to resolve cases 1 and 5 above.
2. Thematic structure and selectional restrictions of the verb for case 4 (e.g. the verb selects a path)
3. Global orientation (positive or negative) of the response: cases 6 and 7.
4. French has distinct prepositions for manner instruments like à (écrire à l'encre rouge: write with red ink) which can be detected via a simple corpus inspection.
5. Causal and agentive instruments can be measured also via corpus inspection on the rate of occurence as subjects of an action.

Given these different types of data, we show below the decision graphs associated with several languages, for any type of instrument and query.

4 Decision Graphs

As we have mentioned earlier, the marker to use for the instrumental role depends on the language and the context. Let us now present the decision graphs for a few different languages. These decision graph show how markers for the instrumental roles are to be generated for various different kinds of instrumental roles. The decision graphs presented here deal with the most common cases, and a common marker for these cases is given.

These graphs are used roughly as follows when processing Web extracts (from search engines). The verb of the query, and possibly a few arguments are used to query the search engine. Then snippets provided by the search engine (Google or Exalead) as response are traversed, looking for an instrumental expression positioned in a close vincinity (language parameter) of the verb present in the query (or a closely related term). The set of responses is collected and fusion techniques (not presented here) are applied. Once a response is choosen, the question verb is considered again and associated with the instrument to form a correct VP. The graphs below are also used for the choice of the right marker to form that question in NL.

4.1 The Decision Graph for Bengali

The instrumental case in Bengali is syntactically realized through suffixation, postposition or a combination of both. The choice of the suffix and/or postposition depends on the nature of the instrument, its relationship with the action, agent and object, and the orientation of the sentence (positive/negative, assertive/exclamatory, formal/informal etc.). Below, is a decision graph for choosing the appropriate suffix and postposition combination; suffixes are preceded by a hyphen. A finer division requires usage frequency analysis, because often there are preferred markers for certain instrument-verb pairs. Under each category below, we list the most preferred possibilities only.

Path:
— Alternative paths exist: *-e, diYe, dhare*
— No alternatives: *-e*
Means of transport: *-e, kare, -e kare, diYe*
Causal:
— Positive orientation: *-r kripAYe, -r kalyANe*
— Neutral orientation: *-r janya, -r phale*
Agentive:
— Specific Agent: *-ke diYe, -r dwArA*
— Non-specific Agent: *diYe*
Indirect instrument: *-r sAhAyye, -r sahayoge*
Resource Instrument: *-e, diYe, khATiYe*
Direct Instrument:
— Verb is “play” or “beat”: null
— Instrument can *contain* the object: *-e, -e kare*
— Agent has control: *diYe*
— Agent has no control: *-e*
Medium: *-r mAdhyame*
Action instrumentalized: Verb-*e* (non-finite)

4.2 The Decision Graph for Hindi-Urdu

Like Bengali, Hindi also uses a combination of suffixes and postpositions to denote the instrumental case, where the choice of the appropriate combination depends on the relationship between the verb, instrument, agent and the object, and the orientation of the speaker. Note that unlike Bengali where in certain cases the instrumental case does not feature any suffix or postposition marker, in Hindi there is always a non-null postposition.

Path: *se*
Means of transport:
— Animals: *se, pe/para*
— Others: *se, me.n*
Causal:
— Positive orientation: *ke badaulata*
— Neutral orientation: *le lie, ke vAste, ke nAte, ke kAraNa*
Agentive:
— Specific Agent: *ke dwArA, se*
— Non-specific Agent: *se, dwArA*
Indirect instrument: *ke sahAre, ke sAtha, sahita*
Resource Instrument: *dekara, lagAkara, se*
Direct Instrument: *se*
Medium: *ke mAdhyama se, ke zariye*
Action instrumentalized: Verb-*ke/kara* (non-finite)

4.3 The Decision Graph for Spanish

In Spanish simple or complex prepositions are used to mark instrumentality. Distinctions are made between direct and indirect, prototypical and non-prototypical, and positive and negative orientations. The decision graph is shown below.

Path:
— Spatial:*por, a travès de*
— Temporal: *a travès de, con*
— Source: *de* + determiner
Means of transport:
— Means, path: *por*
— Container: *en*
Causal: *a causa de*
Manner: *a, en*
Indirect + abstract instrument:
— Agentive: *mediante, por medio de*
— Other cases: *con, por medio de, mediante*
Direct Instrument:
— Container: *en*
— Focus *con la ayuda de*
— Difficulty to realize: *a base de*
— Lack of prototypicality: *por medio de, mediante*
Orientation:
— Positive: *gracias a*
— Negative: *por culpa de*

4.4 The decision graph for French

French is also a Romance language, as Spanish, but it has a number of differences.

Path:
— Spatial:*par*
— Temporal: *avec*
— Source: *de*
Means of transport:
— Means, path: *par, en*
Causal: *à cause de*
Manner: *à, en*
Indirect + abstract instrument:
— Agentive: *au moyen de*
— Other cases: *avec*
Direct Instrument:
— Container: *en*
— Focus *à l’aide de*
— Lack of prototypicality: *avec*
Orientation:
— Positive: *grâce à*
— Negative: *à cause de*

4.5 The Decision Graph for Thai

In Thai, prepositions are used to indicate the instrumental case. Although the use of prepositions is optional in colloquial Thai, they are nevertheless essential in NLG for unambiguity.

Path:
— Direct:*tam*
— Metaphorical or channel: *thang*
— Source: *chak*
Means of transport: *doi, thang*
Causal: *duai*
Manner: *duai, doi*
Indirect + abstract instrument:
— Agentive: *doi*
— Other cases: *tam*
Direct Instrument:
— Default: *duai*
— Part of the body: *kap*
Positive orientation: *khop khun krap* (Masculine) /*kah* (feminine)

5 Conclusion

In this paper we have outlined a method for cross-lingual question answering for instrumental questions. The system is under develoment at the moment (integrated into a procedural question answering system) and thus evaluations could not be carried out. The decision graphs presented in this paper should suffice for most cases for parsing and generation. However, a coverage testing and some statistics identifying main uses needs to be carried out, if one attempts to use these decision graphs for answer extraction.

Acknowledgements

This work has been made possible via the STIC-Asia cooperation framework.

References

[1] Dorr, B., Olsen, M.B., (1997), *Deriving Verbal and Compositional Verbal Aspect for NLP Applications*, proc. ACL'97, Madrid.

[2] Dorr, B. J., Garman, J., and Weinberg, A., (1995), *From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT*, Machine Translation, 9:3-4, pp.71-100.

[3] Fellbaum, C., (1993), *English Verbs as Semantic Net*, Journal of Lexicography, vol. 6, Oxford University Press.

[4] Harabagiu, S., Lacatusu., F., Strategies for Advanced Question Answering. In Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL, Boston, USA, 2004.

[5] Jackendoff, R., (1990), *Semantic Structures*, MIT Press.

[6] Kawtrakul, A., et alii. (12 authors): A multilingual Analysis of the notion of Instrumentality, 2006. *Proc. EACL workshop on prepositions*. Trento, Italy.

[7] Moriceau, V., Generating Intelligent Numerical Answers in a Question-Answering System. 4th International Natural Language Generation Conference (INLG), p103-110, Sydney (Australia), 15-16 juillet 2006.

[8] Spark-Jones, K., Boguraev, B., A note on a study of cases, research note, Journal of the ACL, dec. 85 now available electronically at www.acl.org.

[9] Talmy L., 2001. *Towards a Cognitive Semantics*, vol. 1 and 2. MIT Press.

[10] Webber, B., Gardent, C., Bos, J., Position statement: Inference in Question Answering. In Proceedings of LREC, 2002.

[11] Wierzbicka, A. (1992), *Semantic Primitives and Semantic Fields*, in A. Lehrer and E.F. Kittay (eds.), Frames, Fields and Contrasts. Hillsdale: Lawrence Erlbaum Associates, pp. 208-227.