Incorporating Contextual Information for Language-Independent, Dynamic Disambiguation Tasks

Tobias Staron, Özge Alaçam, Wolfgang Menzel
Department of Informatics
University of Hamburg
{staron, alacam, menzel}@informatik.uni-hamburg.de

Abstract

Humans resolve various kinds of linguistic ambiguities by exploiting available external evidence that has been acquired from modalities besides the linguistic one. This behavior can be observed for several languages, for English or German for example. In contrast, most natural language processing systems, parsers for example, rely on linguistic information only without taking further knowledge into account. While those systems are expected to correctly handle syntactically unambiguous cases, they cannot resolve syntactic ambiguities reliably. This paper hypothesizes that parsers would be able to find non-canonical interpretations of ambiguous sentences, if they exploited external, contextual information. The proposed multi-modal system, which combines data-driven and grammar-based approaches, confirmed this hypothesis in experiments on syntactically ambiguous sentences. This work focuses on the scarcely investigated relative clause attachment ambiguity instead of prepositional phrase attachment ambiguities, which are already well known in the literature. Experiments were conducted for English, German and Turkish and dynamic, i. e. highly dissimilar, contexts.

Keywords: disambiguation, context, language-independence

1. Introduction

In psycholinguistics, there is substantial empirical evidence suggesting that human language processing successfully integrates available information acquired from different modalities in order to resolve fully as well as temporally ambiguous linguistic input, e. g. on the syntactic level, and predict what will be revealed next in the unfolding sentence (Tanenhaus et al., 1995; Altmann and Kamide, 1999; Knoeferle, 2005). During spoken communication, disambiguation and prediction processes allow for more accurate understanding. In contrast, natural language processing (NLP) systems are still not able to achieve that level of accuracy concerning challenging linguistic situations.

For example, a parser that processes linguistic information is expected to successfully handle syntactically unambiguous sentences by applying knowledge derived from training data or linguistic rules. However, neither parsers nor humans can resolve syntactic ambiguities without additional information. They may only have preferences. But, humans will use external information from other modalities for disambiguation successfully if it becomes available. Therefore, we expect to resolve syntactic ambiguities via multi-modal disambiguation by exploiting external knowledge, i. e. contextual information, that is derived from another modality, e. g. from visual scenes of the described events. This work proposes two hypotheses.

H1) A parser will resolve linguistic ambiguities reliably and reach correct interpretations if contextual information derived from additional modalities besides the linguistic one, i. e. visual scenes, are exploited.

H2) In addition, this behavior is expected to be observed independent of the language.

The contributions of this work are twofold. First, a language independent, data-driven parser has been modified to employ a grammar-based approach that will incorporate the contextual information even if it is previously unseen. Secondly, that system is used to validate the hypotheses for multiple languages: English, German and Turkish.

One of the most frequently investigated cases of syntactic ambiguity are prepositional phrase (PP) attachment ambiguities, where different semantic and syntactic interpretations are possible depending on assigning different thematic roles (Tanenhaus et al., 1995). The example *"the woman shoots the man with the pistol"* can be interpreted in different ways. Either, the woman is using the pistol to shoot the man or the man is holding the pistol. Instead, this work investigates the attachment ambiguity concerning relative clauses (Alaçam et al., 2018). In both cases, a multi-modal setting where the visual information constrains the referential choices helps the disambiguation process.

This paper is structured as follows. Section 2. describes the multi-modal data-set that contains syntactically ambiguous sentences and respective, disambiguating contextual information and that has been used to validate our hypotheses. Section 3. proposes a multi-modal disambiguation scheme, which has been used for the experiments, the results of which are shown in Section 4. and analyzed in Section 5.. Next, Section 6. describes related work followed by the discussion of the results.

2. Multi-Modal Data-Set

There are only few data-sets available that address complex linguistic ambiguities. The corpus of language and vision ambiguities (LAVA) (Berzak et al., 2016) contains 237 ambiguous sentences for English, which can only be disambiguated using respective external knowledge provided as short videos or static visual images with real world complexity. The LAVA corpus addresses a wide range of syntactic ambiguities including prepositional as well as verb phrase attachments and ambiguous interpretations of con-

Language	Voice	Exemplary Sentence	PoS Template
English	active	The woman carves the head of the bed, which the man paints.	NP1$_{nom}$ VP1 NP2$_{acc}$ PP1, WDT$_{acc}$ NP3$_{nom}$ VP2.
	passive	The woman carves the head of the bed, which is painted by the man.	NP1$_{nom}$ VP1 NP2$_{acc}$ PP1, WDT$_{nom}$ VP2 PP2.
German	active	Die Frau schnitzt das Kopfende des Bettes, das der Mann bemalt.	NP1$_{nom}$ VP1 NP2$_{acc}$ NP3$_{gen}$, WDT$_{acc}$ NP4$_{nom}$ VP2.
	passive	Die Frau schnitzt das Kopfende des Bettes, das von dem Mann bemalt wird.	NP1$_{nom}$ VP1 NP2$_{acc}$ NP3$_{gen}$, WDT$_{nom}$ PP1 VP2.
Turkish	active	Kadın adamın boyadığı yatağın başını oyuyor.	NP1$_{nom}$ NP2$_{gen}$ VP1(verb+adj/relativiser) NP3$_{gen}$ NP4$_{acc}$ VP2.
	passive	Kadın adam tarafından boyanan yatağın başını oyuyor.	NP1$_{nom}$ NP2$_{ablative}$ VP1(verb+verb+adj/relativiser) NP3$_{gen}$ NP4$_{acc}$ VP2.

Table 1: Exemplary sentence for ambiguity **A1)** for different languages in active as well as passive voice, including the respective part-of-speech (PoS) templates.

junctions. However, it does not take relative clause attachment ambiguities, which we are concerned with, into account. Also, it addresses only English. For human language processing, a recent study on the resolution of relative clause attachment ambiguities for different languages, e. g. English and German, can be found in (Hemforth et al., 2015), but, in general, the reference resolution in this case and the effect of its complexity in visually disambiguated situations addressing various languages have been scarcely investigated although ambiguities concerning relative clause attachments are quite common.

In the Hamburg Dependency Treebank (HDT) (Foth et al., 2014) part A, which contains $\approx$ 100k German sentences that were collected from the news website heise online, there are $13,256$ relative clauses that contain a relative pronoun that is supposed to have an antecedent and its reference resolution is ambiguous in $2,418$ cases (18.24%). While the nearest attachment has been chosen $1,907$ times (78.87%) in the HDT, an alternative respectively farther attachment occurs in 511 cases (21.13%). Therefore, we created a multi-modal data-set addressing these kinds of ambiguities among other things: the Linguistic Ambiguities in Situated Contexts (LASC) data-set (Alaçam et al., 2017; Alaçam et al., 2018). It contains challenging linguistic cases including ambiguous relative clause attachments and scope ambiguities for conjunctions as well as negations, which become fully unambiguous in the presence of visual stimuli. This work focuses on the relative clause attachment ambiguities. The multi-modal data, i. e. the syntactically ambiguous sentences and the corresponding scenes, are discussed in this section.

2.1. Linguistic Input

The LASC data-set (Alaçam et al., 2018) provides three types of fully ambiguous relative clause attachments, which are listed below.

A1) RPA[1] - a Genitive Modifier *(English, German, Turkish - active & passive voice)*
The woman carves the head of the bed, which the man paints.
Int. 1[2]: The man paints the bed. *(low attachment)*
Int. 2: The man paints the head of the bed. *(high attachment)*

A2) RPA - a Prepositional Phrase *(English, German)*
It is a mug on a coffee table, which she damages carelessly.
Int. 1: She damages the coffee table. *(low attachment)*
Int. 2: She damages the mug. *(high attachment)*

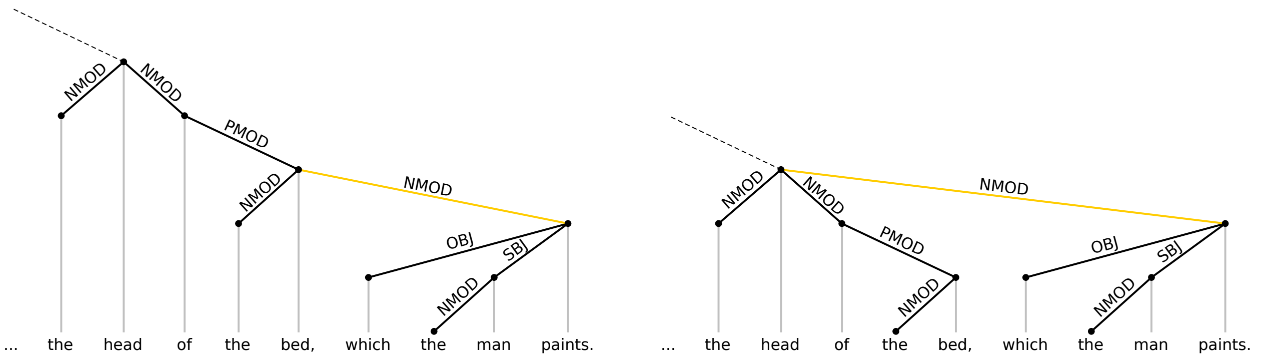
A3) RPA - Scope Ambiguities *(English, German)*
I see apples and bananas, which lie on the table.
Int. 1: Only bananas lie on the table. (low attachment)
Int. 2: Both apples and bananas lie on the table. (high attachment)

In ambiguity **A1)** and **A2)**, the relative clause is either attached to a preceding nominal phrase (NP) (high attachment) or to a genitive modifier respectively a PP of that NP (low attachment). In **A3)**, the relative clause either refers to the preceding conjunction of two NPs (high attachment) or to the latter NP only (low attachment). Syntactically, both the low, i. e. the nearest, and high, i. e. the farther, attachment are always possible in all examples. Hence, all sentences of this LASC subset are ambiguous. The contrary interpretations are equally distributed in this subset. Figure 1 shows the two plausible interpretations of the exemplary sentence of **A1)** by depicting parts of the respective dependency trees. In Figure 1a, the relative clause is low-attached to *bed* while it is high-attached to *head* in Figure 1b.

All examples are provided in English as well as German and with the relative clause in active voice. **A1)** is also available in Turkish and with passive instead of active voice. Table 1 shows the different configurations for the example from ambiguity **A1)** and, in addition, the part-of-speech (PoS) templates that are used to generate the sentences. There are corresponding templates for each type of ambiguity for each language respectively voice. The number of sentences per test set, depending on ambiguity type, language, voice and target interpretation, can be found in Table 2. Overall, this LASC subset contains 458 examples.

[1] RPA = Relative Pronoun Ambiguity

[2] Int. = Interpretation


NMOD
NMOD
PMOD
NMOD
NMOD
OBJ
SBJ
NMOD
... the head of the bed, which the man paints.
NMOD
NMOD
PMOD
NMOD
NMOD
OBJ
SBJ
NMOD
... the head of the bed, which the man paints.

(a) low attachment

(b) high attachment

Figure 1: Partial dependency trees for the two interpretations, i. e. the different attachments of the relative clause, of the exemplary sentence of ambiguity **A1)**.

2.2. Semantically Annotated Visual Scenes

NLP systems are often not able to correctly establish reference resolution in case of linguistic ambiguities like the ones that are described before in this section because they are based on linguistic information alone. Hence, they choose interpretations with respect to statistical distributions in their training data or explicitly stated rules, e. g. the data-driven parser that our system is based on (see Section 3.) originally prefers the low attachment for the German examples of ambiguity **A1)** (see Table 2) even if the high attachment is supposed to be chosen. But, respective visual scenes eliminate interpretations and favor the target one, assuming the scenes themselves are unambiguous.

There are different scenarios in the multi-modal LASC data-set, which involve several people, objects and actions. Each interpretation of a sentence belongs to one scenario and has a corresponding visual scene that visualizes the relations between agents and objects mentioned in the sentence and that supports the target interpretation. Figure 2 shows the images that belong to the different interpretations of the example of ambiguity **A1)**. In Figure 2a, the man paints the bed, which supports the low attachment of the relative clause seen in Figure 1a, while he paints the head of the bed in Figure 2b, which corresponds to the high attachment (Figure 1b).

Since we investigate the effect of external knowledge like visual scenes on language processing, information are not derived from the images automatically. Instead, the semantic annotations that are provided for each image from the LASC data-set are taken as external, contextual information. Those annotations were created with clear knowledge of the scenes, their Agents and Patients (Alaçam et al., 2018) in order to determine the upper bound of the performance of our computational model. Nevertheless, the images are part of the LASC data-set in order to conduct comparable studies with humans, which is not addressed here because it would exceed the scope of this work.

In parts, people, objects and actions in the images are manually annotated with semantic roles, similar to the approach of McRae et al. (2001), see also (Mayberry et al., 2006). Semantic roles are linguistic abstractions to distinguish and classify different functions of a predicate in a sentence, so they specify *"who did what to whom"*, and they establish a relation between the semantic and syntactic level of an analysis. The most common semantic roles include Agent, Theme, Patient, Instrument, Location, Source and Goal (Palmer et al., 2010). Figure 3 exemplarily shows some semantic roles for the images in Figure 2. There, the *man* is the Agent, who paints something, in both images. In Figure 3a, the *bed* is the Patient, the entity undergoing a change of state caused by the painting action, which supports the low attachment of the relative clause in Figure 1a, while the *head* is the Patient in Figure 3b reinforcing the high attachment in Figure 1b.

Both the sentences and the semantic annotations of the corresponding images serve as input to the system that is introduced in Section 3. and that enables multi-modal disambiguation based on sentences as linguistic and visual scenes as contextual information.

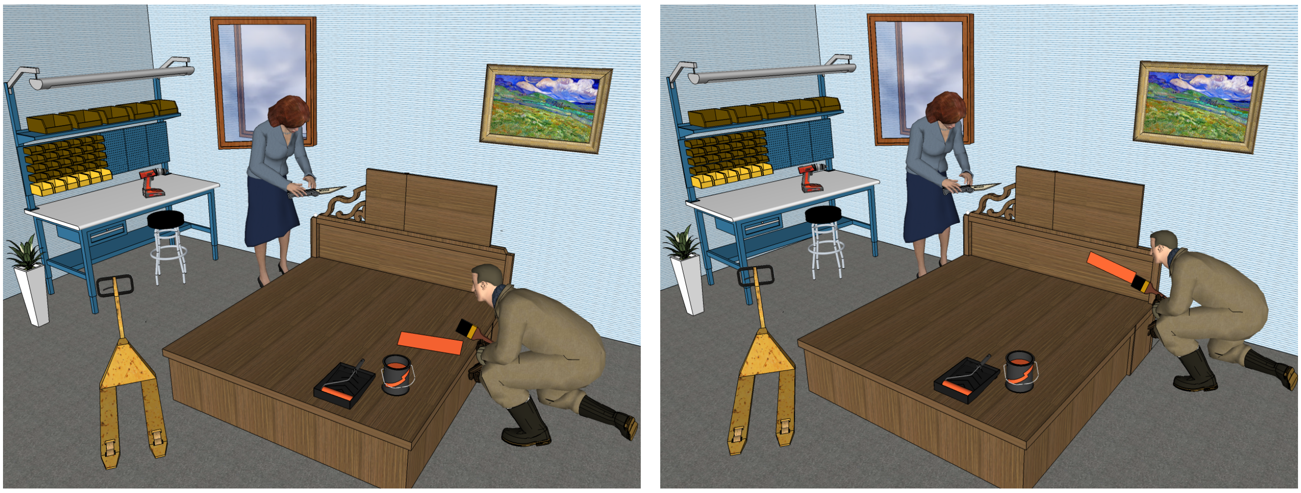
3. Multi-Modal Disambiguation

The previous section describes the multi-modal data-set that contains exemplary linguistic ambiguities whose disambiguation is investigated in this paper. Each example consists of a sentence and a respective visual scene, which is incompletely annotated with semantic roles, as contextual information. While most state-of-the-art parsers, which rely on the linguistic input only, are not able to resolve ambiguities reliably, this section introduces a parsing scheme that reaches correct disambiguation results by exploiting the contexts.

The final system has to meet the following requirements:

R1) Examples with sentences from several languages, i. e. English, German and Turkish, are part of the data-set. Thus, the system is supposed to be language-independent instead of -specific.

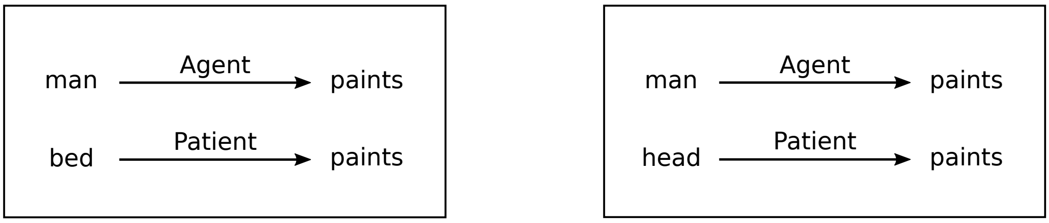
R2) The context is dynamic, i. e. the scenarios, which are displayed by the visual scenes, take place in different environments. Therefore, the contexts are highly dissimilar and the system has to account for this.



(a) The man paints the entire bed - corresponding to 1a.

(b) The man paints only the head of the bed - corresponding to 1b.

Figure 2: The visual scenes that correspond to the two interpretations of the exemplary sentence of ambiguity **A1)**, the dependency trees of which are partially displayed in Figure 1.



(a) corresponding to 2a

(b) corresponding to 2b

Figure 3: Some semantic role annotations for the scenes in Figure 2.

R3) The parsing results for the overall sentences, besides the investigated ambiguities, are supposed to be state-of-the-art results.

Subsection 3.1. addresses requirements R1) and R3) by basing the system on a state-of-the-art, data-driven parser. To deal with the dynamic nature of the context, requirement R2), a grammar-based approach is taken, which is being outlined in Subsection 3.2..

3.1. Language-Independent Parsing

Our system is based on the data-driven, syntactic RBG-Parser (RBG) (Zhang et al., 2014), which performs graph-based dependency parsing. It possesses a scoring function to evaluate entire dependency trees respectively their edges and learns that function based on training data that are annotated dependency trees from treebanks, which exist for various languages. Furthermore, RBG does not require language specific knowledge like hand-written grammars. Therefore, it fulfills the requirement of language-independence, i. e. R1). Also, requirement R3) is met because RBG achieves state-of-the-art results for several languages (Zhang et al., 2014).

RBG extracts up to third-order local features, like sibling or grand-grandparent structures, as well as global features, e. g. span lengths, from input sentences. For a complete list of all possible features, see (Zhang et al., 2014). Hill climbing is applied if all features are exploited to approximate the most plausible dependency tree. First, a random tree is uniformly sampled. Next, the heads of all dependents are exchanged so the edges of the tree are changed until a local optimum is reached. To increase the likelihood of finding the global optimum, hill climbing repeatedly restarts, always with random trees that are sampled independently of previous solutions, until the best solution converges. During hill climbing, edges as well as entire trees are scored by RBG's scoring function, see (Zhang et al., 2014; Lei et al., 2014) for details.

3.2. Inclusion of Dynamic Contexts

While RBG fulfills the requirements of language-independence and state-of-the-art results, it is not able to deal with dynamic contexts, i. e. requirement R2). In theory, features could be extracted from the contexts, for example features capturing the relations the semantic roles, which the visual scenes are annotated with in our data-set, express. Either an RBG model might be trained on those features combined with the ones extracted from the input sentences or separate models might be trained. In both cases, the disambiguation results for the unseen test data might deteriorate and the linguistic ambiguities, which are investigated in this paper, might not be resolved because the contexts of the test data would differ highly from the contexts of the training data. Instead, a grammar-based approach utilizes the contextual information to improve the disambiguation results.

First, the semantic roles the visual scenes are annotated with are grounded. So, the contextual information is aligned with the linguistic one. In Figure 3a, *man* is the Agent of *paints* while *bed* is its Patient. All three instances are grounded by connecting them to the respective words of the input sentence. To focus the effects of multi-modal disambiguation and to determine its upper bounds, grounding is simply based on lexical agreement to exclude possible errors. In future experiments, the lexical-based grounding will be replaced with a concept-based one. An ontology will contain a conceptional hierarchy and each concept will provide a multitude of possible lexicalizations, e. g. singular and plural forms for different cases. This way, instances of the semantic roles and words of the input sentences can differ conceptually, e. g. *chair* occurring in a sentence may refer to *piece of furniture* in the context.

In order to incorporate contextual information into the disambiguation process, our system employs jwcdg (Beuck et al., 2013), which is a graph-based dependency parser. As opposed to RBG, it is grammar-based. It possess the ability to only evaluate dependency trees respectively their edges with respect to its grammar in addition to actually parsing sentences. Possible grammars contain two types of constraints, hard ones, which are not allowed to be violated, and soft constraints, which may be violated. The result of the evaluation is a score between 0, at least one hard constraint is violated, and 1, no constraints are violated at all.

Originally, there is a jwcdg grammar available for German only. Thus, jwcdg does not fulfill the language-independence requirement. But, our system uses jwcdg only for evaluating dependency trees with respect to the context so that the original grammar is not required. Instead, a new grammar has been constructed that links semantic roles with their respective syntactic structures. For example, if the predicate of a semantic role refers to the main verb of a relative clause, which is in active voice, and if the Patient of that predicate is known and grounded as well, that relative clause is supposed to be attached to that Patient. In the example of Figures 1a and 3a, both *paints* and *bed* can be grounded, the former referring to the verb of the relative clause, which is in active voice. Thus, the relative clause is supposed to be attached to *bed*. Our linking grammar only covers semantic roles and their respective syntactic structures relevant with respect to the data-set (Section 2.) used in the experiments (Section 4.). While individual grammars have been used for the different languages in those experiments, those grammars are not lexicalized and there is an overlap of the constraints, e. g. comparable rules have been used for German and Turkish or the rules for English and German do not contradict each other, which suggests that a single grammar is able to cover several languages. Therefore, our system maintains its language-independence.

The score jwcdg determines by evaluating a dependency tree that is passed on to it by RBG is combined with the respective RBG score, which is normalized first because it has a different domain than jwcdg. Min-Max normalization (Priddy and Keller, 2005) is applied to map all RBG scores to the range of 0 to 1. After the normalization, the inverse jwcdg score, interpreted as a penalty, is subtracted from the RBG score. In case neither the dependent nor the head of an edge of a dependency tree refer to any instance of the context, that edge will not be able to violate any constraints. The same will hold true for an entire tree if none of its dependents or heads refer to the context. In those cases as well as in case references are established between the input sentence and the context but no constraints are violated, a score of 1 will be returned by jwcdg, with its inverse score being 0, and the RBG score remains unchanged. Otherwise it is decreased by the inverse jwcdg score.

Figure 4 visualizes this process. While RBG performs hill climbing for the linguistic input $x_{linguistic}$, its scoring function is repeatedly called to evaluate entire trees $t_{current}$ respectively their individual edges. jwcdg is called to evaluate whether the semantic roles $x_{semanticAnnotations}$ the corresponding visual scenes are annotated with can be linked to $t_{current}$ via the $linking\ grammar$, which links semantic roles to syntactic structures. The jwcdg score s_{jwcdg} is converted into a penalty by taking its inverse and this penalty is subtracted from the normalized RBG score $norm(s_{RBG})$. The combined scores $s_{RBG\&jwcdg}$ are returned so that the external knowledge guides the hill climbing. The best dependency tree that is found during hill climbing is returned as final solution t_{best}.

Instead of being trained on the context, our system employs a grammar-based approach to link linguistic with contextual information in order to incorporate previously unseen, i. e. dynamic, context. Therefore, our system meets requirement R2). The following sections evaluate the experiments and analyze the results.

4. Evaluation

This section presents the evaluation results of parsing linguistically ambiguous sentences both without and with exploiting contextual information. The original RBG parser, which relies on linguistic input only, is compared to our system, which performs multi-modal disambiguation and is outlined in Section 3.. The exemplary linguistic ambiguities are described in Section 2.. Three cases of relative clause attachments have been selected from the LASC data-set (Alaçam et al., 2018) that are ambiguous both in English and German. The case of attaching a relative clause to an NP or its genitive object is also ambiguous in Turkish. On the syntactic level, there are two possible antecedents for each relative clause and its relative pronoun agrees with both in number and gender. Therefore, a disambiguation is not possible without further evidence.

For the experiments, full RBG models, which use all available RBG features, i. e. global as well as up to third-order local features, have been trained for English, German and Turkish, respectively. For English, sections $0-22$ and 24 of the Wall Street Journal (WSJ) corpus of the Penn Treebank (Marcus et al., 1994) constitute the training set (≈ 46k sentences, duplicates excluded). They have been converted to dependency structure using the LTH converter (Johansson and Nugues, 2007). For German, the first ≈ 98k sen-

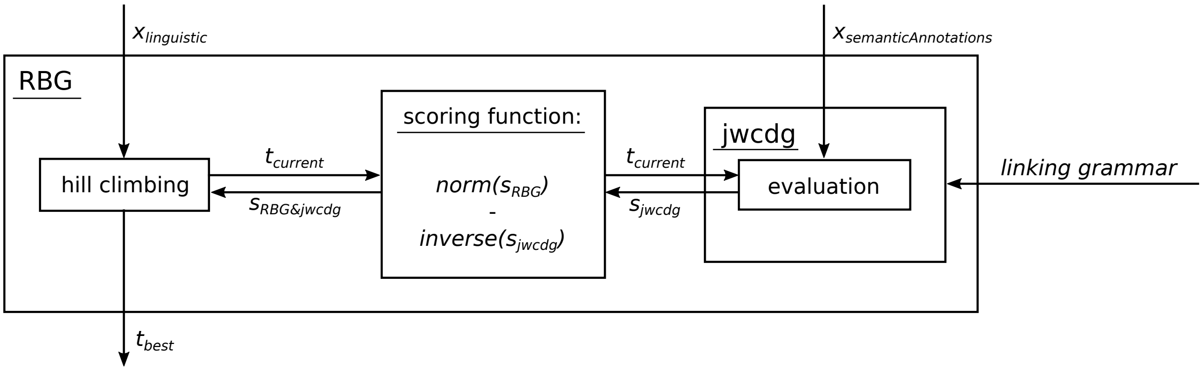


Figure 4: RBG applies hill climbing for the linguistic input $x_{linguistic}$, its scoring function is repeatedly called to evaluate entire trees $t_{current}$ or their edges. jwcdg is called to evaluate whether the semantic roles $x_{semanticAnnotations}$ from the context can be linked to $t_{current}$ via the $linking\ grammar$. The normalized RBG score $norm(s_{RBG})$ is penalized by the inverse jwcdg score s_{jwcdg} and the combined score $s_{RBG+jwcdg}$ is returned. The best dependency tree found during hill climbing is returned as final solution t_{best}.

tences (duplicates excluded) of the HDT part A are used for training. For Turkish, the training data are the first 5k sentences of the METU-Sabanci Turkish Dependency Treebank (Oflazer et al., 2003). If a word from that corpus is separated into different inflectional groups, they will be concatenated as described in (Nivre et al., 2007) so that parsing is performed on word basis. All three corpora provide word forms, gold PoS tags and gold standard annotations. While the gold PoS tags have been used for Turkish, all German and English sentences have been PoS tagged by the TurboTagger (Martins et al., 2013). Ten-way jackknifing has been performed for the training sets, i. e. each is split into ten partitions and each partition is tagged by a model trained on the other nine partitions. The test sets have been PoS tagged by models trained on the entire training set of the respective corpus.

Table 2 lists all experimental results. The accuracies of the relative clause attachments predicted by the original RBG are compared to the multi-modal disambiguations of our system. In general, RBG always favors one attachment, mostly the low one, irrespective of the supposed attachment because it does not have access to the contextual information. In contrast, our system, which exploits that external knowledge, always predicts the correct attachment.

Several observations can be made. First, while RBG mostly favors low attachments, it will predict high ones instead for ambiguities for some English examples if the relative clause is in active voice. Secondly, RBG makes some unusual predictions, sometimes attaching relative clauses to improbable antecedents, for English and Turkish. Section 5. discusses these results.

5. Analysis

This section analyzes the evaluation results, which are described in Section 4. and listed in Table 2. First, the fact that RBG always prefers either the low or the high attachment in case there are two possible antecedents for a relative pronoun is due to statistical distributions in the data the data-driven parser has been trained on. In German sentences, relative clauses are more frequently low-attached than high-, although there are examples for both. In the HDT part A, the nearest plausible attachment is chosen in 78.87 percent of all cases that resemble the examples of our experimental data, i. e. relative pronouns with ambiguous antecedents (see Section 2.). Comparable observations can be made for Turkish and it will also hold for English in case of ambiguity **A1)**. For **A2)** and **A3)**, the opposite can be observed. They are more often high- than low-attached in English. In those cases, RBG prefers the high attachment.

The disambiguation results of RBG do not match the intended interpretations, with respect to the corresponding visual scenes, because it does not take any external knowledge into account. In contrast, our system, which incorporates contextual information, always resolves the linguistic ambiguities in the experimental data correctly without exception. This result proves our hypothesis H1) that the disambiguation of linguistic ambiguities will improve if it is not only based on analyzing the linguistic input but also takes external knowledge respectively contextual information, like the semantically annotated visual scenes in this work, into account.

Furthermore, the evaluation shows that our system fulfills the requirements R1) - R3). It correctly resolves ambiguities independent of the languages although those show some major differences. As discussed before, in case of ambiguity **A2)** and **A3)**, the high attachment of the relative clause is preferred in English compared to the low one in German. Also, relative clauses usually appear after their possible antecedents in English as well as German while they precede them in Turkish. Neither of those properties influence the ability of our system to correctly resolve linguistic ambiguities. Therefore, it is language-independent, which validates hypothesis H2). Furthermore, the contexts respectively the visual scenes differ significantly, i. e. they are dynamic. Nevertheless, our system achieves correct disambiguations, which proves that it is able to deal with dynamic contexts. In addition, it is based on RBG, which is a parser that enables state-of-the-art results. Therefore, all

					# Relative Clause Attachments					
					RBG			Multi-Modal Disambiguation		
Ambiguity Type	Language	Voice	Supposed Attachment	# Examples	low	high	misc	low	high	misc
A1) RPA - a Genitive Modifier	English	active	low	24	8	16	-	24	-	-
			high	24	8	16	-	-	24	-
		passive	low	24	24	-	-	24	-	-
			high	24	24	-	-	-	24	-
	German	active	low	24	24	-	-	24	-	-
			high	24	24	-	-	-	24	-
		passive	low	24	24	-	-	24	-	-
			high	24	24	-	-	-	24	-
	Turkish	active	low	24	19	-	5	24	-	-
			high	24	19	-	5	-	24	-
		passive	low	24	19	-	5	24	-	-
			high	24	19	-	5	-	24	-
A2) RPA - a Prepositional Phrase	English	active	low	20	5	14	1	20	-	-
			high	17	3	13	1	-	17	-
	German	active	low	20	18	-	2	20	-	-
			high	17	16	-	1	-	17	-
A3) RPA - Scope Ambiguities	English	active	low	24	1	23	-	24	-	-
			high	24	1	23	-	-	24	-
	German	active	low	24	24	-	-	24	-	-
			high	24	24	-	-	-	24	-

Table 2: Comparison of the evaluation results of the original RBG and our multi-modal disambiguation regarding the ambiguous attachment of the relative clauses for the ambiguities **A1)** - **A3)** in different configurations (language, voice of the finite verb of the relative clause).

requirements are fulfilled.

RBG sometimes predicts attachments that do not correspond to the ones it generally favors. For example, it predicts the high attachment 16 times for the English version of **A1)** with the relative clause in active voice although it favors the low attachment. This is due to wrongly assigned PoS tags. If the gold PoS tags are used, the low attachment is consistently predicted. The same holds true for the German examples of **A2)**, in which RBG assigns alternative, improbable heads to the relative clauses. For Turkish, RBG always chooses the low attachment, but, in case of compound nouns, it refers to their first part although it is supposed to choose the second part.

6. Related Work

In this work, the contextual information stems from visual scenes and helps to disambiguate linguistic ambiguities. A similar effect can be observed in human language processing. Tanenhaus and his colleagues (Tanenhaus et al., 1995) showed that visual information influences how humans disambiguate linguistic input. While the authors of that study focused on PP attachment ambiguities, we used the problem of attaching relative clauses because it has been less frequently investigated than the former. Further evidence that supports the conclusion of (Tanenhaus et al., 1995) was provided by Knoeferle (Knoeferle, 2005), whose work also indicates that visual information influences language processing independent from the experiment language, she conducted experiments for English and German. In addition to those languages, some of our tests were also carried out for Turkish. Furthermore, Coco and Keller (Coco and Keller, 2015) investigated the interaction between language and vision and its influences on the interpretation of syntactically ambiguous sentences in a simple real-world setting. Their study provided further evidence that not only linguistic but also visual information influences the interpretation of a sentence. While the aforementioned empirical studies were psycho-linguistically motivated and provided insights how humans resolve linguistic ambiguities by exploiting visual cues, our work provides evidence that similar effects can be observed for automatically parsing syntactically ambiguous sentences in the presence of contextual information.

An early approach to exploit external knowledge during parsing was proposed by (McCrae and Menzel, 2007; McCrae, 2009; Baumgärtner et al., 2012). They suggested to utilize high-level representations of visual information from related scenes to resolve linguistic ambiguities in German, e.g. Genitive-Dative ambiguity of feminine nouns or PP attachment, and developed a syntactic parsing architecture for the integration of cross-modal information. Moreover, McCrae (2009) hypothesized multiple requirements for such a system. Like ours, their system uses visual scenes as context and the author did not discuss the automatic derivation of information from those scenes. Also, the instances of the context are mapped onto the words of the corresponding sentence and linked via semantic relations to the syntactic structure by applying constraints. But, that system is solely constraint-based and does no longer produce state-of-the-art results. Furthermore, it requires a complete grammar of the respective language it is applied to. Thus, that approach is language specific. In contrast, our system is based on a data-driven parser, which achieves state-of-the-art results and whose model is trained on annotated data that is provided by treebanks for various languages and which is, thus, language-independent.

In a more recent work, Christie et al. (2016) proposed to jointly segment an input image semantically and resolve PP attachment ambiguities in its caption. Their approach generates several hypotheses for both segmentation and disambiguation and a model scores pairs of hypotheses for both tasks in order to find the most plausible hypotheses for both the visual and linguistic task. Compared to our work, they tried to improve not only the disambiguation but also the processing of the visual input. Furthermore, they do not assume the visual input to be perfect. Instead, we are interested in ambiguous attachments of relative clauses and, while Christie et al. (2016) employed a two-staged approach, i. e. reranking, the visual information guides the disambiguation process in this work, so the different modalities are employed at the same time, not sequentially.

7. Conclusion

This work addresses the hypotheses that linguistic ambiguities can be resolved if contextual information derived from additional modalities besides the linguistic one are exploited, independent of the language. In order to test these hypotheses, a data-driven, syntactic parser has been modified by repeatedly calling a grammar-based parser that evaluates current analyses with respect to the context, which consists of semantically annotated visual scenes. A grammar that contains constraints to link those semantic roles to the respective syntactic structures has been constructed so that the context guides the data-driven parser towards the most plausible solution given both the input sentence and the corresponding visual scene, i. e. the context.

Our hypotheses have been evaluated for several examples of relative clause attachment ambiguities under varying conditions, e. g. active versus passive voice in the relative clause, and for multiple languages, namely English, German and Turkish. While the original data-driven parser, which does not use any external knowledge, has not reached the intended interpretations consistently, instead it always preferred either low or high attachments, our system disambiguates the examples correctly, which is evidence of the hypotheses being true. Furthermore, the resulting system fulfills several requirements it is subjected to. It proofed to be language-independent and able to deal with dynamic context, e. g. highly dissimilar visual scenes.

For the future, our hypotheses will be verified for Chinese, too. Also, the influence of possible error sources will be evaluated, e. g. spelling mistakes or erroneous context that match none of the possible interpretations of the respective sentence. In addition, these images together with the corresponding sentences are going to be employed for comparative studies with humans to enable a comparison between human and machine disambiguation abilities.

Acknowledgments

This research was funded by the German Research Foundation (DFG) in the project Crossmodal Learning, TRR-169.

8. Bibliographical References

Alaçam, Ö., Staron, T., and Menzel, W. (2017). Towards a systematic analysis of linguistic and visual complexity in disambiguation and structural prediction. In *Proceedings of the 5th International Conference on Information Management and Big Data*, Lima, PERU, September.

Alaçam, Ö., Staron, T., and Menzel, W. (2018). A multimodal data-set for systematic analyses of linguistic ambiguities in situated contexts. In Juan Antonio Lossio-Ventura et al., editors, *Information Management and Big Data*, Communications in Computer and Information Science. Springer. In press.

Altmann, G. T. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Baumgärtner, C., Beuck, N., and Menzel, W. (2012). An architecture for incremental information fusion of cross-modal representations. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 498–503, Hamburg, Germany, September. IEEE.

Berzak, Y., Barbu, A., Harari, D., Katz, B., and Ullman, S. (2016). Do you see what i mean? visual resolution of linguistic ambiguities. *arXiv preprint arXiv:1603.08079*.

Beuck, N., Köhn, A., and Menzel, W. (2013). Predictive incremental parsing and its evaluation. In *Computational Dependency Theory*, volume 258 of *Frontiers in Artificial Intelligence and Applications*, pages 186 – 206. IOS press.

Christie, G., Laddha, A., Agrawal, A., Antol, S., Goyal, Y., Kochersberger, K., and Batra, D. (2016). Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. arXiv preprint arXiv:1604.02125, September.

Coco, M. I. and Keller, F. (2015). The interaction of visual and linguistic saliency during syntactic ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, 68(1):46–74.

Foth, K. A., Köhn, A., Beuck, N., and Menzel, W. (2014). Because size does matter: The Hamburg Dependency Treebank. In *Proceedings of the Language Resources and Evaluation Conference 2014*, Reykjavik, Iceland, may. LREC.

Hemforth, B., Fernandez, S., Clifton, C., Frazier, L., Konieczny, L., and Walter, M. (2015). Relative clause attachment in german, english, spanish and french: Effects of position and length. *Lingua*, 166:43–64.

Johansson, R. and Nugues, P. (2007). Extended constituent-to-dependency conversion for english. In *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia, May.

Knoeferle, P. S. (2005). *The role of visual scenes in spoken language comprehension: Evidence from eye-tracking*. Ph.D. thesis, Universitätsbibliothek.

Lei, T., Xin, Y., Zhang, Y., Barzilay, R., and Jaakkola, T. (2014). Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391, Baltimore, Maryland, June. Association for Computational Linguistics.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, March.

Martins, A., Almeida, M., and Smith, N. A. (2013). Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August.

Mayberry, M., Crocker, M. W., and Knoeferle, P. (2006). A connectionist model of the coordinated interplay of scene, utterance, and world knowledge. In *Proceedings of the 28th annual conference of the Cognitive Science Society*, pages 567–572.

McCrae, P. and Menzel, W. (2007). Towards a system architecture for integrating cross-modal context in syntactic disambiguation. In *Proceedings of the 4th International Workshop on Natural Language Processing and Cognitive Science, Madeira (NLPCS 2007)*, pages 228–237.

McCrae, P. (2009). A model for the cross-modal influence of visual context upon language procesing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2009)*, pages 230–235, Borovets, Bulgaria, September.

McRae, K., Hare, M., Ferretti, T., and Elman, J. L. (2001). Activating verbs from typical agents, patients, instruments, and locations via event schemas. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 617–622. Erlbaum Mahwah, NJ.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Oflazer, K., Say, B., Hakkani-Tür, D. Z., and Tür, G. (2003). Building a Turkish treebank. *Treebanks*, pages 261–277.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Priddy, K. L. and Keller, P. E. (2005). *Artificial Neural Networks: An Introduction*, volume 68. SPIE press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632.

Zhang, Y., Lei, T., Barzilay, R., Jaakkola, T., and Globerson, A. (2014). Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Baltimore, Maryland, June. Association for Computational Linguistics.