Integrating a Model for Visual Attention into a System for Natural Language Parsing

Christopher Baumgärtner and Wolfgang Menzel

University of Hamburg, Hamburg, Germany

Abstract. We present a system for integrating knowledge about complex visual scenes into the process of natural language comprehension. The implemented system is able to choose a scene of reference for a natural language sentence from a large set of scene descriptions. This scene is then used to influence the analysis of a sentence generated by a broad coverage language parser. In addition, objects and actions referred to by the sentence are visualized by a saliency map which is derived from the bi-directional influence of top down and bottom-up information on a model of visual attention highlighting the regions with the highest probability of attracting the attention of an observer.

1 Motivation

We introduce a system for natural language processing that integrates knowledge about visual context into the task of parsing sentences. This integration of contextual information can result in different interpretations of the same sentence depending on the knowledge of a listener, as well as on the selectively perceived surroundings at the moment of processing language input.

An artificial system that incorporates contextual knowledge should utilize descriptions of scenes in order to modify the interpretation of a given sentence. This is straightforward as long as it is evident which parts of contextual information has to be used to influence processing. The task gets more difficult when a wide range of possibly influential information units are accessible to the system. In this case processing all available contextual information might not be feasible.

Our system finds a scene of reference for a given sentence and uses this scene to improve the processing of a language parser. For this purpose we adopt findings about context integration in humans as an inspiration for properties our system should exhibit.

2 Related Work

Several models were built that integrate language and vision, attempting to realize language driven reference resolution albeit not directly addressing the issue of attention focusing. Winograd's SHRDLU [18] relates written language to objects in a blocks world domain. Knowledge about this domain is stated in a database of facts. The system receives instructions about what to do with the objects of its domain, parses these,

Baumgärtner C. and Menzel W..
Integrating a Model for Visual Attention into a System for Natural Language Parsing.
DOI: 10.5220/0004088100240033
In *Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science* (NLPCS-2012), pages 24-33
ISBN: 978-989-8565-16-7

finds referents for parts of the parsed sentence and acts accordingly, asking questions whenever an instruction is not understandable or ambiguous.

The system Bishop [6] is based on studies of how people describe objects and their spatial relations. It operates in a domain of cones with different colors. The system receives descriptions like: *the purple cone in the middle to the left of the green cones* parses them and chooses the correct referents accordingly. This is done by incorporating information about color (green or purple), grammatical number (cone or cones) and spatial relations (middle, left, front-most).

In [12] language ambiguity is resolved by integrating external contextual knowledge into a language parser. The system is able to use the context information while processing German utterances by adopting a predictor-based approach: Contextual knowledge is provided prior to processing a sentence, thus influencing the results of subsequent parsing decisions. This influence is mediated by a semantic level which was incorporated into a system originally intended solely for syntactic analysis. So far, however, the system has not been shown to be able to select referents from competing alternative context items according to their relevance for language comprehension.

3 System Architecture

To implement the described behavior in a computational system at least four subsystems are required: a system for natural language processing, a model for the representation of the visual context, a system determining the focus of attention in a picture that shows the visual context and a component for managing hypotheses of referential context information. Our model assumes the role of an agent listening to a sentence while looking at a visual scene.

3.1 WCDG

Contextual knowledge need not be fully consistent with the propositional content of an utterance. Minor deviations or even serious conflicts can be observed in almost any communicative setting. To be able to benefit from information which is partly unreliable, a component for natural language processing is required that allows the analysis to make use of such evidence by arbitrating between competing interpretations based on individual estimations of confidence.

Weighted Constraint Dependency Grammar (WCDG) [16] seems to be a promising candidate for such a task. It has been shown to be able to successfully integrate cues from a wide range of external predictors in a robust manner [5], it can be used to capture the relationships between parallel levels of representation (syntax and semantics) without enforcing a rigid mapping between them [12] and it can be easily interfaced to external reasoning components.

WCDG is based on constraints which are used to disambiguate structural interpretations of a given sentence. The system receives a lexicon, a grammar and a written sentence as input. The constraints in the grammar are weighted with values from zero to one indicating their relevance, with zero implying a hard constraint that must be satisfied by any solution and one indicating an unimportant constraint. The output is a forest of parse trees of word-to-word dependencies.

Parsing in WCDG is a global optimization task. The standard solution procedure is a transformation-based algorithm (frobbing) [4]. It starts from an initial analysis for the input sentence which is likely to violate at least some of the constraints imposed by the grammar. The algorithm repairs the most important violation by exchanging those edges of the analysis that are accountable for violating the constraint. If the repaired analysis still violates important constraints, the algorithm starts over by repairing the next violation. By this WCDG can even revoke decisions previously made.

As the grammar contains weighted constraints the algorithm does not need to repair all violations of constraints contained in a particular analysis to obtain a solution to the parsing problem. In most cases any analysis will violate some constraints. As long as these do not include hard constraints (those with value zero) the analysis is still a viable solution. The system will proceed to repair this analysis in order to optimize the solution to the problem.

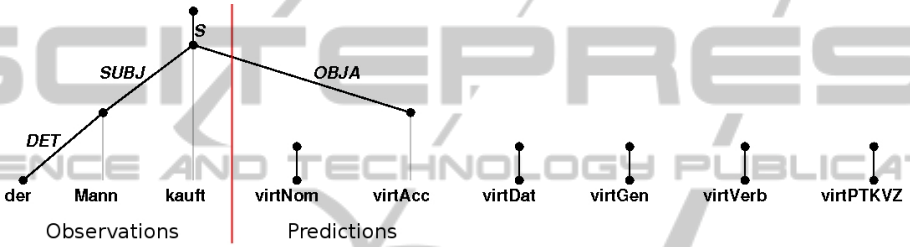

Fig. 1. Prediction of the upcoming object for the fragment *Der Mann kauft* (The man buys) by selecting the not yet instantiated virtAcc node from a set of possible candidates.

The integration of contextual information is a bi-directional process where language processing influences search for information, which in turn influences language processing, which again might influence visual search. In order to implement this kind of behavior, a language processing system that outputs only the final analysis of a sentence without giving intermediate results is of little use. In order to influence context processing, we need access to analysis results at an early stage when processing is not finished yet.

WCDG is useful in this regard because of the anytime property of the frobbing algorithm: As the system generates a final solution to a problem by continuously improving suboptimal solutions, it is possible to access crucial information contained in these less-than-perfect analyses at a very early point in time. In particular, such information can be used to resolve referential ambiguity among a large set of available visual entities.

One feature of the parser is its ability to deal with sentences incrementally, i.e. word by word [1]. In this mode of operation it processes a sentence-prefix to a certain degree, adds a new word and uses the results of the already analyzed sentence fragment as a starting point of further processing. If additional nodes for yet unseen words are included into the model the parser is even able to predict upcoming elements of the sentence(Figure 1).

3.2 Knowledge Representation

We use OWL [14] to manually build a formal description of the visual context of a scene. The representation describes objects in the scene including their properties (like **color(cube, red)**). These descriptions also include spatial relationships between objects. In addition to descriptions of processes solely by spatial terms, for our system we also use descriptions of actions and the persons and objects that are engaged in them. The system also has access to knowledge about taxonomic relations between concepts of objects (i.e. that a poodle is a dog; a dog is a mammal; a mammal is an animal).

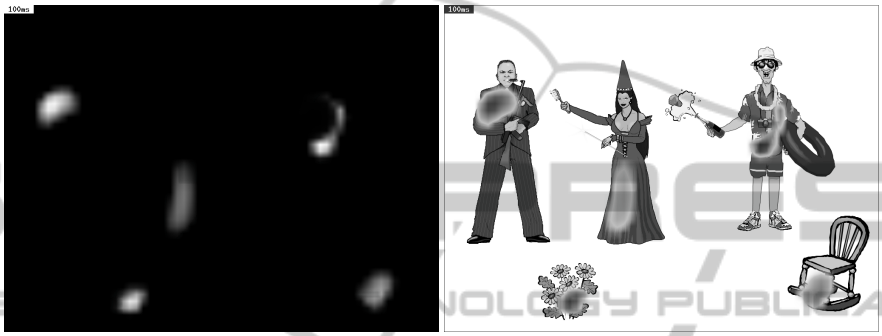

Fig. 2. Saliency Map for a scene consisting of three characters (Underlying picture adapted from [9]).

3.3 Modeling Bottom-up Visual Attention

While the top-down influence of the visual context on language processing is working with symbolic descriptions of a given scene, a model driven by sensor-data is used to capture the bottom-up influence on attention [8] and to visualize the effects of combining top-down and bottom-up information on visual attention. In this model, visual attention of subjects is predicted, depending on low-level features of the picture. It extracts visual cues in the dimensions of color, intensity and orientation. Saliency maps are computed from these features, indicating regions most likely to attract the attention of an observer (Figure 2).

Humans do not stick to just one region in the picture but move their eyes over time, thus exploring different regions. To model this behavior, the saliency of a region is changed, depending on which parts of the picture are currently in focus. This is accomplished by a mechanism called ”Inhibition of Return”: saliency is reduced for the region currently in focus. As soon as the saliency of this region drops below the saliency of another part of the picture, attention will switch to the new saliency maximum, whose saliency will then be reduced subsequently. As the former region is not in focus any more, saliency will start to increase again. Due to this dynamics of the saliency landscape, focus always moves towards the region that is currently the one with the global maximum, resulting in a sequence of predictable eye-movements.

3.4 Finding a Scene

The information from different modalities is encoded by means of different representations, which are tailored to meet the specific requirements of the different input channels. Unfortunately, the differences between the representations make it difficult to integrate the information they contribute. Therefore, a component has been included that mediates between these representations in both directions.

On the one hand, the information integration component makes a choice about which subset of contextual representation fits the given linguistic information best. In order to do this, the possible referential scenes must be graded with regard to how good they match a given sentence.

On the other hand this component computes a score describing how strong the influence of the chosen visual context will be. For any given sentence a scene is identified that describes the content of the pictures fitting the sentence best.

4 Benefits

The system architecture presented provides the benefit of being able to predict upcoming parts of the utterance, to combine top-down and bottom-up evidence in order to determine the focus of attention and to capture the dynamic aspects of language processing as the sentence unfolds over time.

4.1 Predicting Elements

During incremental parsing upcoming elements of a sentence can be predicted by investigating those parts of the context which are connected to the parts chosen as referents for the already parsed subsentence. For example in Figure 1 the subsentence *Der Mann kauft* (The man buys) has already been parsed and referents have been found in the context. As the context model contains a description of a man buying a book, the missing object of the sentence is inferred to be the book.

4.2 Shifting Focus

When a picture is input without any kind of additional language information, the saliency of regions depends only on the visual elements. An initial saliency landscape is given in (Figure 2 left side) which evolves gradually over time, driven by bottom-up cues. Additional top-down influence might change the saliency of a picture region depending on the linguistic input and the chosen referents for parts of a sentence (Figure 3).

To model the influence of the language on the saliency of regions in a picture (and therefore on the choice of the focus of attention) WCDG constraints are used. They select the linguistic input which is likely to provide top-down information for guiding attention to a specific part of the picture and connect the information contained in WCDG's parsing results (words, lexical features or word-to-word dependencies) with parts of the context information. Thus, the occurrence of certain linguistic phenomena analyzed by WCDG will result in an increase or decrease of saliency for certain picture regions.

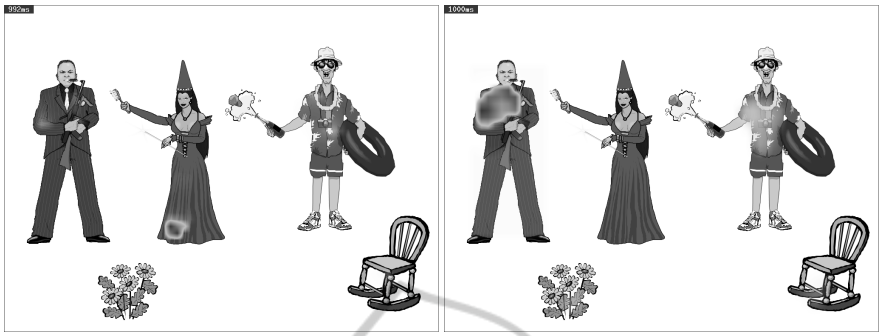

Fig. 3. Saliency before and after processing the sentence *The fairy brushes the gangster* (Underlying picture adapted from [9]).

4.3 Cyclic Refinement

A simple constraint stating that during incremental parsing each new word would increase saliency of the referent of this word in the picture can already be used to model the linguistic influence. The integration of language and vision, however, requires a more complex approach which can be illustrated by the sentence *Der Gangster bedroht die Person neben dem Stuhl* (The gangster threatens the person next to the Chair.) referring again to Figure 3. Finding the correct references for this sentence is not a straightforward task for a number of reasons.

First of all there is the structural ambiguity caused by the different attachment possibilities of the PP *neben dem Stuhl* (next to the chair). It might be stating that the event of threatening happens next to the chair or, as in our case, the person threatened is located somewhere next to the chair. This problem is solved by inserting a constraint into the grammar making the attachment of a prepositional phrase dependent on the existence of a relationship between visually present persons in the corresponding context. So, without context the sentence will be parsed to the default analysis in (Figure 4). Introducing, however, the context represented in (Figure 5 right side)the parser will switch to the analysis in Figure 5 as its final result. The reason for choosing different attachments in the two cases is the contextual description of a person (the tourist) standing next to the chair.

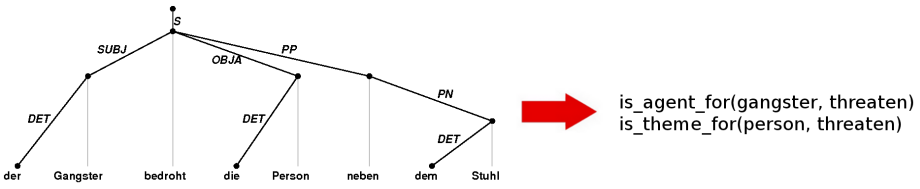

Fig. 4. Parsing the sentence *The gangster threatens the person next to the chair* without context.

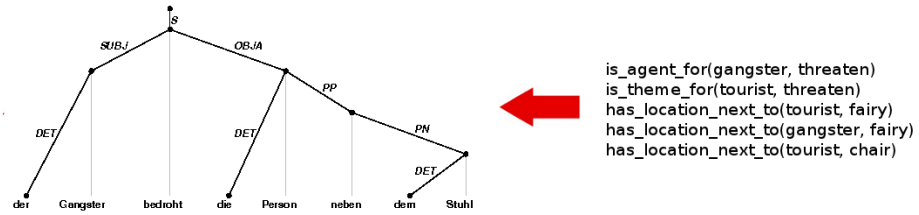

Fig. 5. Parsing the sentence *The gangster threatens the person next to the chair* with context.

A more serious issue regarding the connection between the subsystems is finding the correct referent for the word *Person* (person). Obviously this word is ambiguous given the context illustrated in the picture, as it might refer to any of the three individuals since all of them are related in the taxonomic hierarchy mentioned in Section 3.2 to the concept **person**. It is disambiguated by a constraint in the grammar, that prefers PP-attachment if there is some kind of corresponding connection in context as well, be it a spatial relationship or the participation in an action depicted. Using this constraint as part of the evaluation process, the system will choose as referent for the word *Person* the individual stated as having a connection with the chair, indicated by the prepositional phrase of the dependency tree (Figure 5 left side). This individual is, in our case, the tourist.

Unfortunately the two problems just addressed are connected in a way that each cannot be solved without first solving the other one. The system finds the correct attachment for the prepositional phrase by finding a spatial connection between the referents for the words *Person* and *Stuhl* but it can only find the correct referent for *Person* by first having the prepositional phrase attached in a way compatible with contextual information. This dilemma is solved by constantly exchanging information between the subsystems while the sentence unfolds over time. Due to our additional visual constraint that makes the PP-attachment between connected referents more likely, each time the parser builds such an edge between words with unconnected referents, this solution is penalized. This will force the parser to search for a less penalized analysis with a different attachment-point for the preposition.

When, at some point of processing, the parser attaches the PP/PN-edges between *Person* and *Stuhl* this analysis might be penalized too, as the referent for *Person* can be incorrect or not chosen at all. But in this case the second visual constraint increases the likelihood that the referent for *Person* is an individual that is spatially connected to the tourist, ensuring that the tourist will be chosen as the correct individual.

5 Results

We evaluated our system with pictures and sentences taken from [9] where they have been used to investigate he interaction of language and visual context of human participants.They show scenes like the one in Figure 3 with three characters: one that is the agent of an action (the tourist), one that is the patient of an action (the gangster) and one that is ambiguous with regard to agent/patient-role (the fairy). Each picture is accompanied by two sentences describing the actions depicted. For Figure 3 sentences were *Die*

Fee bürstet hier den Gangster. (Literally: The fairy brushes here the gangster.) and *Die Fee bespritzt hier der Tourist.* (Literally: The fairy is splashed here by the tourist.) In German, the first noun (the fairy) is ambiguous with regard to its role until the second noun phrase is received during incremental parsing. [9] suggested that humans are able to assign the correct role for the first noun phrase and to anticipate the missing role already after hearing the verb, by incorporating the visual knowledge of the perceived scene as a substitute for the not yet available parts of the sentence.

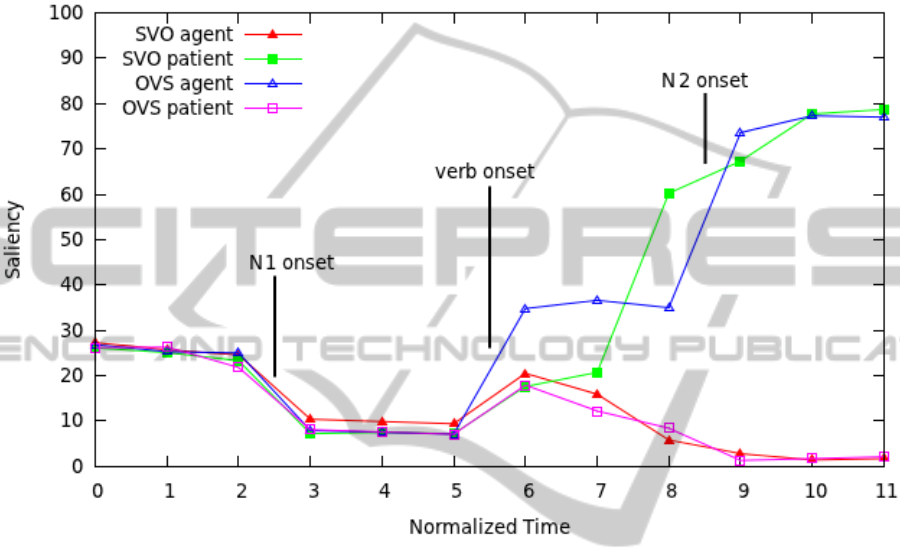

Fig. 6. Development of saliency for agent and patient in SVO and OVS sentences.

Figure 6 shows the development of saliency for the different characters. The graphs present the saliency for agent and patient of the picture for OVS and SVO sentences. The time course is normalized over all trials, such that any sentence processed is divided into twelve time steps, depending on the progress of the incremental parsing. Time steps 0-2 show the saliency before the onset of the first noun. Time steps 3-5 depict the development after the first noun has been added to the sentence fragment. This results in a drop of the saliency for the agent as well as the patient, because saliency is increased for the ambiguous character in the picture, as denoted by the first noun of the sentence. Time steps 6-8 show saliency after verb onset. The crucial effect is the development of the saliency of the character not yet addressed (i.e. the patient for SVO-sentences, the agent for OVS-sentences). Although the second noun is not processed until steps 9-11, saliency already increases for the corresponding character. This development is in agreement with findings about human eye-movements as described in in [10]. Indeed, a closer inspection of the parsed sentences shows that this development is caused by the fact, that the first noun and the verb enable the system to assign the correct roles to both characters, even if the second noun is not yet part of the sentence fragment. This prediction of the correct role for an unnamed referent results in an increase in saliency for the corresponding region.

6 Future Work

We have described a complex system which integrates visual context into language processing. The system is able to differentiate between relevant and irrelevant context items, choosing only those entities of the described visual scene that seem to be important for a given sentence. In contrast to previous approaches our system can find the applicable information in a diverse, dynamic context, switching between its choice whenever additional information is received, thereby guiding the focus of a model of human visual attention.
So far, the integration of context information is restricted to descriptions of static scenes. In the future, we will also investigate changing representations of context to model the influence of a dynamic environment.
At the moment the selected referents influence saliency of a picture over time, but once a scene of reference has been selected for parts of a sentence, even major changes in saliency due to low-level cues in the picture will not be able to question this choice. To remove this limitation a goal of future research will be to model the influence of bottom-up attention on the choice of visual referents for linguistic input.

References

1. Beuck, N., Köhn, A., Menzel. W.: Incremental parsing and the evaluation of partial dependency analyses In Proceedings of the 1st International Conference on Dependency Linguistics, DepLing-2011 (2011) 290–299
2. Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C:, Tanenhaus, M. K.: Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. Journal of Psycholinguistic Research 24 (1995) 409–436
3. Egeth, H. E., Yantis, S.: Visual attention: Control, representation, and time course. Annual Review of Psychology 48 (1997) 269-297
4. Foth, K.: Transformationsbasiertes Constraint-Parsing Diplomarbeit Universität Hamburg (1999)
5. Foth, K.: Hybrid Methods Of Natural Language Analysis PhD Thesis Universität Hamburg (2006)
6. Gorniak, P., Roy, D.: Grounded Semantic Composition for Visual Scenes. Journal of Artificial Intelligence Research 21 (2004) 429–470
7. Haddock, N. J.: Computational models of incremental semantic interpretation. Language and Cognitive Processes 4(3) (1989) 337–36.
8. Itti, L.: Models of Bottom-Up and Top-Down Visual Attention. California Institute of Technology Ph.D. Thesis (2000)
9. Knöferle, P.: The Role of Visual Scenes in Spoken Language Comprehension: Evidence from Eye-Tracking. PhD thesis Universität des Saarlandes (2005).
10. Knöferle, P., Crocker, M. W., Scheepers, C., Pickering M. J.: The influence of the immediate visual context on incremental thematic role-assignement evidence from eye-movements in depicted events. Cognition 95 (2005) 95–127
11. Knöferle, P., Crocker M. W.: The influence of recent scene events on spoken comprehension: evidence from eye movements. Journal of Memory and Language 57(2) (2007) 519-543
12. McCrae, P.: A model for the cross-modal influence of visual context upon language processing. Proceedings of the International Conference Recent Advances in Natural Language Processing (2009) 230–235

13. Menzel, W.: Towards radically incremental parsing of natural language. Current Issues in Linguistic Theory 309 (2009) 41–56
14. W3C-World Wide Web Consortium. OWL Reference, 10.02.2004. http://www.w3.org/TR/2002/REC-owl-ref-20040210 (2004).
15. Scheutz, M., Eberhard, K., Andronache, V.: A Real-time Robotic Model of Human Reference Resolution using Visual Constraints. Connection Science Journal 16(3) (2004) 145–167
16. Schröder, I.: Natural Language Parsing with Graded Constraints PhD Thesis Universität Hamburg (2002).
17. Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., Carlson, G. N.: Achieving incremental semantic interpretation through contextual representation. Cognition 71 (1999) 109–147
18. Winograd, T.: A Procedural Model of Language Understanding. Computer models of thought and language (1973)