Artif Intell Rev (2006) 25:21–35
DOI 10.1007/s10462-007-9022-9

Attention driven reference resolution in multimodal contexts

J. D. Kelleher

Published online: 25 August 2007

Abstract In recent years a number of psycholinguistic experiments have pointed to the interaction between language and vision. In particular, the interaction between visual attention and linguistic reference. In parallel with this, several theories of discourse have attempted to provide an account of the relationship between types of referential expressions on the one hand and the degree of mental activation on the other. Building on both of these traditions, this paper describes an attention based approach to visually situated reference resolution. The framework uses the relationship between referential form and preferred mode of interpretation as a basis for a weighted integration of linguistic and visual attention scores for each entity in the multimodal context. The resulting integrated attention scores are then used to rank the candidate referents during the resolution process, with the candidate scoring the highest selected as the referent. One advantage of this approach is that the resolution process occurs within the full multimodal context, in so far as the referent is selected from a full list of the objects in the multimodal context. As a result situations where the intended target of the reference is erroneously excluded, due to an individual assumption within the resolution process, are avoided. Moreover, the system can recognise situations where attention cues from different modalities make a reference potentially ambiguous.

Keywords Reference resolution · Situated dialog · Attention · Salience · Vision and language · Natural language processing

1 Introduction

Many modern natural language processing applications (human-robot collaboration, computer games, navigation aids etc.) share a visualised space with the user. In these applications the user interacts with the system using *situated language*. Situated language is spoken from a particular point of view within a physical or simulated context. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution.

J. D. Kelleher (✉)
School of Computing, Dublin Institute of Technology, Dublin 8, Ireland
e-mail: john.kelleher@comp.dit.ie

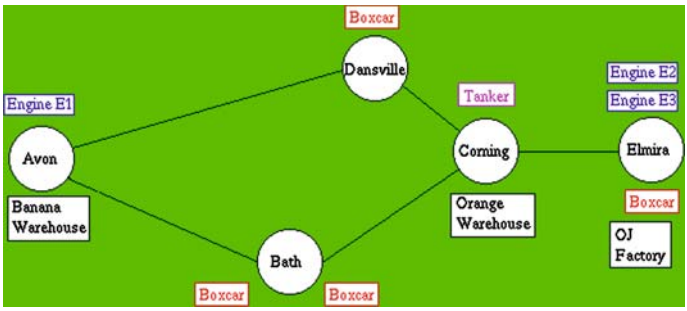

Fig. 1 Map of TRAINS domain

A referring expression is a natural language expression that denotes an entity called a *referent*. Referring expressions come in a variety of *forms* including: definite descriptions, indefinites, pronouns, demonstratives. Each referring expression introduces a representation into the semantics of its utterance and this representation must be bound to an element in the context for the utterance's semantics to be fully resolved. Consequently, from a computational perspective reference resolution involves two main tasks:

1. creating and maintaining a model of the discourse context (this model should contain representations for all the entities that are available for reference)
2. matching/binding the representation introduced by a given referring expression to an element (or elements) in the set of possible referents

Most forms of referring expression have a preferred *mode of interpretation*: *anaphoric, exophoric*, etc. In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Referring expressions that access a representation in the linguistic context are interpreted anaphorically. However, in a situated dialog human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. Referring expressions that access a representation of an object that has not previously been referred to in the dialog but has entered the context through a non-linguistic modality (such as vision) are interpreted exophorically.

The following dialog excerpt, taken from the TRAINS-93 corpus (Allen and Schubert 1991), illustrates the distinction between anaphoric and exophoric references. The excerpt is taken from a collaborative dialog between two participants, S1 and S2, who are trying to ship goods within a railroad freight system. Figure 1 illustrates the schematic representation of the railroad freight system that provided the visual context for the dialog. In this example, the indices i, j, k and l indicate that all the referring expressions marked by a particular index refer to the same entity.

1. **Visual context**: See Fig. 1.

S1.1 "aha ... I see *an engine*$_i$ and *a boxcar*$_j$ both at *Elmira*$_k$"
S2.1 "right"
S1.2 "this looks like the best thing to do ... so we should get *the engine*$_i$ to pick up *the boxcar*$_j$ and head for *Corning*$_l$ ... that sound reasonable"
S2.2 "sure ... that sounds good"

The references *an engine*, *a boxcar*, and *Elmira* in S1.1 and *Corning* in S1.2 are examples of exophoric references. The entities these expressions denote have not been previously

mentioned in the dialog. As a result, these reference must be resolved relative to a set of entity representations in the context model that entered the model via the non-linguistic modalities, in this instance the visual context of the dialog. By contrast, the references *the engine* and *the boxcar* in S1.2 are examples of anaphoric references. The reference *the engine* can be resolved relative to the linguistic context by binding it to the representation of *an engine* introduced to the linguistic context by the resolution of S1.1. Similarly, the reference *the boxcar* can be resolved relative to the linguistic context by binding it to the representation of *a boxcar* introduced by the resolution of S1.1.

However, there is no one-to-one relationship between form and mode of interpretation. For example, definite descriptions can be used either anaphorically or exophorically. Indeed, the two most common cases of definite descriptions in the TRAINS corpus of situated dialogue were anaphoric and exophoric definites (Poesio 1993). One consequence of the one-to-many relationship between referential form and mode of interpretation, is that a multimodal reference resolution process should define a strategy to deal with cases where different mode of interpretations are suggested for the same reference. One solution, to this issue, is to define a preference ordering over the different interpretation rules. A second alternative is to use a probabilistic approach, where each interpretation of a reference is assigned a probability score that is used to rank the interpretations. The approach to reference resolution developed in this paper adopts a probabilistic solution to this issue. The framework uses the relationship between referential form and preferred mode of interpretation as a basis for a weighted integration of linguistic and visual attention scores for each entity in the context model. The resulting integrated attention scores are then used to rank the candidate referents during the resolution process, with the candidate scoring the highest selected as the referent.

One advantage of this approach is that resolution process occurs within the full multimodal context of the dialog, in so far as the the referent is selected from a full list of the objects in the multimodal context ordered by a model of integrated salience. Consequently, none of the objects in the context are excluded from consideration. As a result situations where the intended target of the reference is erroneously excluded, due to an individual assumption within the resolution process, are avoided. Also, the framework can recognise cross-modal ambiguity by comparing the integrated salience of the primary candidate with the integrated salience of all the other objects in the context. In these ambiguous cases the initiation of a clarification dialog may be a better system response rather than the selection of the primary candidate referent. By contrast, many of the previous multimodal resolution frameworks exclude entities in the multimodal context model from consideration before the selection of the referent. In some cases, for example Kievit et al. (2001), Salmon-Alt and Romary (2001), Landragin and Romary (2003), Kelleher et al. (2005), the initial set of candidates referents is restricted to a sub-set of the context based on preferences with respect to the mode of interpretation relative to the form of reference. In other frameworks, for example Gorniak and Roy (2004), candidate referents are incrementally excluded from consideration as the resolution process progresses due to the sequential manner that the semantics of the terms within the reference are processed.

Moreover, from a functional perspective this approach has the advantage of modularity and the potential to accommodate learning within the system. The modularity of the framework stems from the fact that the only information required by the resolution process from each of the information sources (language and vision) within the context are the attention scores for each entity. As a result, the resolution process is, to a large extent, decoupled from the representations and processes used within the linguistic and visual context models. The learning aspect of the system arises from the ease (relative to rule based approaches) with which the integration weightings associated with a particular form of reference could be updated, for

example, using machine learning techniques such as reinforcement learning. Finally, from a cognitive perspective, an attention based model fits the theoretical and psychological data that points to the role of attention within human reference resolution. Grosz (1977) was first in observing the relationship between focus of attention and exophoric definite descriptions. More recently, psycholinguistic studies, such as Duwe and Strohner (1997), have shown that people often use perceptual salience to resolve linguistic references.

The paper is structured as follows: §2 reviews related work; §3 presents the data structures and algorithms used in the framework; §4 contains a worked example illustrating the functioning of the framework; the paper finishes, in §5, with conclusions.

2 Related work

Grosz (1977) is arguably the seminal work on language and vision integration. This work highlighted that attention constrained and structured the processing of discourse. Moreover, Grosz was the first to observe the relationship between focus of attention and the use of exophoric definite descriptions: when an object is in the current mutual focus of attention it can be referred to by means of a definite description even though other objects fulfilling the description have been introduced into the linguistic discourse or are present in the shared visible context.

Building on this work, Grosz and Sidner (1986) developed a focus stack model of global discourse attentional state. According to this model the common ground[1] can be divided into three parts: *the linguistic structure*, which contains information about the linguistic structure of utterances in the dialog; *the intentional structure*, which contains information about the goals of the participants in the conversation; and *the attentional structure*, which contains information about the objects introduced into the discourse and their relative salience. Furthermore, due to attentional constraints, discourse is segmented or chunked and when a definite description is used anaphorically, the only *antecedents*[2] considered are those in the same discourse segment.

Assuming Grosz and Sidner's (1986) focus stack model to be generally correct as a model of global discourse structure,[3] the issue of how focus of attention and reference interact within a discourse segment must still be addressed. Several frameworks have been proposed, for example Alshawi (1987), Hajicová (1993), Lappin and Leass (1994) and Grosz et al. (1995).[4] However, none of these models explicitly accommodate multimodal contexts.

Poesio (1993) reformulates the attentional model in Grosz and Sidner (1986) in situation theoretic terms. Interestingly, Poesio's framework separates the attentional common ground into several *anchoring resource situations*. For example, one anchoring resource is called the *discourse situation* and consists of a record of what has been said. This anchoring resource is used to interpret anaphoric references. Another anchoring resource situation called the *situation of attention* models the subset of information in the visual field of the discourse participants that they are attending to and is used to interpret exophoric[5] definite descriptions. Furthermore, he defines rules within a default logic, called *principles for anchoring resource*

[1] The dialog participants mutually developed public view of what they are talking about.

[2] The antecedent of an anaphoric reference is the representation of the reference's referent that was introduced to the discourse model by a prior referring expression.

[3] For alternate models see Hobbs (1985), Mann and Thompson (1987) and Asher and Lascarides (2003).

[4] See Kruijff-Korbayová and Hajicová (1997) for a comparison of these approaches.

[5] Poesio uses the term *visible situation use* to describe to exophoric definite descriptions.

situations, that predict whether a definite description is going to be interpreted anaphorically or exophorically. However, one of the issues with this approach is how to deal with conflicting defaults. Consequently, the framework cannot handle situations in which two principles of anchoring resource situations apply, one suggesting an anaphoric interpretation the other an exophoric interpretation.

Many computational frameworks for multimodal reference resolution have also been developed. McKevitt (1996) provides an excellent collection of papers on early systems. Recent systems that focus on multimodal reference resolution include: Kievit et al. (2001), Salmon-Alt and Romary (2001), Landragin and Romary (2003), Gorniak and Roy (2004) and Kelleher et al. (2005).

Kievit et al. (2001) define separate resolution strategies for each form of referring expression. A strategy consists of one or more resolution steps applied in a predefined order. A resolution step consists of 4 stages: (1) the selection of possible referents from a single sub-context (dialog, visual domain, etc.), (2) the filtering of this set of candidates, (3) the ordering of the candidates based on saliency, (4) an evaluation of the result. The algorithm halts as soon as one of the resolution steps finds a unique object or finds several objects and cannot choose which is the intended one. This approach is equivalent to a preference ordering being defined over the different modes of interpretation for each form of reference. One issue with this approach is that the set of candidates considered during any one resolution step is constrained to the set of entities within the sub-context the resolution step uses to construct the initial set of candidates. As a result, the system cannot recognise situations where a reference may be ambiguous between two entities in different sub-contexts, and, consequently, it may resolve a reference incorrectly rather than initiate a clarification process.

Gorniak and Roy (2004) focus on the resolution of references containing spatial descriptions. They propose a feed-forward filtering process to reference resolution. In their framework, each lexical item in the system's lexicon is associated with one or more *composer functions*. A composer function takes one or more candidate referents as input and filters this set of candidates by computing how well each of the candidates fulfills the semantic model defined for the lexical term. Reference resolution is carried out by chaining the composer functions associated with the lexical terms in the reference together, i.e. the filtered set of candidates output by one composer function is used as the input set by the next composer function in the chain. Gorniak and Roy note that this strategy can fail if one of the composer functions excludes the target object from the set of candidates. For example, when interpreting "the leftmost one in the front" the composer for "leftmost" selects the leftmost objects in the scene, not including the obvious example of "font" that is not a good example of "leftmost".

The reference resolution frameworks presented in Salmon-Alt and Romary (2001), Landragin and Romary (2003) and Kelleher et al. (2005) use the notion of a reference domain. A reference domain is a structured contextual subset of the multimodal dialog context. Reference domains are created in the context model due to perceptual or linguistic events or conceptual knowledge and are intended to reflect the mental representation of the event they model. In these frameworks the resolution process involves: (1) the construction of an underspecified reference domain, using templates associated with the form of the reference given; (2) the unification of this underspecified domain with a suitable reference domain within the context model; (3) the selection of one of the elements within the unified reference domain to function as the referent. However, similar to the frameworks proposed in Kievit et al. (2001) and Gorniak and Roy (2004), there is the potential for these frameworks to overcommit to a particular subset of the context during the resolution process. As the resolution process occurs within a sub-context, whose selection is at least partially driven by the form of the reference

being interpreted, if the wrong reference domain is selected the intended target object and/or plausible distractor referents, that may indicate the need for reference clarification, may be excluded from consideration.

3 Approach

Several theories of discourse reference have attempted to provide an account of the relationship between types of referential expressions on the one hand, and degrees of mental activation of discourse referents on the other (e.g. Ariel 1990; Gundel et al. 1993; Grosz et al. 1995). A common theme among these accounts is that referential expressions need more coding material as the referent is less activated. Following these theories, the basic approach of the framework is to treat a given referring expression as a set of instructions that specifies how the spread of attention across the set of objects within the discourse context should be modified before the selection of the referent. Consequently, the concept of attention is at the core of the framework.

Studies of attention, for example Enns and Rensink (1990), Spivey-Knowlton et al. (1998), Hopfinger et al. (2000), Chum and Wolfe (2001), indicate that both bottom-up and top-down processes affect it. Bottom-up processing guides attention based on low-level perceptual cues. Top-down processing, driven by factors such as intention, also affect attention. Indeed, the results of several eye-tracking experiments, for example Yarbus (1967), Spivey-Knowlton et al. (1998), Tanenhaus et al. (1995), indicate that language comprehension is one of the top-down processes affecting visual attention.

A concept closely related to attention is salience. In this paper salience is used to describe the factors and associated processes that direct attention. The framework distinguishes between three levels of salience.

Level 1 This level includes the basic visual salience (i.e. the prominence of an object due to bottom up visual cues) and linguistic salience (i.e. the prominence of an object due to previous discourse) of an object.

Level 2 This level consists of reference relative visual and linguistic salience. These saliency scores represent the salience of an object within each of the modalities within the context provided by the referring expression that is being resolved.

Level 3 This level represents the integrated salience of objects within the context provided by the referring expression that is being resolved. This is computed using a weighted combination of the object's level 2 salience scores. The weights used in this combination reflect the biasing associated with different forms of reference toward a particular information source.

The flow of information during reference resolution is from level 1 to level 3. Algorithm 1 lists the basic steps in reference resolution.

1. compute the reference relative saliences for each object in the context
2. compute the integrated salience for each object in the context
3. return the object with the highest overall salience as the referent

Algorithm 1. Reference resolution algorithm

In the following sections we describe the data structures, basic visual and linguistic salience algorithms used by the framework, and the algorithms used during reference resolution.

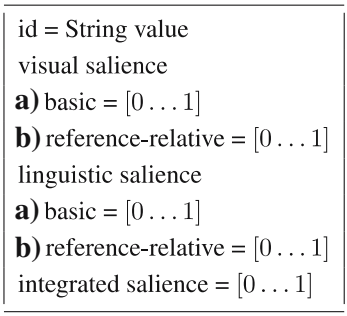

Fig. 2 A coreference class

3.1 Data structures

The basic data structure used by the framework is called an *coreference class*. Each coreference class stores the saliency information for one object in the context model. Figure 2 illustrates the internal structure of a coreference class. The coreference class `id` is a unique string identifier. Each coreference class contains components for storing the basic and reference relative visual and linguistic salience scores and the integrated salience sores for the object the class represents in the context model.

New coreference classes are added to the context model as a result of visual processing. Each time an object is detected in the visual scene the context model is queried for the coreference class representing the object. If there is no coreference class for the object model a new coreference class is created and is assigned the id used by the vision processing. The basic visual salience component is initialised to the value created by vision processing when the object was detected. This is updated after each scene is rendered. All the other salience scores are initialised to 0. These components are updated after each utterance has been processed. Coreference classes are removed from the context model when both their basic visual and linguistic saliences fall below a threshold (0.0001). In the following sections the algorithms that provide and use the information stored in these structures are described.

3.2 Modelling basic visual salience

Most computational models of visual attention focus on bottom-up processing, see Koch and Itti (2001) and Heinke and Humphreys (2004) for recent reviews. In most of these models several feature maps (such as colour, intensity etc.) are computed in parallel across the visual field and these are then combined into a single saliency map. Then a selection process deploys attention to locations in decreasing order of salience. In Kelleher and van Genabith (2004) a simple model of visual salience (based on object size and centrality relative to a focus of visual attention) was presented. In this paper we adopt use this model to capture the information entering the discourse through vision.

The visual salience algorithm uses a *false-colouring* technique. Each object in the simulation is assigned a unique colour or *vision-id*. This colour differs from the normal colours used to render the object in the world; hence the term false colouring. Each frame is rendered twice: first, using the objects' normal colours, textures and shading, and secondly, using the vision-ids. The first rendering is on screen (i.e. the user sees it), the second rendering may be off screen. After each frame is rendered, a bitmap image of the false colour rendering is created. The bitmap is then scanned and a list of the colours in the image is created. Using

Let U_i be a sentence uttered in state s_i, in which reference is made to $\{d_i, \ldots, d_n\} \subseteq D$. Let $C_f(U_i)$ (the forward looking center of U_i) be a partial order defined over $\{d_i, \ldots, d_n\} \subseteq D$. Then the salience weight of objects in s_{i+1} is determined as follows:

$$sf(s_{i+1}, d) = \begin{cases} 1 & \text{if } d = subject(U_i) \\ (sf(s_i, d)/2) + .5 & \text{if } d = object(U_i) \\ (sf(s_i, d)/2) + .25 & \text{if } d = other(U_i) \\ sf(s_i, d)/2 & \text{if } d \notin C_f(U_i) \end{cases}$$

Fig. 3 Linguistic salience weight assignment

this list the system can recognise which objects are visible and which are not. Moreover, the system can identify, at the pixel level, the area covered by each object in the scene. This pixel information is used to compute the basic visual saliency of each object.

Mimicking the spread of visual acuity across the retina, the algorithm weights each pixel in the image based on its distance from the point of visual focus. The weighting is computed using Eq. (1). In this equation, D equals the distance between the pixel being weighted and the point of focus, M equals the maximum distance between the point of focus and any point on the border of the image. The point of focus can be determined using eye tracking technology to compute the user's gaze at each scene rendering. However, if eye tracking is not being used the point of focus defaults to the center of the image or to the center of silhouette of the last object referred to. Algorithm 2 lists the procedure used to compute basic visual saliency and to update the coreference classes. For each scene processed the algorithm returns a list of objects in the scene each with a relative salience between 0 and 1, with 1 representing maximum salience.

$$\text{Weighting} = 1 - \left(\frac{D}{M+1}\right) \tag{1}$$

for each object O_i in the scene **do**
  $AW(O_i) =$ average weighting of the pixels covered by O_i
  $Total_{AW} = Total_{AW} + AW(O_i)$
endfor
for each coreference class CR_i **do**
  if CR_i is the coreference class representing O_i **then**
    $CR_i.basic\_visual = (CR_i.basic\_visual/2) + (AW(O_i)/Total_{AW}$
  else
    $CR_i.basic\_visual = CR_i.basic\_visual/2$
  endif
  $Total_{bvs} = Total_{bvs} + CR_i.basic\_visual\_salience$
endfor
for each coreference class CR_i
  $CR_i.basic\_visual = CR_i.basic\_visual/Total_{bvs}$
endfor

Algorithm 2. The basic visual salience algorithm.

3.3 Modelling basic linguistic salience

The basic linguistic salience of objects in the context are computed using an algorithm that is similar to Krahmer and Theune (2002). The algorithm is based on the ranking of the so-called

forward looking centers (C_f) of an utterance. The set of forward looking centers of an utterance contains the objects referred to in that utterance. This set is partially ordered to reflect the relative prominence of the referring expressions within the utterance. Grammatical roles are a major factor here, so that `subject > object > other`. The central component of the algorithm is a function sf that maps the objects in a domain D to the set $\{0, \ldots, 1\}$, with the intuition that 0 represents non-salience and 1 maximal salience. Figure 3 defines the salience function sf used by the framework. The algorithm assumes that in the initial situation s_0 all the objects in the domain are equally (not) salient: $sf(s_0, d) = 0$ for all $d \in D$.

It is not claimed that the function sf is the best way to assign linguistic salience. However, it does provides a reasonable, transparent and operational model of linguistic salience. Algorithm 3 defines the procedure used to update linguistic salience after an utterance has been processed.

Let $Total_{DS} = 0$
for each coreference class CR_i **do**
    $CR_i.BLS = sf(s_j, CR_i)$
    $Total_{DS} = Total_{DS} + CC_i.BLS$
endfor
for each coreference class CC_i **do**
    $CR_i.BLS = CR_i.BLS / Total_{DS}$
endfor

Algorithm 3. The basic linguistic salience algorithm. BLS = basic linguistic salience.

3.4 Computing reference relative saliences

The first step in resolving a reference is to compute for each object in the context the salience of that object within each modality within the context provided by the reference. These *reference relative saliences* are computed for each object by integrating each object's basic visual and linguistic saliences with a rating of how well the object fulfills the selectional preferences[6] encoded in the reference.

The rating of how well an object fits the description provided by a reference is called an *f-score*. Two f-scores are computed for each object for each reference: a visual and a linguistic f-score. Currently, the system can rate objects relative to their type, colour, size[7] and location.[8] Table 1 lists the ratings ascribed to an object for each type of selectional preference. An object's visual f-score is initialised to 0 and its ratings are integrated using addition. An object's linguistic f-score is initialised to 1 and its ratings are integrated using multiplication.

Once the f-scores have been computed the object's reference relative visual and linguistic saliences are computed by integrating the f-scores with its basic visual and linguistic salience. Again, addition is used for integration in the visual context and multiplication is used for integration in the linguistic context. Consequently, an object's reference relative visual salience will be > 0 if it fulfils any of the selectional preferences in the description, and its linguistic reference relative salience will be $= 0$ if it does not fulfill all of the selectional preferences in the description. Algorithm 4 lists the algorithm for computing the reference relative saliences.

[6] The semantics of the descriptive terms used in the reference.

[7] An objects size rating is based on the number of pixels it covers relative to the other objects in the scene.

[8] An objects location rating is computed using the AVS model described in Regier and Carlson (2001).

Table 1 Selectional preferences scores

	Fulfills		Not fulfill
Type	1		0
Colour	1		0
Size		[1 . . . 0]	
Location		[1 . . . 0]	

for each coreference class CR_i **do**
  $f\_score_{linguistic} = 1$
  $f\_score_{vision} = 0$
  for each selectional preference sp_j in the description **do**
    $f\_score_{linguistic} = f\_score_{linguistic} * rating(CR_i, sp_j)$
    $f\_score_{vision} = f\_score_{vision} + rating(CR_i, sp_j)$
  endfor
  $CR_i.RLS = CR_i.BLS * f\_score_{linguistic}$
  $Total_{rls} = Total_{rls} + CR_i.RLS$
  $CR_i.RVS = CR_i.BLSl * f\_score_{vision}$
  $Total_{rvs} = Total_{rvs} + CR_i.RVS$
endfor
for each coreference class CR_i **do**
  $CR_i.RLS = CC_i.RLS/Total_{rls}$
  STATE $CC_i.RVS = CC_i.RVS/Total_{rvs}$
endfor

Algorithm 4. Computing the reference relative saliences. RLS = reference relative linguistic salience, RVS = reference relative visual salience, BLS = basic linguistic salience, BVS = basic visual salience.

3.5 Creating the integrated context and selecting the referent

The final step before the selection of the referent is the integration of each object's reference relative saliencies. This is done using a weighted combination. The weightings are dependent on the form of referring expression (e.g. definite descriptions versus pronominal references) being resolved and reflect the preferential interpretation associated with each type of reference. For example, in general, a pronoun is used to refer to a referent that is prominent within the linguistic context. By contrast, a definite description can be used to refer to an object from the visual scene and to previously mentioned objects. Currently, the system uses predefined weights for this integration. When resolving a definite description visual and linguistic salience are integrated evenly. When resolving a pronominal reference the integration weightings used biases towards linguistic salience. Algorithm 5 defines the procedure used to construct the integrated context and select the reference. It also defines the mechanism used to check for ambiguous references. This ambiguity check uses a predefined confidence interval and simply checks that within the context provided by the referring expression the integrated salience of the object selected as the referent is sufficiently larger than the other

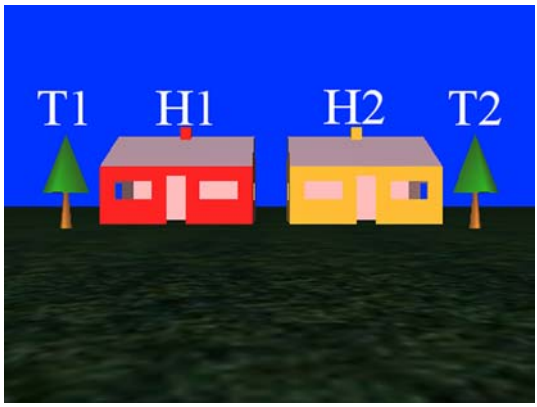

Fig. 4 Example visual context. H1 = red house, H2 = green house

objects in the context to ensure that the reference is not ambiguous. In situations where the ambiguity check fails the algorithm returns 0.

for each coreference class CR_i in the context model **do**
  Let index = 0, max = 0, interval = 0.3
  if reference = definite description **then**
    $CR_i.integrated = (CR_i.RVS * 0.5) + (CR_i.RLS * 0.5)$
  elseif reference = pronominal reference **then**
    $CR_i.integrated = (CR_i.RVS * 0.1) + (CR_i.RLS * 0.9)$
  endif
  if $CR_i.integrated > max$ **then**
    index = i
    $max = CR_i.integrated$
  endif
endfor
for each coreference class CR_j in the context model **do**
  if $j <> index$ **then**
    if $CR_j.integrated > CR_{index}.integrated - interval$ **then**
      return 0 //Reference Deemed Ambiguous
    endif
  endif
endfor
return CR_{index}

Algorithm 5. Constructing the integrated context and selecting the references. RVS = reference relative visual salience, RLS = reference relative linguistic salience.

4 Worked example

The functioning of the framework can be illustrated using a worked example. The example uses Fig. 4 as a visual context, and the utterances (1) and (2) as the example discourse.

1. make *the red house* green
2. make *the tree to the left of it* bigger

Fig. 5 The final state of the visual context

Table 2 lists the saliences scores computed by the framework during the different stages of this interaction. Rows 1 and 2 of the table present the initial basic visual and linguistic salience scores of the objects in Fig. 4. Rows 3 to 7 presents the f-scores and reference and integrated saliences computed for the objects when the system processed *the red house*. The asterix in line 7, H1's column, indicates the highest integrated salience at the end of the resolution process. As a result of obtaining the maximum salience H1 is selected as the referent. Rows 8 and 9 list the basic salience scores for the objects after the basic linguistic salience has been updated and the point of visual focus has been located at the center of H1's silhouette. The movement of the visual focus away from the center of the image is reflected in the increases in the basic visual salience of H1 (0.3271 $\rightarrow$ 0.3581) and T1 (0.1728 $\rightarrow$ 0.2938). Rows 10–14 list the f-scores and reference and integrated saliences computed for the objects when the system processed *it*. The biasing towards linguistic salience is apparent in the dominance of H1's integrated salience. Rows 15–19 list the f-scores and reference and integrated saliences computed for the objects when the systems processed *the tree to the left of it*. The difference between the visual and linguistic f-scores of T1 and T2 is due to the locational description: T1 was judged by the system to fulfill the locational description with a rating of 0.9396, while T2 was judged not to fulfill the description and was ascribed a rating of 0.0000 for this selectional preference. As a result, T1 achieved the highest salience (0.5669) and was selected as the referent. Figure 4 illustrates the visual context at the end of the interaction.

5 Conclusions and future work

This paper presented an attention based reference resolution framework for visually situated discourse. The framework uses a weighted integration of visual and linguistic attention to order the candidate referents within the context. The candidate with the highest integrated attention score is taken to be the referent. One advantage of this approach is that the resolution process occurs within the full multimodal context. As a result situations where the intended target of the reference is erroneously excluded, due to an individual assumption within the resolution process, are avoided. Moreover, the system can recognise situations where attentional cues from different modalities make a reference potentially ambiguous. From a cognitive perspective the framework meshes well with psycholinguistic results that point to the role of attention within human reference resolution processes.

Table 2 Salience scores computed during the example interaction

		H1	H2	T1	T2
Initial context					
1	Basic visual salience	0.3271	0.3272	0.1728	0.1728
2	Basic linguistic salience	0.0000	0.0000	0.0000	0.0000
The red house					
3	Visual *f*-score	2.0000	1.0000	0.0000	0.0000
4	Linguistic *f*-score	1.0000	0.0000	0.0000	0.0000
5	Reference visual salience	0.5818	0.3318	0.0432	0.0432
6	Reference linguistic salience	0.0000	0.0000	0.0000	0.0000
7	Integrated salience	0.5818*	0.3318	0.0432	0.0432
8	Basic visual salience	0.3581	0.2273	0.2938	0.1208
9	Basic linguistic salience	1.0000	0.0000	0.0000	0.0000
It					
10	Visual *f*-score	0.0000	0.0000	0.0000	0.0000
11	Linguistic *f*-score	1.0000	1.0000	1.0000	1.0000
12	Reference visual salience	0.3581	0.2273	0.2938	0.1208
13	Reference linguistic salience	1.0000	0.0000	0.0000	0.0000
14	Integrated salience	0.9358*	0.0227	0.0293	0.0120
The tree to the left of it					
15	Visual *f*-score	0.0000	0.0000	1.9396	1
16	Linguistic *f*-score	0.0000	0.0000	0.9396	0
17	Reference visual salience	0.0909	0.0577	0.5669	0.2845
18	Reference linguistic salience	0.0000	0.0000	0.0000	0.0000
19	Integrated salience	0.0909	0.0577	0.5669*	0.2845

Finally, it should be noted that the framework as it currently stands is intended to represent an abstract and preliminary attempt. Several issues need to be addressed if it is to be used as a component within a dialog systems for less constrained contexts. In particular, the use of predefined weights for salience integration is overly simplistic. This issue could be addressed by using a machine learning algorithm, such as reinforcement learning, to automatically compute these weights. The visual and linguistic salience algorithms should also be improved. For dialog systems interfacing with virtual environments, the visual salience algorithm should be extended to at least handle attentional cues such as color, motion and location of gaze. If the framework was to be used within a real-world system, such as a robot dialog system, a computer vision saliency algorithm, such as Itti and Koch (2000), could be adopted. The linguistic saliency algorithm should also be revised and extended. For example, a information based approach to linguistic attention, such as Hajicová (1993), may be more suitable than the simplified Centering (Grosz et al. 1995) based framework proposed here. Moreover, the relationship between the framework's model of local level attention and a more global model of discourse structure, such as Grosz and Sidner's (1986) focus stack model or Asher and Lascarides' (2003) SDRT framework, should be clarified. Fortunately, the modular nature of the framework makes such modifications possible without major changes to the overall approach.

Acknowledgements The author would like to thank the anonymous reviewers for their valuable comments and feedback.

References

Allen J, Schubert L (1991) The TRAINS Project. Technical report, Department of Computer Science, University of Rochester

Alshawi H (1987) Memory and context for language interpretation. Cambridge University Press, Cambridge, UK

Ariel M (1990) Accessing noun-phrase antecedents. Routeledge, London

Asher N, Lascarides A (2003) Logics of conversation. Cambridge University Press

Chum M, Wolfe J (2001) Visual attention. In: Goldstein EB (ed) Blackwell Handbook of perception, Handbooks of experimental psychology, Chapt. 9. Blackwell, pp 272–310

Duwe I, Strohner H (1997) Towards a cognitive model of linguistic reference. Report: 97/1–Situierte Künstliche Kommunikatoren 97/1, Univeristät Bielefeld

Enns J, Rensink RA (1990) Influence of scene-based properties on visual search. Science 247:721–723

Gorniak P, Roy D (2004) Grounded semantic composition for visual scenes. J Artif Intell Res 21:429–470

Grosz B (1977) The representation and use of focus in dialogue understanding. Ph.D. thesis, Standford University

Grosz B, Joshi A, Weinstein W (1995) Centering: a framework for modelling local coherence of discourse. Comput linguist 21(2):203–255

Grosz B, Sidner C (1986) Attention, intentions, and the structure of discourse. Comput Linguis 12(3):175–204

Gundel J, Hedberg N, Zacharski R (1993) Cognitive status and the form of referring expression in discourse. Language 69:274–307

Hajicová E (1993) Issues of sentence structure and discourse patterns, Theoretical and Computational Linguistics, vol 2. Charles University Press

Heinke D, Humphreys G (2004) Computational models of visual selective attention: a review. In: Houghton G (ed) Connectionist models in psychology. Psychology Press

Hobbs J (1985) On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information

Hopfinger J, Buonocore M, Mangun G (2000) The neural mechanisms of top-down attentional control. Nat Neurosci 3(3):284–291

Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res 40:1489–1506

Kelleher J, Costello F, van Genabith J (2005) Dynamically structuring, updating and interrelating representations of visual and lingusitic discourse context. Artif Intell 167(1–2):62–102

Kelleher J, van Genabith J (2004) Visual salience and reference resolution in simulated 3D environments. AI Rev 21(3–4):253–267

Kievit L, Piwek P, Beun R, Bunt H (2001) Multimodal cooperative resolution of referential expressions in the denk system. In: Bunt H, Beun R (eds) Cooperative multimodal communication: Lecture Notes in Artificial Intelligence 2155. Springer-Verlag, Berlin Heidelberg, pp 197–214

Koch C, Itti L (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203

Krahmer E, Theune M (2002) Efficient context-sensitive generation of referring expressions. In: van Deemter K, Kibble R (eds) Information sharing: reference and presupposition in language generation and interpretation. CLSI Publications, Standford

Kruijff-Korbayová I, Hajicová E (1997) Topics and centers: a comparison of the salience-based approach and the centering theory. Prague Bull Math Linguist 67:25–50. Charles University, Prague, Czech Republic

Landragin F, Romary L (2003) Referring to objects through sub-contexts in multimodal human–computer interaction. In: DiaBruck 7th workshop on the semantics and pragmatics of dialogue, Sept 4th–6th 2003. University of Saarland, Germany

Lappin S, Leass H (1994) An algorithm for pronominal anaphora resolution. Computat Linguist 20(4): 535–561

Mann W, Thompson S (1987) Rhetorical structure theory: description and construction of text structures. In: Kempen G (ed) Natural language generation: new results in artificial intelligence, psychology and linguistics. Nijhoff., Dordrecht, pp 83–96

McKevitt P (ed) (1995/1996) Integration of natural language and vision processing, vols I–IV. The Netherlands: Kluwer Academic Publishers, Dordrecht

Poesio M (1993) A situation-theoretic formalization of definite description interpretation in plan elaboration dialogues. In: Aczel P, Israel D, Katagiri Y, Peters S (eds) Situation theory and its applications, vol 3. CSLI, pp 339–374

Regier T, Carlson L (2001) Grounding spatial language in perception: an empirical and computational investigation. J Exp Psychol Gen 130(2):273–298

Salmon-Alt S, Romary L (2001) Reference resolution within the framework of cogitive grammar. In: Proceedings of the Seventh International colloquium on cognitive science (ICCS-01). Donostia, Spain, pp 284–299

Spivey-Knowlton M, Tanenhaus M, Eberhard K, Sedivy J (1998) Integration of visuospatial and linguistic information: language comprehension in real time and real space. In: Olivier P, Gapp K (eds) Representation and processing of spatial expressions. Lawrence Erlbaum Associates, pp 201–214

Tanenhaus M, Spivey-Knowlton M, Eberhard K, Spivey J (1995) Integration of visual and linguistic information in spoken language comprehension. Science 268:1632–1634

Yarbus A (1967) Eye movements and vision. Plenium Press, New York