

Contextual Coherence in Natural Language Processing

Robert Porzel and Iryna Gurevych

European Media Laboratory, GmbH
Schloss-Wolfsbrunnenweg 31c
D-69118 Heidelberg, Germany
{porzel,gurevych@eml.org}

Abstract. Controlled and restricted dialogue systems are reliable enough to be deployed in various real world applications. The more conversational a dialogue system becomes, the more difficult and unreliable become recognition and processing. Numerous research projects are struggling to overcome the problems arising with more- or truly conversational dialogue system. We introduce a set of contextual coherence measurements that can improve the reliability of spoken dialogue systems, by including contextual knowledge at various stages in the natural language processing pipeline. We show that, situational knowledge can be successfully employed to resolve pragmatic ambiguities and that it can be coupled with ontological knowledge to resolve semantic ambiguities and to choose among competing automatic speech recognition hypotheses.

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. timetable 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. Research projects are struggling to overcome the problems arising with more- or truly conversational dialogue systems, e.g. Wahlster et al. (2001). 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 complex dialogue systems.

In this paper we report on a set of research issues, solutions and results pertinent to the construction of mobile multi-domain spoken dialogue systems. These systems aim at providing conversational speech interfaces to complex and heterogeneous applications and their domains, e.g. touristic, spatio-geographic

or entertainment information as well as various assistance domains such as planning, electronic communication or electronic commercial transactions. A common feature of the solutions, to be described below, is that they involved the inclusion of extra-linguistic contexts into the natural language processing (NLP) pipeline by applying contextual coherence measurements.

In this work, we will focus on two specific extra-linguistic knowledge stores - namely ontological- and situational knowledge - and introduce the corresponding ontological - and situational coherence measurements.¹ Ontological knowledge, for example, may assert that a bakery is a store and that it has specific properties, such as opening times, specific goods for sale etc. Situational knowledge, on the other hand, may assert that the bakery Seitz is located in a specific street and currently open. Given a user utterance such as: *Is there a bakery somewhere around here?*, we ultimately want an NLP system to understand that the user might want to go there in order to buy something to eat and supply corresponding spatial instructions - to the nearest bakery or other shop depending on what is actually open given the situation at hand - rather than answering the question solely with yes or no. While the ontologies employed herein model more or less static world, conceptual and common-sense knowledge concerning *types* and *roles* (Russell and Norvig, 1995) based on the standard combinations of frame- and description logics, situational knowledge is induced in specific *instances* and highly dynamic states of affairs.

Our overall goal is to produce reliable natural language understanding components that increase dialogue quality metrics,² by applying context sensitive analysis such as described below. After a brief outline of contextual processing in spoken dialogue systems in Sect. 2, we will introduce *situational coherence* and the resulting model, employing data, analyses and results from the domain of spatial information in Sect. 3. We will discuss data, results and model for *ontological coherence* scoring applied in automatic speech recognition and semantic interpretation in Sect. 4. A conclusion on contextual coherence scoring is given in Sect. 5.

2 Contextual Interpretation in NLP

Utterances in dialogues, whether in human-human interaction or human-computer interaction, occur in a specific situation that is composed of different types of contexts. A broad categorization of the types of context relevant to spoken dialogue systems, their content and respective knowledge stores is given in Table 1. Following the common distinction between linguistic and extra-linguistic context³ our first category, i.e. the dialogical context, constitutes the linguistic context, encompassing both co-text as well as intertext.

¹ The role of linguistic- and user-context for NLP is included via discourse-, user- and belief-modeling (LuperFoy 1999, Paris 1993, Narayanan 1997).

² Measurable in the PARADISE evaluation framework (Walker et al. 2000).

³ All extra-linguistic contexts are also often referred to as the *situational context* (Connolly, 2001). however, we adopt a finer categorization thereof.

Table 1. Contexts, content and knowledge sources

types of context	content	knowledge store
dialogical context	what has been said by whom	dialogue model
situational context	time, place, etc	situation model
interlocutionary context	properties of the interlocutors	user model
domain context	world/conceptual knowledge	ontology

In linguistics the study of the relations between linguistic phenomena and aspects of the context of language use is called *pragmatics*. Any theoretical or computational model dealing with reference resolution, e.g. anaphora- or bridging resolution, spatial- or temporal deixis, or non-literal meanings, requires taking the properties of the context into account. In current spoken dialogue systems *contextual interpretation* follows *semantic interpretation*, which follows automatic speech recognition (ASR) (additionally fused with other modality-specific information). That is, the modality-specific signals, (e.g. speech or gesture) are transferred into graphical representations (e.g. word- or gesture graphs) and then fused and mapped onto some meaning representation followed by contextual interpretation (Allen, 1987). Computationally, this implies that context-independent graphical and semantic representations can be computed and the context-dependent contributions are associated with the semantic interpretation thereafter, resulting in the final representation.

This so-called *modular* view supports a distinct study of meaning (corresponding to the semantic representation) without having to muck around in the mirky waters of language use. This view is supported by the claim that some semantic constraints seem to exist independent of context. In this work we propose a different view that also allows for context-independent constraints, but offers a less modular point of view of contextual interpretation. We will show that, given the notion of context introduced above, contextual analysis can be employed already at the level of speech recognition, during semantic interpretation and, of course, thereafter. The central claim is being made, that - as in human processing - contextual knowledge can be used successfully in a computational framework in all processing stages.⁴ While most research in linguistics has consequently departed from this view, most computational approaches still feature a modular pipeline architecture in that respect.

In linguistics utterances which are context-dependent are called *indexical* utterances (Bunt, 2000). Computationally they exhibit a difference in their semantic and final representation. Indexical utterances are - by virtue of the pervasiveness of contextual knowledge - the norm in discourse, with linguistic estimations

⁴ In recent times the so-called *modular* theory of cognition (Fodor, 1983) has been abandoned more or less completely. The so-called *new look* or modern cognitivist positions hold that nearly all cognitive processes are interconnected, and freely exchange information; e.g. influences of semantic and pragmatic features have been shown to arise already at the level of phonological processing (Bergen, 2001).

of declarative non-indexical utterances around 10% (Barr-Hillel, 1954). Without contextual knowledge utterances, or fragments thereof, become susceptible of interpretation in more than one way. Computer languages are designed to avoid anaphoric, syntactic, semantic and pragmatic ambiguity, but human languages seem to be riddled with situations where the listener has to choose between multiple interpretations. In these cases we say that the listener performs *pragmatic analysis*; corresponding to contextual interpretation on the computational side. For human beings the process of resolution is often unconscious, to the point that it is sometimes difficult even to recognize that there ever was any ambiguity.

The phenomenon that this process of resolution, frequently goes unnoticed is due to the fact that in many cases the ambiguity is only perceived if the contextual factors that allowed the listener interpret the utterance unambiguously are missing. For example, if shared ontological and situation-specific knowledge provided information that was elided in the utterance. These utterances/texts, therefore, become ambiguous only after they have been stripped of discourse-, situation-, domain- and speaker-context, and, for example, appeared as a text(-fragment) in a linguistics textbook. The problem for computational linguistics originates at least partially in the fact language understanding has to make do with exactly such a contextually and pragmatically impoverished input.

3 Situational Coherence

In this section we display findings from experiments tailored towards identifying and learning contextual factors relevant to understanding a user's utterance in an uncontrolled dialogue system. That system supplies touristic and spatial information (Porzel and Strube,). In this data we find many instances of phenomena usually labeled as *pragmatic ambiguity*. In our view these examples constitute *bona fide* cases for contextual interpretation after phonological and semantic processing has been concluded. We show how natural language analysis can employ models that incorporate specific situational factors, resulting in a context-dependent analysis of the given utterances, thereby increasing the conversational capabilities of dialogue systems.

Several NLP research efforts have adopted the tourism domain as a suitably complex challenge for an intuitive conversational natural language processing system (Johnston et al. 2002, Wahlster et al. 2001). Supplying spatial information, specifically spatial instructions and spatial descriptions, constitutes an integral part of the functionality of a mobile tourist information system. We regard 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*" - as a felicitous response to a corresponding *instructional* request. A **spatial description** - e.g. "*The Cinema Gloria is near the marketplace on the Hauptstrasse*" - is 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 a descriptive request. Responding with one to

the other 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 additional dialogue turns. Furthermore, satisfaction measures decrease along with perceived task ease and expected system behavior.⁵

The Data: In an initial data collection (Porzel and Gurevych, 2002) 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 are in Table 2.

Table 2. Request types and occurrences

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, e.g., <i>I do not see any bus stops</i>	3	2%

While handling both instructional and descriptive requests for spatial information our parsers 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.⁶ The problem is that the current parser grammars either interpret 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.⁷

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 logged several factors:

⁵ Unfortunately in PARADISE dialogue quality metrics are not effected by intention misrecognitions, as they are not taken into account (Walker et al. 2000).

⁶ Numerous instances of **Where interrogatives** requesting spatial localizations can be found also in other corpora such as the HCRC Map Task Corpus.

⁷ As 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.

- the goal object, i.e. either the castle, city hall, a specific school, a specific discotheque, a specific cinema, a bank (ATM) and a specific clothing store, all of which can be either open or closed depending on the time of day,
- the time of day (i.e. morning, afternoon, evening),
- the proximity to the goal object, i.e. near (< 5 minutes walk), medium (5 - 30 minutes walk) and far (> 30 minutes walk). - additionally we kept track of the approximate age group (young, middle, old) and gender of the subjects.

In this set of contextual features we find that the results of generating decision trees and rules applying a c4.5 learning algorithm (Winston, 1992), show that:

- 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, and one subject gave instructions.

- 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.

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. The following sections will describe how we have chosen to incorporate findings such as the ones described above into the natural language understanding process.

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. 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.

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. 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 requests felicitously. In many cases, e.g. the ones noted above, construal resolution corresponds to an automatic context-dependent generation of 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 a 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 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.

Applying the Contextual Analysis: As a first approach we have chosen Belief- or *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 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 of a **Where interrogative** being construed as an instructional or descriptive request, needs only three *observation nodes*. These nodes observe whether a **Where interrogative** is at hand, the goal object is open or closed and its proximity to the user. The single *decision node* - whether a spatial location or instruction constitutes an appropriate response - is connected to the three observation nodes.

We have linked the network to interfaces providing that contextual information. For example within the SmartKom framework (Wahlster et al. 2001), a database called the *Tourist-Heidelberg-Content Base* supplies information about individual objects including their opening and closing times. A global positioning system built into the mobile device supplies the current location of the user. This is handed to the geographic information system that computes 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

⁸ Here 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.

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 representation, and the given utterance at hand, i.e. the parser output, constitute the input into our belief network. The resulting output constitutes a measurement of the *situational coherence* of the possible alternative readings. In other words it represents a list of ranked construals, e.g. a ranked list of two decisions for a given **Where interrogative** with their corresponding situational coherence scores (e.g. (probability(instruct), 0.64223 p(true | evidence) 0.35777 p(false | evidence))). This can then be employed to interpret requests accordingly, i.e., the parser output is either converted into the system's representation of an instructional or localizational request.

Results: As we have seen 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, 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 additional data 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 unanalyzable for the system.

4 Ontological Coherence

As we have seen above one of the fundamental issues concerning pragmatic ambiguity, is to enable dialogue systems to pick the most appropriate reading given the contextual factors at hand. This is equally true for ambiguities that arise during semantic interpretation and automatic speech recognition.

4.1 Speech Recognition Ambiguities: N-Best Lists

A common phenomena found in different fields of NLP, e.g. automatic speech recognition, information retrieval or question answering, is that current processing techniques seem to hit a ceiling of performance. In ASR systems have progressed to a level where they are close to extracting as much information as possible from the acoustic stream. Some context-dependent features have been added to handle dialectal- and speaker-adaptation and dynamic lexica, to handle novel input (Rapp et al. 2000). However, neither ontological nor situational knowledge is taken into account, which leaves the known problem of dealing with phonetically indistinguishable input, unresolved. The classic example in the community is, that a large vocabulary speech recognition (LVSR) system, as needed for more conversational dialogue systems, could hardly differentiate between homonymic utterances such as: “*it is hard to wreck a nice beach*” and “*it is hard to recognize speech*”. Humans on the other hand *hear* either one or the other depending on the context.

Today's LVSR systems rarely feature simple one-best hypothesis as interface between ASR and NLU. While that may suffice for restricted dialogue systems, most systems either operate on n-best-lists as ASR output or convert ASR word graphs (Oerder and Ney 1993) into n-best lists, given the distribution of acoustic and language model scores (Schwartz 1990). In our data a user expressed in Example (1) the wish to see a specific city map again, leading to the top two speech recognition hypotheses (1a,1b). Annotators found that Example (1a) constituted a pretty much well formed representation of the utterance whereas Example (1b) constituted an inadequate representation thereof:

- (1) *Ich würde die Karte gerne wiedersehen*
 I would the map like to see again
- (1a) - *Ich würde die Karte eine wieder sehen*
 - I would the map one again see
- (1b) - *Ich würde die Karte eine Wiedersehen*
 - I would the map one Good Bye

Facing multiple representations of a single utterance consequently poses the question which of the different hypotheses most likely corresponds to the user's utterance. Several ways of solving this problem have been proposed and implemented in various systems, i.e. to use scores provided by the ASR system, i.e. acoustic and language model probabilities or to use scores provided by the natural language understanding and discourse modeling components, c.f. Litman et al. (1999).

We claim that contextual extra-linguistic knowledge can as well be used at this point to provide further information and to help in solving this task, especially in those cases where ASR and semantic scores fail. In the following we will report on the experimental setup and evaluations of this claim, thereby introducing the central notion of *ontological coherence*.

The Data: An initial experiment was reported in Gurevych et al. (2002) where we tested, whether or not human annotators could reliably classify 2300 speech recognition hypotheses (SRH) in terms of their ontological coherence, i.e. whether or not a given hypothesis constitutes an internally coherent utterance. On an additional corpus of 1400 hypotheses we showed in recent experiments that annotators could also reliably (>94%) identify the best hypothesis, given a transcribed utterance and the corresponding SRHs choices.

Requirements and Application: The corresponding contextual analysis, then, needs to provide a coherence score automatically, that can be employed by any NLU system to select the best hypothesis from the N-best list independently or in conjunction with acoustic or statistical scores. We employ the ONTOSCORE system described in Gurevych et al (2003): Given a frame- and description logic-based ontology - e.g. a semantics as defined in oil-rdfs, daml+oil, or owl -⁹ we map *words* to *concepts* and compute the average path-length of the shortest

⁹ See www.daml.org or www.w3c.org/rdf.

graph found connecting all concepts excluding *isa* relations in the individual path-length measures, that we fitted with a conceptual context addition.

Results: Using the SmartKom system and its pre-existing ontology, ONTOSCORE correctly assigns the highest score to over 84% of the best hypothesis as defined in the merged human *gold standard* (baseline 63.91%). This coherence measuring method has, therefore, been shown to exhibit much greater than baseline-performance in an additional task and performs better or equal compared to the alternative scoring methods.

4.2 Semantic Interpretation and Construal

In much the same way ontological coherence can be employed to disambiguate between multiple representations of a user's input. We will show how it can serve to assist in semantic interpretation, i.e. in resolving semantic construal, that underlies many non-syntactic phenomena involving *unconventional* meaning (Langacker, 1998). Employing simple examples from our tourism domain:

- (2) Goethe often visited the historical museum.
- (3) The Palatine museum was moved to a new location in 1951.
- (4) The apothecary museum was renovated in 1983.
- (5) In 1994 the museum bought a new Matisse.

we find four instances of noun phrases featuring the word *museum* as an argument of four main verbs: *visited*, *moved*, *renovated* and *bought*. Linguistic analyses may vary in their classifications, however, commonly, Example (2) would be regarded as pretty conventional, Examples (3) and (4) as polysemous and Example (5) as metonymical language use with respect to the word *museum*. In many cases we find lexical ambiguities as for *kommen* in the SRH Example (6) and Example (7):

- (6) *was für Spielfilme kommen heute Abend*
 what for films come today evening
- (7) *wie kann ich mir zur Schloss kommen*
 how can I me to castle come

Due to the persuasiveness of construal in natural language, a formal model thereof as well as an account of its mechanisms, constitutes an important part of any approach to natural language understanding.

Data: As shown in Poesio (2002) about 50% of all noun phrases in their corpora are discourse-new, anaphoric noun phrases make up 30% of their data. The remaining 20% are made up by so-called *associative* expressions. In an additional experiment annotators labeled correct word-senses for all cases (1415 markables) of multiple word to concept mappings. For example, in SRHs containing forms of the verb *kommen* (to come/showing on), a decision had to be made whether it is a `MotionDirectedTransliterated`, as in Example (6) or

`WatchPerceptualProcess` as in Example (7) or undecidable - which was only the case in non-best SRHs (see Sect. 4.1).

Requirements and Application: Previous work on resolving ambiguities, metonymic language use and other types of *associative* meaning (Poesio, 2002) exists that also employ various kinds of hierarchical knowledge bases, showing promising results in domain-specific settings. The actual content of an ontology depends on the specific modeling choices made while constructing the ontology. Due to individual differences and even internally heterogeneous modeling choices, we need flexible algorithms for retrieving the appropriate information from the knowledge base, unlike those employed in previous approaches.¹⁰ Additionally, the semantic web projects bring forth a multitude of external ontologies, whose modeling choices need not be known beforehand. Yet, if dialogue systems intent to profit from this undertaking, they will need to be able to extract the necessary information without knowing the specific modeling choices. As proposed in Porzel and Bryant (2003) an extra-linguistic knowledge store - an ontology - can be employed to find sets of alternative readings by searching the conceptual graph in ways as permissive as radial categories suggest. These ontological substitutions constitute an addition to ambiguity mappings from the lexicon.

This has been interfaced to the ontological coherence scoring application, i.e. ONTOSCORE, to calculate how often contextual coherence picks the appropriate reading. In order to aid semantic interpretation by means of contextual knowledge we can apply the same algorithm employed to score sets of speech recognition hypotheses for scoring different potential trigger - target pairings with respect to their ontological coherence. For metonymy or bridging resolution, however, an initial processing step is needed to find sets of possible pairings, i.e. candidates that are potentially more ontologically coherent.

Results: As a result of measuring the ontological coherence of the conceptual representations we get a corresponding ranking for the alternative readings. Looking at the case of *kommen* as *showing (on TV)* versus *coming (to/from)*, given a pre-existing ontology we find 85 occurrences of this ambiguity in which the contextually enhanced ONTOSCORE picked in the correct reading in 72 cases, and not in 2 cases, and mixed in all 11 undecidable cases, which were not in the best SRH set. Baseline, given the majority distribution, was 56.5%.

The inclusion of such contextual interpretation during and before semantic interpretation can enable natural language understanding systems to become more conversational without losing the reliability of restricted dialogue systems. Our work on combining situational coherence measures as reported in Sect. 3 with ontological coherence and discourse coherence has already shown an increase in performance on multiple tasks. We are, therefore, strongly encouraged by the results that this approach constitutes a suitable path towards making natural language processing more robust and human-like.

¹⁰ While it is certainly feasible to limit bridging or metonymy resolution to a pre-defined set of ontological relations, such as *has-part* relations, if the ontology was especially crafted for that type of resolution (Poesio, 2002).

5 Contextual Coherence

Extra-linguistic factors relate not only to the situational context, but also to the other context stores, such as the discourse, interlocutionary and ontological context. For an integrated model of common sense-based contextual coherence, we have introduced a way of integrating diverse knowledge sources into belief networks 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, - or of the *museum* sense - 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 hand 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 situational factors as introduced in Sect. 3 as well as ontological factors as input to the contextually enhanced ON-TOSCORE system. It's integration into the SMARTKOM can be extended as collected data and monitoring capabilities, e.g. for the current weather conditions, become available. 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 resolve different construals of conversational utterances via context-dependent analysis.

6 Conclusion

In this work we focus primarily on contextual interpretation that makes NLP applications more reliable and conversational. We rely on two primary contextual knowledge stores: world- and situational-knowledge captured, herein, by means of formal ontologies and belief networks. We argue that the addition of extra-linguistic knowledge, i.e. situational and ontological knowledge, can represent and integrate the diverse knowledge sources necessary for context-dependent natural language analysis. As a result we showed decreases in the amount of misinterpretations or intention misrecognitions applied at three stages in the processing pipeline of an implemented dialogue system. The application, thereby, increases the systems' performances on features crucial to user satisfaction evaluations, leading to measurable increases in evaluation criteria such as task ease,

¹¹ This addition offers a systematic way of combining evidences from independent factors in belief networks and shrinks the conditional probability tables.

expected behavior as well as dialogue metrics, due to a decrease in the number of turns necessary to achieve task completion.

We introduce contextual coherence measurements, i.e. the output of situational- and ontological coherence measurements. We employed these to find best-speech recognition hypotheses in n-best lists, rank ambiguous, polysemous and metonymical readings and resolve pragmatic ambiguities via inferences from knowledge- and belief-models - based on *common sense* knowledge. The general model introduced shows how such scores reflect a set of additional *common sense* constraints that can be applied as semantic- and pragmatic constraints next to phonological or morpho-syntactic constraints. We can, for example, consider the case of *where questions* as cases where all syntactic and semantic constraints are perfectly satisfied by a proposed filler, while pragmatic constraints concerning the accessibility of the goal object can be violated depending on the situational context.

Since the approach described herein results in ranked lists of possible construals for a given utterance, we can define a threshold for cases where the resulting scores 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, the system can respond 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. This work has been partially funded by the German Federal Ministry of Research and Technology (BMBF) as part of the SmartKom project under Grant 01 IL 905C/0 and by the Klaus Tschira Foundation.

References

- Allen, James and Georga Ferguson and Amanda Stent: An architecture for more realistic conversational system. In Proc. of Intelligent User Interfaces. (2001) 1–8
- Allen, James: Natural Language Understanding. Ben. Cummings (1987)
- Barr-Hillel, Y.: Logical Syntax and Semantics. Language, 20 (1954) 230–237
- Bergen, Ben: Of sound, mind, and body: neural explanations for non-categorical phonology. PhD Thesis UC Berkeley (2001)
- Bunt, Harry: Dialogue pragmatics and context specification. Computational Pragmatics, Abduction, Belief and Context; Studies in Computational Pragmatics. John Benjamins (2000) 81–150
- Connolly, John: Context in the Study of Human Languages and Computer Programming Languages: A Comparison. Modeling and Using Context, Springer, LNCS (2001) 116–128
- Cozman, Fabio: Generalizing Variable Elimination in Bayesian Networks. In Proc. of the IBERAMIA Workshop on Probabilistic Reasoning in Artificial Intelligence (2000)
- Ebert, Christian and Shalom Lappin and Howard Gregory and Nicolas Nicolov: Generating Full Paraphrases out of Fragments in a Dialogue Interpretation System. In Proc. 2nd SIGdial Workshop (2001) 58–67

- Fodor, Jerry: *The Modularity of Mind*. MIT Press (1983)
- Gurevych Iryna, Rainer Malaka, Robert Porzel and Hans-Peter Zorn: Semantic Coherence Scoring Using an Ontology. In Proc. of the HLT-NAACL Conference (2003)
- Gurevych Iryna and Michael Strube and Robert Porzel: Automatic Classification of Speech Recognition Hypothesis. In Proc. of the 3rd SIGdial Workshop (2002) 90–95
- Johnston, Michael and Bangalore, Srinivas and Vasireddy, Gunaranjan and Stent, Amanda and Ehlen, Patrick and Walker, Marilyn and Whittaker, Steve and Maloor, Preetam: *MATCH: An Architecture for Multimodal Dialogue Systems*. Proceedings of ACL'02. (2002) 376–383
- Langacker, Ronald: *Conceptualization, symbolization and grammar. Cognitive and functional approaches to language structure*. Laurence Erlbaum (1998)
- Litman, D. and Walker, M. and Kearns, M.: Automatic Detection of Poor speech recognition at the dialogue Level In Proc. of ACL'99 (1999)
- LuperFoy, Susann: The representation of multimodal user interface dialogues using discourse pegs. In Proceedings of ACL'92 (1992) 22-31
- Narayanan, Srinu and Daniel Jurafsky: Bayesian Models of Human Sentence Processing. In Proc. 20th Cognitive Science Society Conference (1998) 84–90
- Narayanan, Srinivas: *KARMA: Knowledge-Based Active Representations for Metaphor and Aspect*. PhD Thesis. UC Berkeley, (1997)
- Oerder, Martin and Hermann Ney: Word Graphs: An Efficient Interface Between Continuous-Speech Recognition and Language Understanding. In Proc. of ICASSP Volume 2 (1993) 119–122
- Paris, Cécile L.: *User Modeling in Text Generation*. Pinter (1993)
- Poesio, Massimo: Scaling up Anaphora Resolution. In Proc. of the 1st Workshop on Scalable Natural Language Understanding (2002) 3–11
- Porzel, Robert and Iryna Gurevych: Towards Context-adaptive Utterance Interpretation. In Proc. of the 3rd SIGdial Workshop (2002) 90–95
- Porzel, Robert and John Bryant: Employing the Embodied Construction Grammar Formalism for Knowledge Representation: The case of construal resolution. In Proc. of 8th International Conference on Cognitive Linguistic (2003)
- Rapp, Stefan and Torge, Sunna and Goronzy, Silke and Kompe, Ralf: Dynamic speech interfaces In Proc. of 14th ECAI WS-AIMS (2000)
- Russell, Stuart J. and Norvig, Peter: *Artificial Intelligence. A Modern Approach*. Prentice Hall (1995)
- Robert Porzel and Michael Strube: Towards Context Adaptive Natural Language Processing Systems. In Proc. of the International Symposium: Computational Linguistics for the New Millenium (2000)
- R. Schwartz and Y. Chow: The N-best Algorithm: an Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses. In Proc. of ICASSP'90 (1990)
- Wolfgang Wahlster and Norbert Reithinger and Anselm Blocher: *Smartkom: Multimodal Communication with a Life-Like Character*. Proc. of the 7th Eurospeech (2001) 1547–1550
- Marilyn A. Walker and Candace A. Kamm and Diane J. Litman: Towards Developing General Model of Usability with PARADISE. *Natural Language Engineering* 6 (2000)
- Patrick Henry Winston: *Artificial Intelligence*. Addison-Wesley (1992)