Evaluation of the KomParse Conversational Non-Player Characters in a Commercial Virtual World

Tina Klüwer, Feiyu Xu, Peter Adolphs and Hans Uszkoreit

German Research Center for Artificial Intelligence (DFKI) GmbH
Language Technology Lab, Germany
{tina.kluewer,feiyu,peter.adolphs,uszkoreit}@dfki.de

Abstract

The paper describes the evaluation of the KomParse system. KomParse is a dialogue system embedded in a 3-D massive multiplayer online game, allowing conversations between non player characters (NPCs) and game users. In a field test with game users, the system was evaluated with respect to acceptability and usability of the overall system as well as task completion, dialogue control and efficiency of three conversational tasks. Furthermore, subjective feedback has been collected for evaluating the single communication components of the system such as natural language understanding. The results are very satisfying and promising. In general, both the usability and acceptability tests show that the tested NPC is useful and well-accepted by the users. Even if the NPC does not always understand the users well and expresses things unexpected, he could still provide appropriate responses to help users to solve their problems or entertain them.

Keywords: virtual worlds, conversational agents, dialogue system evaluation

1. Introduction

Commercial online games and social virtual worlds such as *Second Life* or *World of Warcraft* attract an enormous amount of users and open new perspectives in human-machine interaction. In addition to the game players, non-player characters (NPCs) are essential for some game types, in order to support the game plot and to create an immersive environment. Several research systems provide stand-alone embodied conversational agents (ECA), for example, the Companion project (Cavazza et al., 2010), Justina (Kenny et al., 2008) or Max (Kopp et al., 2005), and recently even virtual characters in 3D-worlds with conversational capabilities are developed, e.g. the NICE fairy-tale game (Gustafson et al., 2005) and the Mission Rehearsal Exercise System (Hill et al., 2003). However, in all these systems, the capabilities of dealing with unrestricted natural language input and dialogues with users are still very limited.

The *KomParse* system is a natural-language dialogue system embedded in the commercial three-dimensional virtual world *Twinity*[1]. *KomParse* realizes a flexible and hybrid approach to dialogue processing combining knowledge-intensive domain-specific question answering, task-specific and domain-specific dialogue with some small talk functionalities. One NPC in *KomParse* is a barkeeper who recommends and sells cocktails to the users and can entertain his guests by providing trivia-type information about celebrities.

Since it is essential for our industry partner *Metaversum*[2] to know the usability and acceptability of the NPCs in their platform, we conduct an evaluation of the *KomParse* system from the end user's point of view. The evaluation method applied here is a further adaptation of the field test and the evaluation methods realized in (Uszkoreit et al., 2007) and measures subjective usability, acceptability, naturalness and efficiency/effectiveness of the system. The overall evaluation result is very satisfying and promising.

Section 2 describes the relevant technologies developed in the pilot system. In section 3, we give an overview of the field test organization. Section 4 defines our evaluation approach and shows the evaluation results. In section 5 we close off with a short conclusion.

2. NPC as Conversational Agent in a Virtual World

Our NPCs provide various services pertinent to virtual worlds through conversation with game users. Therefore, the requirements for the system include e.g. the ability of producing pragmatically adequate responses as well as the interpretation of the user's utterances and the spatial situation. Furthermore, NPC systems always need to preserve the robustness of real-world applications. Being confronted with these challenges, the *KomParse* system builds on theoretical and experimental insights from linguistic pragmatics, uses novel techniques from computational linguistics and combines them with robust baseline technologies to provide NPCs which possess the necessary intelligence to act and talk in a virtual world. Furthermore, the utilization of detailed knowledge representations enables semantic search and inference. We use the virtual world *Twinity* as a testbed for our system, which simulates 3D models of cities such as Berlin, Singapore, London and New York. Users can log into *Twinity* where they can meet other users and communicate with them using the integrated text chat function. They can style their virtual appearance, rent or buy their own flats and decorate them as to their preferences and tastes.

[1] http://www.twinity.com
[2] http://www.metaversum.com

2.1. Barkeeper NPC

In this paper, we present the evaluation of Hank Slender, a barkeeper NPC. Hank Slender owns a bar in the virtual Berlin, where he sells cocktails to the users and entertains his guests with conversations about pop stars, movie actors and other celebrities as well as the relations between these people. He is able to access several knowledge bases and to handle questions about a domain called the "gossip domain". The barkeeper agent has to be able to offer and sell virtual drinks and to understand the wishes of the users. Moreover, the main role of the barkeeper is a conversation companion, which allows us to study and model small-talk strategies.

Figure 1: Barkeeper NPC in interaction with a human customer

3. The KomParse Architecture

The NPC consists of an "avatar", which is the physical appearance of the NPC in the virtual world, and the conversational agent which provides the dialogue system, the control logic of the agent's behavior. The agent is hosted by a multi-client, multi-threaded server written in Java, whereas the NPC's avatar is realized by a modified *Twinity* client. It sends all in-game events relevant to our system to the server and translates the commands sent by the server into *Twinity*-specific actions. The system can be naturally extended to other platforms. In addition to the *Twinity* platform, we realized a web interface. In the following, we will give a brief description of the key technologies.

3.1. Knowledge Representation and Acquisition

We use Semantic Web technology for building the knowledge base for our agents. This knowledge base is in particular important for the barkeeper scenario, where the NPC has to be able to interpret user statements and queries about the world. We created a biographical ontology, the "gossip ontology", defining biographical and career-specific concepts for people. This ontology is accompanied by a huge data resource of celebrities by combining i) existing Semantic Web resources, ii) Semantic Web resources which have been created from semi-structured textual data in the Web and iii) with relation information extracted from free natural language texts (Xu et al., 2007; Xu et al., 2010). This resource covers nearly 600,000 persons and relations between them such as family relationships, marriages and professional relations.

3.2. Natural Language Understanding

Natural Language Understanding focuses on the recognition of the dialogue acts expressed by the user's utterances. Dialogue acts are verbal or nonverbal actions that incorporate participant's intentions and represent the functional level of an utterance, such as a greeting, a request or a statement. The dialogue act recognition in the described system is a hybrid component, which uses patterns, statistical methods and rules to detect the appropriate intention belonging to the input. Patterns contain assignments of regular expressions matching the input to a special dialogue act. This is useful for highly predictable, idiosyncratic or one word input such as "hi" or "you are welcome". The rules and the statistical model use a cue-based method for dialogue act classification with various features from multi-level knowledge sources. Features include information extracted from the incoming utterance as well as information about the previous dialogue. In contrast to existing systems using bag-of words representations of the utterances (Crook et al., 2009), lexical features, i.e. words, single markers such as punctuation (Verbree et al., 2006) or combinations of various features (Stolcke et al., 2000), the results of our interpretation component are based on syntactic relations and a minimal dialogue context (Klüwer et al., 2010b). Relations are extracted from the utterance through a predicate argument analysis. Each utterance is parsed with a predicate argument parser and annotated with syntactic relations organized according to PropBank (Palmer et al., 2005) containing the following features: Predicate, Subject, Objects, Negation, Modifiers, Copula Complements. The parser is a rule-based predicate argument parser. The rules utilized by the parser describe subtrees of dependency structures in XML by means of relevant grammatical functions. The rules deliver raw predicate argument structures, in which the detected arguments and the verb serve as hooks for further information lookup in the input. In a second step all modifier arguments existing in the structure are recursively acquired. The same is done for modal arguments as well as modifiers of the arguments such as determiners, adjectives or embedded prepositions. The added context features are the last preceding dialogue act, a value for equality between the last preceding topic and the actual topic and the sentence mood. Our statistical model is generated by a Bayesian classifier on the basis of the corpus annotated with dialogue acts and relational information. The model is integrated into the interpretation component and deals with input which does not match a pattern or a rule. The dialogue act recognition process is described in detail in (Klüwer et al., 2010b).

3.3. Dialogue Management

The core of the dialogue system is a finite-state graph. The graph determines the next action according to the results of the natural language understanding component, database queries or other environment parameters. The decision to use a finite-state approach was made in order to find a balanced agreement between complexity, ease-to-use, robustness and flexibility regarding the demands of the real-world massive multiplayer online game, which carries a need of high robustness. However, we have integrated several ad-

ditional components, which enable further flexibility of the system, moving more into the direction of an information state update approach (Larsson and Traum, 2000). The dialogue system manages a handful of variables which form a type of information state. Moreover, frames are used to handle the task model of the dialogue system. They store the objects the users have discussed and bought so far (e.g. cocktails), together with the relations belonging to these objects. This information is taken from the knowledge bases at runtime. Thus, the finite-state graph is made very flexible. Nevertheless, the positive characteristics of the finite-state approach such as robustness are maintained.
The used graph formalism belongs to the finite state graph toolkit "SceneMaker" (Klesen et al., 2003; Gebhard et al., 2003). SceneMaker offers a very clear user interface for authoring of graphs incorporating Harel's state charts and a fast graph interpreter including a Java API.

3.4. Question Answering

Interaction data show that users ask NPCs about all sorts of things, as for instance their personal information and their doing, the surrounding, available choices in selling situations, or the state of affairs in the world. Question-answering (QA) technology forms a central pillar for handling many of these information-seeking requests. Following a general design paradigm for building our system, we use detailed knowledge representations for solving this task (Adolphs et al., 2010). On top of this knowledge base, we built a question answering module, which allows users to access information in a smooth natural language dialogue.
The QA module is embedded into our input processing pipeline: each user input is first linguistically analysed and interpreted with respect to the current dialogue context. The result of the input interpretation is a dialogue act and a more fine-grained semantic representation of the user input in case of information-seeking requests. This question semantics is turned into a query to the knowledge base. The query results are then turned into an abstract representation of the answer and spelt out by a natural-language generator (Adolphs et al., 2010). Our module is able to provide the QA functionality in a smooth and connected dialogue. A dialogue memory and the dialogue state allow the module to resolve pronouns from previous entity mentions as well as to pose and react to clarification questions in case of a possible misunderstanding.
In order to achieve robustness and accuracy for question processing, we take a hybrid approach by combining two strategies. One is a fuzzy pattern matching algorithm which utilises regular lexico-syntactic patterns based on surface strings and recognised named entities, while another makes use of dependency tree structures as patterns. The lexico-syntactic patterns are very robust and not dependent on the performance of a parser. However, the enhancement of the pattern set with the dependency trees allowed us to reduce the 1067 lexico-syntactic patterns to 212 dependency tree patterns, with almost the same linguistic coverage. Thus the utilisation of syntactic parsing results eases maintenance of the pattern set in a significant way (Klüwer et al., 2010a).
While tree-based rules help to abstract from surface variations and thus to reduce the number of rules that have to be coded, relevant expressions for the domain still have to be found and formalised for a specific domain. In order to discover all the different means to express a certain fact or question, we drastically extended our initial pattern base by using crowd-sourcing methods in a further effort. We built a platform for acquiring paraphrases of seed questions for biographical facts from multiple human annotators. The seed questions are associated with their semantic arguments and functions. As before, the user input is mapped against this set of paraphrases using fuzzy pattern matching. The resulting resource can be both used for deriving further modular pieces of expression for a compositional input analysis or it can serve as a gold resource for pattern acquisition and system evaluation.

4. Evaluation

Since it is important to know whether our NPCs will be accepted by the end users of the virtual online game, we conducted an evaluation to test usability and acceptability of our system. Similar to well-known dialogue system evaluation frameworks such as PARADISE (Walker et al., 1997), our evaluation measures system-performance regarding user satisfaction and task success. Therefore, our evaluation methods belong to the group of subjective and user-oriented evaluations (Dybkjr et al., 2008).
For entertaining applications, the usual evaluation measures, which target task-based systems, are not necessarily useful. As (Gustafson et al., 2004) point out, computer games are usually evaluated by professional game reviewers, since e.g. the user satisfaction may increase rather than decrease with task completion time. Unfortunately, professional reviewing was not available for the described scenario. Therefore, the post-test questionnaire contains additional questions about naturalness, personality of the agent and fun while using the system to indicate the positive or negative perception of the entertaining aspects of the barkeeper.
Our method is built on top of a successful field test and evaluation of the compass2008 system (Uszkoreit et al., 2007), which combines "field test" with "acceptability test" methods and is based on two standard evaluation instruments: SUMI (Software Usability Measurement Inventory)[3], the de facto industry standard questionnaire for analyzing users' opinions towards software products, and the ISO NORM 9241/10, which checks compliance to ISO Norm 9241/10 (ergonomic requirements for screen work places).

4.1. Evaluation of the compass2008 System

A so-called *Field Test* normally takes more than ten users to test an software application in a real environment. The compass2008 system reported in (Uszkoreit et al., 2007) is a mobile location-based information and communication system, which helps foreigners in China to overcome language barriers and to access needed information in situations such as restaurants, taxis or shops. Therefore, a field test is reasonable for this system, which works in the real environment. In the field test, the usability problems are

[3] http://sumi.ucc.de

Dimension	Sample Item	Scale
Task completion	Could you complete your task	Y/N
Dialogue Efficiency	It took me too much time to complete the task	5-step Likert Scale
Dialogue Control	I always knew what to say next	5-step Likert Scale
Reliability	NPC did what I expected	5-step Likert Scale

Table 1: The Post-Task Questionnaire

Dimension	Sample Item	Scale
Usability		
Dialogue Control	I felt the conversation with the barkeeper being under my control	5-step Likert Scale
	I always knew what to say next	5-step Likert Scale
Reliability	The barkeeper has done something unexpected at some time	5-step Likert Scale
Aesthetics	The barkeeper has a virtually appealing presentation	5-step Likert Scale
Cognitive Demand	There have been times during talking with the barkeeper when I have felt quite tense	5-step Likert Scale
	I had to look for assistance when I talked to the barkeeper	5-step Likert Scale
Acceptability		
Satisfaction	The barkeeper is a nice person	5-step Likert Scale
	I would like to visit the bar again	5-step Likert Scale
	I liked the barkeeper's behavior	5-step Likert Scale
Naturalness	The barkeeper behaved naturally (like a real barkeeper)	5-step Likert Scale
Entertainment	Conversation with the barkeeper was fun!	5-step Likert Scale
	Talking to the barkeeper was boring	5-step Likert Scale
Usefulness	The barkeeper makes virtual worlds like Twinity more interesting	5-step Likert Scale
Improvements/Remarks	What has to be changed/What did you like	open-end

Table 2: The Post-Test Questionnaire

collected by subjective reports of test users (e.g. online questionnaires). Logging data can be used to assess usage duration and quality.

The second method, namely, the *acceptability* test can be run in real and laboratory environments. Acceptability is measured by test users self reports (questionnaires, interviews).

The experts for software quality and usability at the German Telekom were responsible for the design of the evaluation methods for the compass2008 system. They referred to two standard evaluation instruments which are wide-spread for evaluating commercial software applications: 1) SUMI (Software Usability Measurement Inventory)[4]; 2) ISO NORM 9241/10. SUMI is the de facto industry standard questionnaire for analyzing users' responses to desktop software or software applications provided through the internet. SUMI emphasizes the five dimensions of subjectively experienced usability (Efficiency, Affect, Helpfulness, Control, Learnability), while ISO NORM 9241/10 checks compliance to ISO Norm 9241/10 (ergonomic requirements for screen work places) and tests the following additional features: suitability, self-descriptiveness, controllability, conformity with user expectations, error tolerance, suitability for individualization and learn-ability.

[4]http://sumi.ucc.de

4.2. Evaluation Design for the KomParse Evaluation

In comparison to the compass2008 system, the *KomParse* system tries to assist and entertain users in a virtual environment. Both field test and acceptability test can take place in a laboratory environment, namely, in a room with computers where the software programs are running. However, since the virtual environment in *KomParse* simulates the real world, we can still conduct a kind "virtual field test". Therefore, we developed a field test containing a list of tasks to solve in the virtual world through a conversation with the barkeeper NPC. Twelve people were selected as our subjects for the test, mainly students with different knowledge and experience with virtual games. The tests take place in front of personal computers. The subjects have to fulfill a list of tasks with the *KomParse* system via conducting dialogues with the NPC barkeeper. The tasks included but were not restricted to ordering cocktails and asking for biographic information of a famous personality. We also added "small talk with the Barkeeper" to the list of the tasks to ensure that the users take the opportunity to chat with the NPC. For each task, the subjects had to fill out a post-task questionnaire. The used Likert scales contain values for agreement, namely "totally agree", "agree", "neutral", "disagree" and "totally disagree".

Table 1 shows the post-task questionnaire for a single task. Additionally, the evaluation questionnaire covers questions regarding the performance of the overall system as well as the system's single communication components.

Dimension	Sample Item	Scale
understanding	I had the feeling the barkeeper understood me well	5-step Likert Scale
response generation	What the barkeeper said made sense to me	5-step Likert Scale
dialogue moves	The barkeepers reactions are appropriate	5-step Likert Scale
	NPC has done unexpected things	5-step Likert Scale
Large scale knowledge	The barkeeper seemed informed about the world	5-step Likert Scale

Table 3: The Post-Test Questionnaire for Components

Question	Average Results Task 1	Average Results Task 2	Average Results Task 3
Could you complete your task?	yes: 8, problems; 3, no:1	yes:10, problems: 1, no:1	yes:6, problems: 6, no:0
It took me much time to complete the task	-0.917	-0.750	-0.750
I always knew what to say next	+0.333	+0.667	+0.667
The agent has always done what I was expecting	-0.417	0	0

Table 4: The Results of the Post-Task Questionnaire

The post-test questionnaire for the system evaluation includes the items in table 2.
The dimensions in the post-task and post-test questionnaires are selected carefully from SUMI and ISO NORM 9241/10.
In addition to the above general usability and acceptability tests, we are also interested in the user opinions of communication capabilities of the NPC. Table 3 show the features which are included in the components questionnaire.

4.3. Evaluation Results

The evaluation results are very useful and provide valuable insights into the acceptability of the system. Table 4 shows the results of the post-task questionnaire. The average values from the Likert scale are calculated by translating the agreement values to numerical values: "totally agree" (+2), "agree" (+1), "neutral" (0), "disagree" (-1), "totally disagree" (-2).
The users report positive results for all given tasks in the post-task questionnaire. Even for the small talk task which was given to guarantee that the users will make use of the small talk opportunity, feedback is positive. Most users could complete all tasks, only 2 users report that they could not fulfill a task. That means that *task completion* is very high. However, a comparatively huge number of users report problems which occurred during a task (10 against 24 without problems). Some of the problems are mentioned in the general positive/negative remarks in table 6.
The average of the users stated that they did not feel that it takes too much time to complete the tasks. Especially the task 1 "Ordering a drink" gets a good result with -0.917 in average. Therefore, the *dialogue efficiency* of the *KomParse* system is satisfying. The *dialogue control* in the task seems to be a little better in the two other tasks, "Information about a celebrity" and "Small Talk". Both get an average value of $+0.667$ for the statement "I always knew what to say next". The *reliability* (NPC did what I expected) of the agent is rated rather neutral in average. The drink-task even gets an average value of -0.417. That indicates that the ability of the agent to act appropriately to the user's expectations is still to improve.
Additionally to the post-task questionnaire, the evaluation includes a post-test questionnaire covering questions regarding the usability of the overall system as well as the system's single communication components. Table 5 shows the overview about the results from the post-test questionnaire.
The users report positive feedback in the post-test questionnaire, too. The system gets very good satisfaction results: Eight of twelve users would visit the bar again, the remaining four have a neutral opinion, no subject disliked the possibility to come back. The average value is $+1.083$. The *interest to use the system* is encouraging.
Likewise, most people like the barkeeper's behavior ($+0.667$) and agree with the opinion that the barkeeper is a nice person ($+0.833$). This indicates that the *satisfaction and the attitude towards NPC* among the users is very high. Moreover some users agree that the barkeeper behaves *naturally* like a real barkeeper. The average value is $+0.250$.
The barkeeper also does a good job on *entertaining* the users: nine of twelve users agreed with the opinion, that the conversation with the barkeeper was fun - the average scale value is $+0.833$ - and deny the statement that talking to the barkeeper was boring (-0.750).
Understandably, ten of twelve subjects estimated the barkeeper to make virtual worlds more interesting, what acknowleges the *usefulness* of the system.
The *usability* of the system seems to be satisfying, too. Most users like the appearance of the NPC. The average scale value is $+0.917$. The *aesthetics and naturalness* of the NPC is hence confirmed by the users. However, the *dialogue control* seems to be problematic. The average value for the statement "I felt the conversation with the barkeeper being under my control" is slightly negative with -0.333. On the other hand the decision what to say next is uncomplicated ($+0.667$). Regarding the general remarks (6), it seems that the system initiative sometimes is too fast for the user. We may have to adjust the time and selection values for the system initiative. This is also reflected in the comparatively high value for the *reliability* of the barkeeper: the average value is $+1.167$. However, this does not become noticeable in the *cognitive demand*. Very few users (2 of

Question	Average Results
The barkeeper makes virtual worlds like Twinity more interesting	+1.250
Conversation with the barkeeper was fun!	+0.833
There have been times during talking with the barkeeper when I have felt quite tense	-0.500
I liked the barkeeper's behavior	+0.667
The barkeeper has done something unexpected at some time	+1.167
I felt the conversation with the barkeeper being under my control	-0.333
The barkeeper has a visually appealing presentation	+0.917
I always knew what to say next	+0.667
I had the feeling the barkeeper understood me well	-0.167
The barkeeper's reactions are appropriate	+0.500
What the barkeeper said made sense to me	+0.250
I had to look for assistance when I talked to the barkeeper	-0.750
The barkeeper seemed informed about the world	+0.833
The barkeeper behaved naturally (like a real barkeeper)	+0.250
Talking to the barkeeper was boring	-0.750
The barkeeper is a nice person	+0.833
I would visit the bar again	+1.083

Table 5: The Results of the Post-Test Questionnaire

Positive User Feedback	**Negative User Feedback**
the smilies	sometimes the answers to questions don't appear
you an answer the questions with numbers	sometime the reaction will need to long
some funny answers	it seems that the NPC ignores the input
the way he offers an answer is very natural	more phrases for interaction will be better, e.g., "I want ...", "Please make me a ..."
an open-minded person, Hank Slender	Hank does not know the ingredients of a cocktail
works well in general	I had the feeling, if I am not fast, he will change the topic.
Hank tried to change the topic himself. It makes the conversation natural	a little less quiz would be better

Table 6: Answers Given to the Open-End Question in the Post-Test Questionnaire

12) felt stressed during talking with the barkeeper (-0.500) and there was no huge need for assistance (-0.750).

The subjective component evaluation provides very useful feedback. Only 3 of 12 users agree to the statement that the NPC understood them well. The scale is more negative with a value of -0.167. This means that the *natural language understanding* component is still the bottleneck of the system. Although dialogue act recognition works well in comparison to the state of the art (Klüwer et al., 2010b), the system needs more paraphrases and rules for the generation of valid syntactic relations and more training data for the classification of rare utterances to dialogue acts.

On the generation side, about 58% of users accept the *response* of the NPC (the scale value is $+0.250$) and 83% of users find the reaction of the NPC appropriate.

The perceived level of information of the NPC gets a positive value of $+0.833$. We take that as an indicator that the *knowledge* bases, namely the cocktail ontology and the celebrity database, create a sufficient basic knowledge for the agent.

The post-test questionnaire includes one open-ended question asking for general positive and negative features of the system. Table 6 shows some answers the users give to the open-end question. There are both positive and negative feedbacks returned by users, which are very important for the further development and research. The positive comments show that the user interface design of the avatar is widely accepted by the users and that they enjoy talking to the barkeeper. The negative rates tell us that a even more fine grained ontology of cocktails is helpful for our scenario and that we have to adjust the system initiative to achieve a more reliable behavior of the barkeeper.

In general, both the usability and acceptability test results show that our NPC, the barkeeper Hank Slender is useful and well-accepted by the users.

5. Conclusion and Future Work

The paper describes the evaluation of the *KomParse* system, which controls conversational non player characters (NPCs) in a virtual environment. The system was evaluated through a field test with users in the virtual world *Twinity*. The evaluation subjects should carry out two dialogue tasks and have a small talk conversation with our barkeeper NPC. The system was evaluated amongst others in terms of acceptability, usability and dialogue efficiency of the overall system as well as the single communication components and in terms of acceptability and usability regarding the single tasks. The results are very satisfying and promising. In general, both the usability and acceptability test results show that our NPC, the barkeeper Hank Slender

is useful and well-accepted by the users. Users regard the NPC as a virtual person, even as an open-minded personality. Furthermore, the active control of the topic switch during conversation by the NPC is also regarded as useful by the users. Even if the NPC does not always understand the users well and said things unexpected, he could still provide appropriate response to help users to solve the problems or entertain them. Nevertheless, a big coverage of paraphrases is needed to make the conversation more robust and natural.

Acknowledgements

This paper is supported by the project KomParse, funded by the ProFIT program of the Federal State of Berlin, co-funded by the EFRE program of the European Union, and the project T4ME, funded by the 7th Framework Programme of the European Commission with the grant agreement no. 249119.

6. References

Peter Adolphs, Xiwen Cheng, Tina Klwer, Hans Uszkoreit, and Feiyu Xu. 2010. Question answering biographic information and social network powered by the semantic web. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), 5.

Marc Cavazza, Raúl Santos de la Cámara, Markku Turunen, José Relaño Gil, Jaakko Hakulinen, Nigel Crook, and Debora Field. 2010. 'How was your day?': an affective companion ECA prototype. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 277–280, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nigel Crook, Ramn Granell, and Stephen G. Pulman. 2009. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 341–348.

Laila Dybkjr, Holmer Hemsen, and Wolfgang Minker, editors. 2008. *Evaluation of Text and Speech Systems*, volume 37 of *Text, Speech and Language Technology*. Springer.

Patrick Gebhard, Michael Kipp, Martin Klesen, and Thomas Rist. 2003. Authoring scenes for adaptive, interactive performances. In *Proc. of the Second International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03)*.

Joakim Gustafson, L. Bell, J. Boye, A. Lindström, and M. Wirn. 2004. The nice fairy-tale game system. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

J. Gustafson, J. Boye, M. Fredriksson, L. Johannesson, and J. Knigsmann. 2005. Providing computer game characters with conversational abilities. In *Proceedings of Intelligent Virtual Agent (IVA05)*, Kos, Greece.

All W. Hill, Jonathan Gratch, Stacy Marsella, Jeff Rickel, William Swartout, and David Traum. 2003. Virtual humans in the mission rehearsal exercise system. *In KI Embodied Conversational Agents*, 17:32–38.

Patrick G. Kenny, Thomas D. Parsons, Jonathan Gratch, and Albert A. Rizzo. 2008. Evaluation of Justina: A Virtual Patient with PTSD. In Helmut Prendinger, James C. Lester, and Mitsuru Ishizuka, editors, *IVA*, volume 5208 of *Lecture Notes in Computer Science*, pages 394–408. Springer.

Martin Klesen, Michael Kipp, Patrick Gebhard, and Thomas Rist. 2003. Staging exhibitions: Methods and tools for modelling narrative structure to produce interactive performances with virtual actors.

Tina Klüwer, Peter Adolphs, Feiyu Xu, Hans Uszkoreit, and Xiwen Cheng. 2010a. Talking npcs in a virtual game world. In *Proceedings of the ACL 2010 System Demonstrations*. Association for Computational Linguistics, 7.

Tina Klüwer, Hans Uszkoreit, and Feiyu Xu. 2010b. Using syntactic and semantic based relations for dialogue act recognition. In *Coling 2010: Posters*, pages 570–578, Beijing, China, August. Coling 2010 Organizing Committee.

S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth. 2005. A conversational agent as museum guide – design and evaluation of a real-world application. In *Proc. of Intelligent Virtual Agents (IVA 2005)*, volume 3661, pages 329–343. Springer.

S. Larsson and D. Traum. 2000. Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering*, 6:323–340.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van, and Ess dykema Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Hans Uszkoreit, Feiyu Xu, Weiquan Liu, Jörg Steffen, Ilhan Aslan, Jin Liu, Christel Müller, Bernhard Holtkamp, and Manfred Wojciechowski. 2007. A successful field test of a mobile and multilingual information service system compass2008. In *Proceedings of HCI International 2007, 12th International Conference on Human-Computer Interaction*.

A.T. Verbree, R.J. Rienks, and D.K.J. Heylen. 2006. Dialogue-act tagging using smart feature selection: results on multiple corpora. In B. Raorke, editor, *First International IEEE Workshop on Spoken Language Technology SLT 2006*, Palm Beach. IEEE Computer Society.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280, Morristown, NJ, USA. Association for Computational

Linguistics.

Feiyu Xu, Hans Uszkoreit, and Hong Li. 2007. A seed-driven bottom-up machine learning framework for extracting relations of various complexity. In *Proceedings of ACL 2007, 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, 6.

Feiyu Xu, Hans Uszkoreit, Sebastian Krause, and Hong Li. 2010. Boosting relation extraction with limited closed-world knowledge. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, o.A.