

Towards Context-adaptive Utterance Interpretation

Robert Porzel and Iryna Gurevych

European Media Laboratory

Schloss-Wolfsbrunnenweg 33

D-69118 Heidelberg

E-mail: {porzel,gurevych}@eml.org

Abstract

In the tourism domain, a simple question such as “Where is the castle?” may be interpreted solely as a request for the castle’s location. More often, as our data indicate, such a question is used to ask for directions to the specified object. A felicitous response to such a request may depend not only on the questioner’s current location but also on other contextual features, such as the weather, traffic conditions, mode of transportation, and time. This paper describes experimental results supplying factors relevant to such context dependent analysis and a corresponding model that can be employed to increase the conversational abilities of dialogue systems.

1 Introduction

Following Allen et al. (2001), we can differentiate between controlled and conversational dialogue systems. Since controlled and restricted interactions between the user and the system decrease recognition and understanding errors, such systems are reliable enough to be deployed in various real world applications, e.g. transportation or cinema information systems. The more conversational a dialogue system becomes the less predictable are the users’ utterances. Recognition and processing become increasingly difficult and unreliable. Numerous research projects are struggling to overcome the problems arising

with more- or truly conversational dialogue systems.¹

Their goals are more intuitive and conversational natural language interfaces that can someday be used in real world applications. The work described herein is part of that larger undertaking: we view the handling of contextual - and therefore linguistically implicit - information as one of major challenges for understanding conversational utterances in dialogue systems. The paramount importance of context for natural language understanding is frequently noted in the literature, albeit few dialogue systems take extralinguistic contexts into account and perform a corresponding context-dependent analysis of the given utterances at hand.² We differentiate between four different types of contexts that contribute information relevant to natural language understanding and are listed in Table 1. In dialogue systems these knowledge stores are commonly assigned to respective models: the situation model, dialogue model, user model and the domain model, e.g. the ontology.

types of context	content
situational	time, place, etc
discourse	what has been said
interlocutionary	user/system properties
domain	ontological knowledge

Table 1: Contexts and content

¹For example within the DARPA Communicator (www.darpa.mil/ito/research/com/index.html) or the SmartKom (www.smartkom.org) research frameworks.

²See also Porzel and Strube (2000) for an overview of context-adaptiveness in natural language processing (NLP) systems.

In this paper we display findings from data collected in experiments tailored towards identifying and learning contextual factors relevant to understanding a user’s utterance in an uncontrolled dialogue system supplying touristic and spatial information. We, then, show how natural language analysis can employ models that incorporate these contextual factors, resulting in a context-dependent analysis of the given utterances, thereby increasing the conversational capabilities of NLP systems.

Our overall goal is to produce reliable natural language understanding components that increase user satisfaction measures - measurable in an evaluation framework such as PARADISE, described in Walker et al. (2000), or the benchmark- and impurity graphs proposed in Paek (2001) - by applying context sensitive analysis such as described below. We will introduce our model employing examples from the domain of spatial information in Section 2 and give an analysis of the collected data and experiments undertaken so far in Section 3. The resulting model will be described in Sections 4 through 6 followed by concluding remarks in Section 7.

2 Instructions and Descriptions in the Tourist Domain

Several NLP research efforts have adopted the tourism domain as a suitably complex challenge for an intuitive conversational natural language processing system. The resulting prototypes - i.e. mobile tourist information systems that can guide users through cities by providing detailed spatial, architectural and historical information as well as topical information from hotel, entertainment and weather services - have been demonstrated on various occasions.³ Supplying spatial information, specifically spatial instructions and spatial descriptions, constitutes an integral part of the functionality of a mobile tourist

³For example the SmartKom system, Wahlster et al. (2001), and Deep Map system, Malaka and Zipf (2000), have been demonstrated at C-STAR II, Eurospeech 2001 and the International Status Conference “Human Computer Interaction” 2001.

information system.

A **spatial instruction**, e.g. *In order to get to the castle you have to turn right and follow the path until you see the gate tower on your left hand side*, instructs the user how to get from one location/object to a different location along a specific path, which can be for example the shortest, nicest or fastest possible. We regard such an instruction as a felicitous response to a corresponding *instructional* request.

A **spatial description**, e.g. *The Elizabeth Gate is 200 meters to your right* tells the user where a location/object is situated with respect to a reference location/object. We consider this type of response appropriate for a *descriptive* request.

We can, therefore, say that a spatial instruction is an appropriate response to an instructional request and a spatial description, e.g. a localization, constitutes an appropriate response to descriptive request. Responding to a descriptive localization request with a spatial instruction or vice versa, however, does not constitute a felicitous response, but can be deemed a misunderstanding of the questioner’s intention, i.e. an intention misrecognition. In all dialogue systems intention misrecognitions decrease the overall evaluation scores, since they harm the dialogue efficiency metrics, as the user is required to paraphrase the question, resulting in at least one additional user- and system turn. Furthermore user satisfaction measures can also be expected to decrease due to factors as perceived task ease and expected system behavior.⁴

3 The Data

In an initial data collection for constructing adequate language models for automatic speech recognition we asked American native speakers to imagine that they are tourists in Heidelberg, Germany, equipped with a small, personal computer device that understands

⁴Unfortunately dialogue quality metrics are not affected by intention misrecognitions, as they are currently not taken into account in the PARADISE framework.

them and can answer their questions. Among tasks from hotel and restaurant domains subjects also had to ask for directions to specific places. In the corpus we find 128 instances of instructional requests out of a total of roughly 500 requests from 49 subjects. The types and occurrences of these categories in our data are listed in Table 2.

Type <i>Example</i>	# %
(A) How interrogatives, e.g., <i>How do I get to the Fischergasse</i>	38 30%
(B) Where interrogatives, e.g., <i>Where is the Fischergasse</i>	37 29%
(C) What/which interrogatives, e.g., <i>What is the best way to the castle</i>	18 14%
(D) Imperatives, e.g., <i>Give me directions to the castle</i>	12 9.5%
(E) Declaratives, e.g., <i>I want to go to the castle</i>	12 9.5%
(F) Existential interrogatives, e.g., <i>Are there any toilets here</i>	8 6%
(G) Others <i>I do not see any bus stops</i>	3 2%

Table 2: Request types and occurrences

Current parsers that handle both instructional and descriptive requests for spatial information (e.g. the Soup Parser described in Gavalda (1999) within the Deep Map system and the SPIN parser within the SmartKom system (Wahlster et al., 2001)) identify types A, C, D and E as instructional request. This corresponds to a baseline of recognizing roughly 63% of the instructional requests contained in our first data sample as such. Changing the grammars to treat type B and F as instructional request would consequently raise the coverage to 98%. However, **Where interrogatives** do not only occur as requests for spatial instructions but also as requests for spatial descriptions, i.e. localizations.⁵

The problem lies in the fact that the current parser grammars can either interpret

⁵Numerous instances of **Where interrogatives** requesting spatial localizations can be found also in other corpora such as the HCRC Map Task Corpus.

all **Where interrogatives** as descriptive requests or as instructional requests. This implies that both systems can either misinterpret 29% of the instructional request from our initial data as descriptive requests or misinterpret all descriptive request as instructional ones. In short, they lack a systematic way of asking which type of **Where interrogative** might be at hand.⁶

Resulting from these observations we conducted an experiment in which we ask people on the street always the same **Where interrogative**, i.e. *Excuse me, can you tell me where X is*. We varied three factors:

- the goal object, i.e. either the castle, city hall, a specific school, a specific discotheque, a specific cinema, an ATM machine and a specific clothing store, all of which can be either open or closed depending on the time of day, except for the ATM,
- the time of day (i.e. morning, afternoon, evening),
- the proximity to the goal object, i.e. near (less than 5 minutes walk), medium (more than 5 and less than 30 minutes walk) and far (more than 30 minutes walk).

Additionally we kept track of the approximate age group (young, middle, old) and gender of the subjects. In this initial and by no means exhaustive set of contextual features we find that the results of generating decision trees and rules applying a c4.5 learning algorithm as described in Winston (1992), follow our basic intuitions, i.e.:

- if the object is currently closed, e.g. a discotheque or cinema in the morning, almost 90% of the *Where interrogatives* are answered by means of localizations, a few subjects asked whether we actually wanted to go there now or not, and one

⁶As the data discussed herein show a simple approach to employ the system's class-based lexicon to make this decision hinge on the object-type, e.g. BUILDING or STREET, will not suffice to solve the problem completely.

subject gave instructions (the object was the cinema).

- if the object is currently open, e.g. a store or ATM machine in the morning, people responded with instructions, unless - and this we did not expect - the goal object is near and can be localized by means of a reference object that is within line of sight, e.g. *an ATM is in that post office over there*

Looking at the problem of analyzing **Where interrogatives** correctly, we can conclude already that, depending on the combination of at least two contextual features, accessibility and proximity, responses were either instructions, localizations or questions. We feel very confident, however, that by means of introducing additional contextual variations, e.g. dressing the questioner a craftsman carrying buckets of paint, we would get instructions to objects such as discotheques or cinemas even if they happen to be closed at present. The following section will describe how we have chosen to incorporate findings such as the ones described above into the natural language understanding process.

4 Requirements for Contextual Analysis

We have noted above that current natural language understanding systems lack a systematic way of asking, for example, whether a given **Where interrogative** at hand is *construed* as an instructional or a descriptive request.⁷ Speakers habitually rely on situational and other contextual features to enable their interlocutor to resolve such construals appropriately. This is not at all surprising, since conversational dialogues - whether in human-human interaction or human-computer interaction - that occur in a specific context are consequently composed of utterances based upon specific knowledge of that context.

⁷In our terminology saying that the questioner intends the question to be an instructional- or descriptive request is equivalent to it being construed as either one.

In order to capture the diverse kinds of contextual information, studies and experiments of the type described above need to be conducted, so that the individual factors and their influences for a set of additional construal resolutions can be identified and formalized, e.g. via machine learning algorithms. Looking at the domain of spatial information alone we find a multitude of additional decisions that need to be made in order to enable a dialogue system to produce felicitous responses. Next to the instruction versus localization decision, we find construal decisions, such as:

- does the user want to enter, view or just approach the goal object
- does the user want to take the shortest, fastest or nicest path
- does the user intend to walk there, drive or take public transportation

as relevant to answering instructional request felicitously. In many cases, e.g. the ones noted above, construal resolution corresponds to an automatic context-dependent generation of non-elliptical paraphrases in the sense of Ebert et al. (2001). That is, to explicate information that was left linguistically implicit, e.g. to expand an utterance such as *How do I get to the castle* depending on the context into *How do I get to the castle by car on a scenic route*.

These decisions hinge on a number of contextual features much like the instruction versus localization decision discussed above.⁸ In our minds a model resolving the construal of such questions has to satisfy the following demands:

- it has to model the data collected in the experiments, which provide the statistic likelihoods of the relevant factors, for example, the likelihood of a **Where**

⁸Here also ontological factors, e.g. object type and role, additional situational factors, e.g. weather, discourse factors, e.g. referential status, as well as user-related factors, e.g. tourists or business travelers as questioners and their time constraints, constitute significant factors.

interrogative being construed as a descriptive or instructional request, given the accessibility of the goal object,

- it has to be able to combine the probabilistic observations from various heterogeneous knowledge sources, e.g. what if the object is currently accessible, but too far away to reach within a given time period,
- it has to be robust against missing and uncertain information, as these contextual features may not always be observable, e.g. in case specific services of the system such as location modules (GPS) or weather information services are currently offline.

5 Applying the Contextual Analysis

As a first approach we have chosen Bayesian networks employing a generalized version of the variable elimination algorithm described in Cozman (2000) to represent the relations and conditional probabilities observed in the data and to compute the posterior probabilities of the decision at hand. Bayesian networks are extremely well-suited for combining heterogeneous, independent and competing input to produce discrete decisions and can even be regarded as suitable mathematical abstractions over the cognitive processes underlying the way human speakers process natural language (Narayanan and Jurafsky, 1998). The simplest network possible, estimating the likelihood a **Where interrogative** being construed as an instructional or descriptive request, is shown in Figure 1. This network observes whether a **Where interrogative** is at hand, the goal object is open or closed and its proximity to the user (near, medium or far).

We have linked the network to interfaces providing that contextual information. For example within the Deep Map framework, a database called the *Tourist-Heidelberg-Content Base* supplies information about individual objects including their opening and

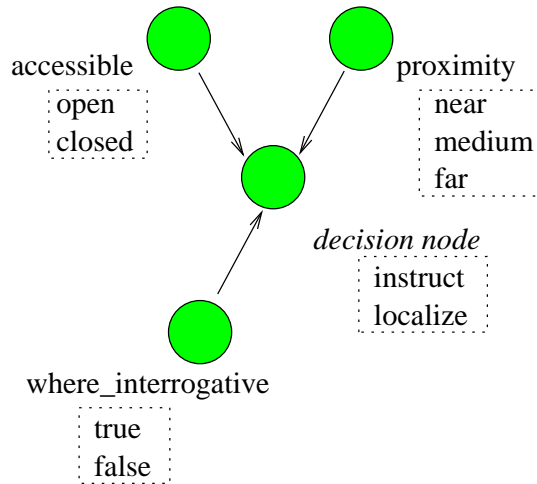


Figure 1: The instruct - localize network

closing times⁹. By default, objects with no opening times, e.g. streets, are considered always to be open. A global positioning system built into the mobile device supplies the current location of the user which is handed to the geographic information system that computes among other things the respective distances and routes to the specific objects. It is important to note that this type of context monitoring is a necessary prerequisite for context-dependent analysis. These technologies enable our model to make dynamic observations of the factors determined as relevant/significant by the data collected.

These observations, captured by the monitoring modules and converted into a context vector defined by a XML schema and the given utterance at hand, i.e. the current parser output, constitute the input into our network¹⁰. The resulting output, consisting of a list of ranked possible construals, e.g. a ranked list of two decisions (e.g. (probability(instruct), 0.64223 p(true | evidence) 0.35777 p(false | evidence))) for a given **Where interrogative**, can then be employed to interpret requests accordingly. That is, the

⁹Additional information extraction agents are able to detect changes in the web and update the local database.

¹⁰We employ a Bayesian interpreter designed for mobile systems called *Embedded Java Bayes*, which can take input as defined in the Bayes Interchange Format (<http://www-2.cs.cmu.edu/javabayes/EBayes/>).

parser output is either converted into the system’s representation of an instructional or localizational request.

As we have seen in Section 3 the current baseline performance results in a misinterpretation rate of 37% of the instructional requests of our initial data set. More specifically, all requests of type B and E, shown in Table 2, will falsely be interpreted as localizational requests and type F is not recognized at all and causes the system to indicate non-understanding. The context-adaptive enhancement described herein, lowers the error rate to 8%, which, in our minds, constitutes a significant improvement. If the ongoing studies indicate that we can treat **Existential Interrogatives** in a similar fashion, this would result in an additional lowering by 6%, leaving only 2% of the initial data set as un-analyzable for the system.

6 The Extended Model

As the data supplies factors related not only to the situational context, but also to the other context stores, such as the discourse, interlocutionary and domain context, we have introduced a way of integrating diverse knowledge sources into graphical models by means of establishing a set of intermediate nodes that form a *decision panel*. In such a panel each weighable *expert node* votes on a common decision, e.g. the posterior probability of a **Where interrogative** being construed as a descriptive or instructional request, as viewed from:

- a situation expert observing, e.g., time, date, proximity, accessibility
- a user expert observing, e.g., interests, transportation, thrift
- a discourse expert observing, e.g., referential status, discourse accessibility
- an ontological expert observing, e.g., object types and object roles

These weights and votes of the experts are, then, combined to achieve resulting posterior

probabilities for the decision at had that equal 1 in their sum.¹¹

In the simple case of a single decision (i.e. instructive versus descriptive requests) we have seen that the model is able to capture the data adequately and behaves accordingly. The full blown model features currently a set of 14 additional discrete decisions and observes over 20 contextual factors. It has not been integrated into the system as the individual data collection for these factors is still ongoing and the integration of some monitoring capabilities, e.g. for the current weather conditions, have just begun. A schematic view of the network with only two decision nodes is given in Appendix 1.

An additional reason for choosing these networks was that even if they become rather complex, they are naturally robust against missing and uncertain data, by relying on the priors in the absence of currently available topical data. This approach, therefore, offers a systematic and robust way of enabling natural language understanding modules to choose among different construals of conversational utterances via context-dependent analysis.

7 Concluding Remarks

In this paper we argue that the implementation of our model can represent and integrate the diverse knowledge sources necessary for context-dependent natural language analysis. As a result it decreases the amount of misinterpretations or intention misrecognitions in conversational dialogue systems, thereby increasing the systems’ performances on user satisfaction evaluations. We expect measurable increases in PARADISE criteria (Walker et al., 2000) such as task ease, expected behavior as well as dialogue metrics, due to a decrease in the number of turns necessary to achieve task completion. An additional experiment based on the Wizard-of-Oz paradigm is currently in the transcription and labeling process to serve as a *gold standard*

¹¹This addition also constitutes an novel systematic way of combining evidences from independent factors in Bayesian networks and keeps the conditional probability tables from becoming exponentially big.

for an evaluation in the framework of Paek (2001).

Since the approach described herein results in a ranked list of possible construals for a given utterance we also defined a threshold for cases where the posterior probabilities can be considered too close. If, for example, the difference of the posterior probabilities of the `instruct - localize` decision is between 0.1 and -0.1, then the system responds by asking the user: *Do you want to go there or know where it is located?*, which incidentally is also a response we found in our initial experiments. This, in turn, would result in more mixed initiative of conversational dialogue systems next to increasing their understanding capabilities and robustness.

Acknowledgments

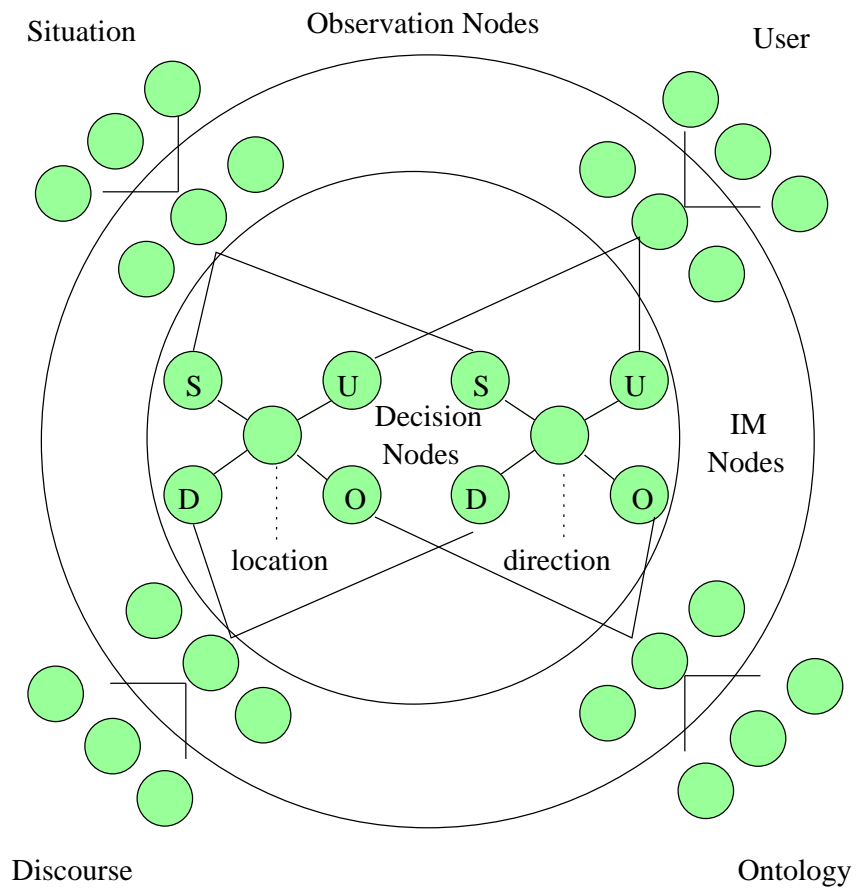
The work described herein was conducted within the SmartKom and EDU projects partly funded by the German ministry of Research and Technology under grant 01IL95I7 and by the Klaus Tschira Foundation.

References

- James F. Allen, Georga Ferguson, and Amanda Stent. 2001. An architecture for more realistic conversational system. In *Proceedings of Intelligent User Interfaces*, pages 1–8, Santa Fe, NM.
- Fabio Cozman. 2000. Generalizing variable elimination in Bayesian networks. In *Proceedings of the IBERAMIA Workshop on Probabilistic Reasoning in Artificial Intelligence*, Sao Paulo, Brazil.
- Christian Ebert, Shalom Lappin, Howard Gregory, and Nicolas Nicolov. 2001. Generating full paraphrases out of fragments in a dialogue interpretation system. In *Proceeding of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 58–67, Aalborg, Denmark.
- Marsal Gavaldà. 1999. SOUP: A parser for real-world spontaneous speechgrowing semantic grammars. In *Proceedings of the 6th International Workshop on Parsing Technologies*, Trento, Italy.
- Rainer Malaka and Alexander Zipf. 2000. Deep Map - challenging IT research in the framework of a tourist information system. In D.R. Fesenmaier, S. Klein, and D. Buhalis, editors, *Information and Communication Technologies in Tourism*, pages 15–27. Springer.
- Srini Narayanan and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. In *Proc. 20th Cognitive Science Society Conference*, pages 84–90. Lawrence Erlbaum Associates.
- Tim Paek. 2001. Empirical methods for evaluating dialog systems. In *Proceeding of the 2nd SIGdial Workshop on Discourse and Dialogue*, pages 100–107, Aalborg, Denmark.
- Robert Porzel and Michael Strube. 2000. Towards context adaptive natural language processing systems. In Manfred Klenner and Henriette Visser, editors, *Proceedings of the International Symposium: Computational Linguistics for the New Millennium*.
- Wolfgang Wahlster, Norbert Reithinger, and Jochen Mueller. 2001. Smartkom: Multimodal communication with a life-like character. In *In Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 1547–1550.
- Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. 2000. Towards developing general model of usability with PARADISE. *Natural Language Engineering*, 6.
- Patrick Henry Winston. 1992. *Artificial Intelligence*. Addison-Wesley.

Appendix 1:

Network Overview



A schematic overview of the network showing the four types of observation nodes (situation, user, discourse and ontology), intermediate nodes (IM) combining several context-specific observation nodes, the expert nodes (labeled *S* for situation, *U* for user, *D* for discourse and *O* for ontology) as well as two decision nodes.