Disambiguation between Visual Display and Represented Domain in Multimodal Interfaces

Daqing He* **Graeme Ritchie*** **John Lee†**

Abstract

In natural language queries/commands to an intelligent multimodal system, a particular kind of ambiguity – *source ambiguity* – can arise between items in a visual display and entities in the domain being represented. We associate with each symbolic entity an attribute called its *source*, which indicates which model (screen or domain) the entity belongs to. Also, each referring expression is assigned two representations: the *described entity sets* and the *intended referents*. The process of resolving source ambiguity then happens as a part of the referent finding process. This is work in progress, with many interesting aspects needing further development.

1 Source ambiguities

In research into natural language (NL) input to intelligent multimodal interfaces, there is little attention paid to allowing the display attributes (visual characteristics) of entities on the screen to be used in natural language referring expressions. (There is some work on NL generation which considers this issue; see Section 6 below.) But problems that are interesting in both their linguistic and their multimodal aspects are encountered as soon as the display attributes of entities are allowed to be used in the NL queries or commands, especially in those systems where there is a representation mapping between visual entities and domain objects. Ambiguities can arise involving the exact meaning of a word; for example, **blue** in the phrase "`a blue car`" could be a feature in the represented world, or a screen attribute.[1] This can lead to ambiguity of reference; for example, the phrase "`a blue car`" could refer to a blue car in the represented world, or a screen entity that is blue and represents a car. The meaning of the whole sentence can be affected; for example, the sentence "`move the blue car here`" could request an event on the screen or an action in the represented world. In some situations, there may be conflicts to be resolved. For instance, a car which is blue in the domain may be represented by a red icon on the screen, which raises a question whether and when it should be called "`a blue car`" or "`a red car`"; or both the domain and the screen may have a **move** operation resulting in ambiguity.

These problems, though appearing in several forms, share a common cause: ambiguity about the relevant model. He et al. [He et al., 1997] argued that there are two models – the *world model* (domain) and the *display model* (screen) – in this kind of system, which we call *IMIG*s – Intelligent Multimodal Interfaces employing Graphics. All the entities in the system are either *domain entities* or *screen entities*. Accordingly, the attributes of the entities and the actions applicable to the entities are divided into *screen/domain attributes* and *screen/domain actions*.

Although it is important to distinguish

*Department of Artificial Intelligence, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland. `daqingd,graeme@dai.ed.ac.uk`

†Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland. `john@cogsci.ed.ac.uk`

[1] In our examples, we shall assume an interactive system displaying details of cars for sale, where icons on the screen represent individual cars, and various characteristics of the icons convey attributes of the corresponding cars.

between screen entities and domain entities (for example, the former may have a "`colour`" attribute even if the domain makes no use of colour), various processes (e.g. reasoning) are simplified if we represent both using some uniform scheme, but allow for an explicit relation (of denotation) between a screen symbol and the corresponding domain entity, which we shall refer to here as the *representation mapping*. The screen symbol and the domain entity connected by this mapping are called the *corresponding entity* of each other.

If we adopt the term *source* for the particular model that an attribute, entity or action belongs to, then the various ambiguities outlined above can be referred to as *source ambiguities*.

Some people may argue that this problem is not strictly an "ambiguity". There is no single general accepted definition of ambiguity, but it seems reasonable to describe a sentence or phrase as ambiguous if there is more than one clearly defined meaning for it. Source ambiguities occur at the semantic processing level (cf. [Gazdar and Mellish, 1989, Section 10.1]), and as shown by the examples given above, they have effects on various aspects of that level. Since individual words with source ambiguities have effects on referent identification, and it is not possible to eliminate all such ambiguities at a lexical level, there is a need for a referent resolution process which can allow for source ambiguities at the word, phrase and clause levels. For simplicity, we will restrict our discussion to a subset of noun phrases: singular expressions referring to specific objects which exist in the models, in which the premodifiers are epithet adjectives and/or a noun classifier [Halliday, 1994], and the only type of post-modifier is a prepositional phrase. In addition, we assume that there are no examples of "one"-anaphora in our referring expressions.

We call phrases whose words are all from the screen or all from the domain *screen phrases* and *domain phrases* respectively. It is possible that some words of a phrase are from the domain, while others are from the screen; we call that kind of phrase a *mixed-source phrase*. For instance, "$\texttt{this}\uparrow\ \texttt{small}_s\ \texttt{icon}_s$", "$\texttt{the expensive}_d\ \texttt{car}_d$" and "$\texttt{the blue}_s\ \texttt{car}_d$" are a screen phrase, a domain phrase and a mixed-source phrase respectively.[2]

2 Referring expressions with sources

Described entity set

According to typical treatments of reference, most referring expressions have a meaning which is a function of the meaning of the words in them, and which can be called its *descriptive content* [Kronfeld, 1990] (this is essentially the traditional philosophical notion of *sense*). This content, like pragmatic restrictions (such as, constraints from world knowledge and from discourse, etc. [Appelt, 1989]), provides the constraints that the referent must satisfy.

However, this scheme is not wholly appropriate for our problem. In an IMIG system, the descriptive content of a referring expression may appear to describe an entity in one model, but when considered in a larger context, such as with other components of a sentence, it may actually refer to the corresponding entity in the other model. For example, we can say that the phrase "$\texttt{the expensive}_d\ \texttt{car}_d$" describes a domain entity which is a car and expensive, and it also refers to such an entity in the sentence "$\texttt{buy}_d\ \texttt{the expensive}_d\ \texttt{car}_d$". However, it does not thus refer in the sentence "$\texttt{move}_s\ \texttt{the expensive}_d\ \texttt{car}_d\ \texttt{to the top}_s\ \texttt{of the screen}_s$". In order to make sense of the operation **move**, which is a screen action, the only entity that the phrase can possibly refer to is a screen entity, the corresponding entity of the expensive car in the domain. This is in spite of the fact that the descriptive material in the phrase (e.g. "$\texttt{expensive}_d$") applies to the domain entity.

Another problem suggests that it is not enough in an IMIG system to let each phrase simply have a descriptive content (sense) and

[2] The subscript attached to a word shows the source of the word: s and d mean screen and domain respectively. "$\uparrow$" indicates a pointing action.

a referent. Consider the example "move$_s$ the expensive$_d$ car$_d$ to the top$_s$ of the screen$_s$ and place an order$_d$ for it". According to linguistic intuition, the pronoun "it" seems to have an antecedent in the sentence, which is "the expensive$_d$ car$_d$". Unless we posit such a relation, it is quite difficult to interpret the pronoun "it". However if we examine the referents, the problem appears. For the pronoun "it", its referent is an expensive car in the domain, while that of the phrase "the expensive$_d$ car$_d$" is the icon corresponding to that expensive domain car.

All of these suggest to us that as well as the eventual referent for each phrase, there must be some other form of representation to cope with this kind of situation. It is accepted that the same descriptive content of a phrase can have a different referent in different situations, which can be shown in the following example (see also figure 1), which suggests that the descriptive content is not the appropriate representation to make this distinction.

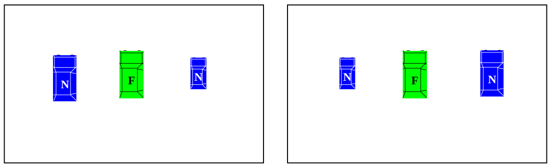
User: Is the dark$_s$ car$_d$ at the right$_s$ of the grey$_s$ car$_d$ small$_d$?

System: Yes, it is.

Action: *The user swaps the two dark cars by using the mouse. The screen shot then changes from (A) to (B) in figure 1.*

User: Is the dark$_s$ car$_d$ at the right$_s$ of the grey$_s$ car$_d$ small$_d$?

System: No, it is a large$_d$ car$_d$.



(A) (B)

Figure 1: *Two screen shots for the car browsing*

Informally speaking, we need a representation which is more precisely specified than the descriptive content, but not necessarily as exactly defined as the ultimate referent object. The representation stands for a set of entities described by the descriptive content, which could be the set or a subset of the potential referents in the current situation. This allows both the described item and its corresponding entity to be considered in the process of finding the ultimate referent object.

To do this, we associate with each referring expression a *described entity set*,[3] which can be thought of as the set of item(s) described by the descriptive content (sense) of the phrase, in the current situation. Each such expression will also have an *intended referent*, the item which is ultimately denoted within the context, for example as argument to a verb or preposition. The intended referent is essentially what would be called the *referent* in a traditional (non-IMIG) situation; the complication here is the lack of a direct descriptive relation between the substance of the phrase and that entity.

Some definitions

The following semi-formal definitions elaborate the relationship between described entity sets and intended referents. Although all these definitions would have to be generalised to cover plural noun phrases in future, the concepts are clearer using our simplifying assumption of considering only singular noun phrases.

Definition 1 *A* described entity set *is a set with one of the following forms:*

1. *a singleton set containing either a domain entity or a screen entity;*
2. *a two-element set containing a domain entity and a screen entity related by the representation mapping.*

[3] In an earlier paper [He et al., 1997], we called these entities the *described referents*, but we found this term was contentious, since there could be philosophical disputes about their status as referents. Since they are merely a convenient formal construct in our theory, we have modified the name to avoid irrelevant arguments.

Definition 2 *The semantic form P of a noun phrase* describes *a described entity set R iff both:*

1. *every predicate in P is true of an element in R, with source taken into account (i.e. domain predicates can only be true of domain entities, screen predicates of screen entities);*
2. *for every element in R, there is a predicate in P which is true of it.*

The details of Definition 2 are dependent on the simplifying assumption that the modifiers in an NP are semantically a conjunction of independent properties. As an example, "`the small`$_d$ `expensive`$_d$ `car`$_d$" would "describe" the described entity set **{car1}** if **car1** was an entity which was classed as a **car** and had the properties **small** and **expensive** asserted of it in the domain model. Notice that this relationship between phrases and entities is relatively context-free, in that it does not take into account the surrounding sentence structure (if any).

Definition 3 *A* described entity set assignment *is a mapping which allocates to each noun phrase N a described entity set R such that the semantic form of N describes R.*

Informally, the described entity set of an NP contains either an entity that is completely described by the content of a single-source NP (e.g. **car1** for the phrase "`the small`$_d$ `expensive`$_d$ `car`$_d$"), or the two corresponding entities that might each be partly described by a mixed-source phrase (e.g. **{car1, icon1}** for "`the blue`$_s$ `car`$_d$"). Definition 3 is intended to capture the notion of a "reading" or "interpretation", so that we can discuss the phenomena we are interested in without the complication of other forms of ambiguity (not source); these other factors may give rise to alternative described entity set assignments.

To capture the effect that a phrase can refer to the entity that corresponds to (one of) the entities that it actually describes, we need two further definitions.

Definition 4 *Given a described entity set R, the* potential referents *of R is the set:*
$R \cup \{y \mid$ *there is an x in R which is related to y by the representation mapping*$\}$.

Definition 5 *Given a described entity set assignment F, an* intended referent assignment based on F *is a mapping G which allocates to each noun phrase N an element of the potential referents of F(N).*

In Definition 5, G(N) is a formalisation of the notion of "intended referent" of a phrase N.

For phrases where there is a single source for words in the expression, there is no doubt about the source of the entities in the described entity set; however, the intended referents may be either the same as the entities in the described entity set (as in "`Buy`$_d$ `the expensive`$_d$ `car`$_d$") or may correspond, via the representational mapping, to the entities in the described entity set (as in "`Move`$_s$ `the expensive`$_d$ `car`$_d$ `to the top`$_s$ `of the screen`$_s$").

Coreference in an IMIG system

In a dialogue system, the relationship between anaphora and their antecedents is sometimes known as *cospecification* [Sidner, 1987], or *coreference* [Lyons, 1977], and is usually taken to mean that the two phrases refer to the same object [Allen, 1995]. But in IMIG systems this informal definition is not general enough. Consider the dialogue:

User: What is the most expensive$_d$ car$_d$?

System: The ACME-4000$_d$.

User: Delete$_s$ that icon$_s$.

Here, the phrases "`the ACME-4000`$_d$" and "`that icon`$_s$" could be argued to be related coreferentially, in the sense that the latter phrase takes on its full meaning anaphorically using the former phrase an antecedent. However, the two phrases have different intended referents – the first denotes a domain entity, the second a screen entity.

In a situation where there are various sorts of referents for each referring expression, we replace the definition of *coreference* with the following:

Definition 6 *Noun Phrases N1 and N2 are* quasi-coreferential *iff either they have the same intended referents or their intended referents correspond under the representation mapping.*

As in the definition of coreference in previous research, quasi-coreferential is defined without any details of computation. In order to identify whether or not two phrases are quasi-coreferential, we have to develop our own computational rules for doing so. Following the discourse approach to resolving anaphoric phrases [Sidner, 1987], our system uses a *context model*, a store of recently mentioned entities organised by the focus theory of attention [Grosz and Sidner, 1986], to narrow down the potential antecedent candidates. However, suitable heuristic rules about syntactic, semantic and general attributes of antecedents also have to be constructed.

From definition 6 and definition 4, we can draw a result that *two phrases are quasi-coreferential iff their potential referents are the same*, and then using the definition 1, the following rules can be constructed:

> Two phrases are quasi-coreferential iff: *either* 1) the described entity set of one phrase is a subset of that of the other; *or* 2) the entity in one described entity set is the corresponding entity of an entity in the other described entity set.

In the case of the anaphoric pronoun "`it`", because it lacks detailed descriptive content, we shall adopt the convention that its described entity set is the same as that of its antecedent. For instance, in the sentence "$\texttt{move}_s$ `the` $\texttt{expensive}_d$ $\texttt{car}_d$ `to the` $\texttt{top}_s$ `of the` $\texttt{screen}_s$ `and place an` $\texttt{order}_d$ `for it.`", in the reading where "`it`" and "`the` $\texttt{expensive}_d$ $\texttt{car}_d$" are related as anaphor and antecedent, they are defined to have the same described entity set. However their intended referents do not have to be the same, and they are not in this example.

With the discussion above and some heuristics mentioned in the next section, we can start to develop the computational process for finding an antecedent.

3 Linguistic factors

Since an IMIG system has to compute (intended) referents in order to respond to questions, carry out commands, etc., we need to develop some referent-finding procedures. In particular, we need to depend on linguistic regularities as the basis for heuristics, much as in referent-identification within non-IMIG NLP systems.

Restrictions inside phrases

The representational relation between screen entities and domain entities is not symmetrical. For example, we can say that "`a square represents a car`", however we cannot say that "`a car represents a square`". This asymmetric aspect restricts the use of head nouns in the interaction. Typically, the head noun of a referring expression is a common noun, which refers to a class of entities. A screen entity, such as a **square**, can be referred to using a head noun naming its screen category, i.e. "$\texttt{square}_s$", or by a word naming the category of the domain entity that it represents, such as "$\texttt{car}_d$". However a domain entity, such as a **car**, cannot be called by the head noun from the corresponding screen category; that is, it is very odd to refer to a **car** as "`a` $\texttt{square}_s$". From the above discussion, we can summarise that if the head noun of a phrase is a screen word, the phrase can only refer to a screen entity, while if a domain head noun is encountered, it can refer to either a domain entity or a screen entity.

If the determiner of the phrase is a demonstrative word (e.g. "`this`") accompanied by a pointing action, it acts as a link between the language modality and some operation in the graphic modality [Neal et al., 1988]. Since the graphic operation acts on some screen entity, which it identifies through its basic semantic form, we say that the phrase "describes" that screen entity. Hence the described entity set of the phrase should include the screen entity indicated by the pointing.

According to Halliday [Halliday, 1994], modifiers in a noun phrases have the function of providing extra information about an entity in the class specified by the head noun to discriminate it from other potential referents. Hence the attributes provided by the modifiers should be appropriate to the class of entities specified by the head. There appear, from our observations, to be further regularities. For instance, if the head noun of a phrase denotes a screen category, the modifiers can only be screen words otherwise they are not appropriate to modify the screen entities. For example, phrases "a red$_s$ icon$_s$", "a small$_s$ circle$_s$" are meaningful expressions, while phrases "*the expensive$_d$ icon$_s$" and "*every large$_d$ square$_s$" are unacceptable. In contrast, if the head noun describes a domain category, there is not any strong restriction on modifiers. So phrases "an expensive$_d$ car$_d$ in$_s$ the corner$_s$", "the red$_s$ car$_d$", "the small$_d$ saloon$_d$" are all acceptable.

Sometimes, a word from what Halliday [Halliday, 1994] calls the *Classifier* group (especially if it is a noun) may be so closely bonded to a head noun of a fairly general screen class that the whole phrase is still meaningful even though the classifier is a domain word, as in "the Nissan$_d$ icon$_s$". This could be because this Classifier is not functioning as a modifier, rather this is an example of a compound head noun.

In summary, the constraints on a referring expression are:

1) if its determiner is a demonstrative word with a pointing action, the screen entity pointed to is in the described entity set;

2) if its head noun unambiguously names a screen category, its non-Classifier modifiers describe screen attributes;

3) if its head noun unambiguously names a screen category, and it does not have a Classifier modifier, its described entity set does not contain any domain entities and its intended referent is a screen entity.

For example, the words "blue" and "small" in phrases "the blue icon" and "a small square" are all from the screen if the head nouns "icon" and "square" are both from the screen.

Mutual influence among phrases

In the past decades, a great many psycholinguistic observations of human generation of referring expressions have been performed. In Levelt's review of those achievements [Levelt, 1989], one point is relevant to our current issue: *"the speaker apparently contrasted the new referent object with the previous one"*. From this, we know that comparisons between features are used in referring. Our commonsense knowledge tells us that the features in a comparison are different but comparable, so it is much easier to understand a comparison if the attributes are from the same source rather than different sources. Thus we can propose the following heuristic rules:

Same predicate rule If a predicate of a phrase is known to come from a source, the predicates of the same type in other phrases in the same sentence are probably from the same source too. Two predicates are in the same type if their semantic categories share a common direct supercategory in the sort hierarchy.

Same structure rule If there are two phrases with the same structure in a sentence, and two predicates are at the same positions in these phrases, the sources of these predicates are probably the same. The criteria for defining two phrases as having the same structure are 1) they have same type of head noun; and 2) either they both have post-modifiers, or both do not.

Although the above two rules sound similar, they actually aim for different situations. For instance, the two predicates "`red`" and "`blue`" in the sentence "`is the upholstery in the red car`$_d$ `blue?`" are probably from the same source because their relation satisfies the *same predicate* rule. An example where the relation satisfies the *same structure* rule can be found in the sentence "`move`$_s$ `the small car`$_d$ `to the right`$_s$ `of the expensive`$_d$ `car`$_d$`.`" Here, the predicates of "`small`" and "`expensive`" occur in phrases having the same structure as each other, which suggests that they probably share the common source. In some situations, both rules can be satisfied.

4 Representation of meanings with sources

Our work on representing the source is to build a two-layer meaning representation language (MRL). An original query input by the user does not provide explicit source information, so at the initial stage of the process the first layer of the MRL is used to represent the meaning in a form which may lack source information. An essential characteristic of the second layer, which occurs just before the execution of the query against a database, is that it can represent the meaning of the input sentence with details of "sources" indicated explicitly. This is crucial since the evaluation of the query/command must be sensitive to the sources of predicates and constants involved. For example, a particular predicate-name, say **is-blue**, may have a different interpretation depending on whether the screen or domain model is used. There is therefore a gap between the meaning representation used in the first layer and that used in the second layer, and our main work is to bridge that gap. In providing this bridge, all the necessary source information is filled in so that the source ambiguities are solved accordingly.

5 Resolution process

As well as the MRL which can represent the meanings before and after the resolving of source ambiguities, we have devised several rules inside and among phrases which can give guidance in finding sources, and definitions which relate described entity sets and intended referents.

Our approach is to treat the source as a necessary attribute of entities otherwise they cannot be distinctly identified. Under our treatment, the source ambiguity resolving process is a part of the referent resolving process. The referent resolving process in an IMIG system is very different from that of previous systems. In the following, we will talk about the whole referent resolving process which is divided into three parts.

Source guessing

In section 1, we said that without a clear source for each word, the meaning of a phrase is vague, and thus it is very difficult to find its described entity set and intended referent. So the first step of the resolution process is to find the sources of words in the phrases – the *source guessing process*.

The basis of the source guessing process is the heuristic rules we discussed in section 3. The main idea is to use the known sources of some words to guess the sources of others in the same sentence. We assume here and elsewhere that the NL lexicon can indicate the source for a word, so that if the source of a word is unambiguous, the NL processes will have access to that information; in the case of source-ambiguous words, such as "`blue`" in some of our examples above, the lexicon will not be able to help. In the initial semantic processing, the system will assign an appropriate source to semantic symbols wherever there is clear information. Those sources are the known sources in the initial stage of the source guessing process. Later whenever the source of a predicate has been found and verified, it can be used for further guessing. In the final state of the guessing some predicates of the words in some phrases are assigned a source value, and a set of ordered possible meanings of those phrases is generated.

As the name of this process implies, all the results of this process are hypotheses that need to be verified, otherwise it is not certain which one is actually the correct meaning of a phrase. The verification is done in the next step, the *described entity set finding* process.

Described entity set finding

The described entity set finding process has two roles. First it is a referent retrieving process that will find the entities described by the input phrases. The phrases it has to deal with are definite descriptions, deictic phrases and pronouns. Two of the usages of definite noun phrases mentioned by Hawkins [Hawkins, 1978] are considered here: the *anaphoric use* and the *visible situation use*. In the case of anaphoric use of definite descriptions, the context model is accessed, and during the retrieval, the various referential constructs of section 2 above are used to guide the search. The visual use of a definite description – that is to say, referring to an entity which is visible to the user, e.g. "`the leftmost`$_s$ `red`$_s$ `car`$_d$" – is one of the most distinctive characteristics of an intelligent multimodal system. In contrast to the anaphoric use of a definite description, in which a mentioned entity is retrieved from the context, the visual use of a definite description will directly bring into the dialogue a screen entity which has not been mentioned previously. The display model is the place to find that entity.

For deictic phrases, the display model is the only resource to be searched. There has been a large amount of discussions about resolving deictic phrases (e.g. [Neal et al., 1988]). The only thing worth mentioning here is that in an IMIG system, sometimes the world model has to be checked to obtain relevant information about the corresponding entity if the phrase is a mixed-source one (e.g. "`this`$\uparrow$ `blue`$_s$ `car`$_d$").

The process for finding described entity sets of pronouns is similar to that for anaphoric use of definite descriptions, where the context model is searched for the described entity sets of their antecedents.

The second function of the described entity set finding process, as we said in the last section, is verification of the results generated by the source guessing process. The correctness is decided by whether or not a possible meaning can find an entity in the source to be the element in the described entity set. Our motivation for this criterion is the assumption that in the system a meaningful referring expression will always lead to an entity in the models.[4] So if an interpretation of the phrase cannot find an entity in the model, it is wrong, and thus should be deleted from the possible meanings.

Certain types of noun phrase (indefinite noun phrases and interrogative phrases) would not normally be subject to the referent resolving process. However, since the meanings of those phrases are the results of source guessing process, they have to be tested too, in order to verify their source assignments. Although their information is not sufficient to find a particular entity to be the element of their described entity sets, they do describe what kind of entities could be the elements of their described entity sets. So for these phrases the criterion mentioned above – that phrases should have possible referents – is effected by checking the display model, the world model or both models (according to whether the phrases are screen phrases, domain phrases, or mixed-source phrases), to see whether there is at least one entity in the models satisfying their restrictions. The result is either success, which can be interpreted as this result of guessing is acceptable, or fail, which means that this result is invalid.

Intended referent finding

After the described entity set finding process, the verification of the current source guessing is complete, and the source ambiguities in those phrases are resolved. However the intended referents of the phrases are still unknown. The intended referent is the key to resolving the source ambiguities at the sentence level or in a more complex phrase. For example,

[4] We acknowledge that this assumption is probably too strong since it means that the system cannot interpret a query like "`is there a red`$_d$ `car`$_d$`?`" if there is no red car in the databases.

in a simple sentence "delete the expensive$_d$ car$_d$", although the described entity set of the phrase "the expensive car" is found to be, say, **car1**, without knowing the source of the operation **delete** and the intended referent of the phrase, the meaning of the sentence is still unclear. In the case of more complex noun phrases, after the described entity set finding process, it may be that only a part of the phrase has found its described entity set, while the intended referent of this part and the described entity set of the whole phrase are unknown. For example, after the described entity set of "the red$_s$ car$_d$" in the phrase "the blue car near the red$_s$ car$_d$" has been found, the source of the word "near" and the intended referent of the phrase "the red$_s$ car$_d$" are needed for finding the described entity set of the whole phrase. The fact that we cannot solve the referents of all phrases together means that the resolving process has to run repeatedly. In each run, based on the results of the last run, some source ambiguities are resolved and the referents of some phrases are found. Then a new run can start on the new results until all source ambiguities are resolved and all the phrases have found their referents.

The key to intended referent resolving is to find the source of the surrounding construct (e.g. verb, preposition). Clues can be found in the rules we found in the section 3 about the head noun, or the attributes of the construct itself, such as the fact that the action **buy** cannot happen in the display model, and thus its source is domain.

6 Related work

There are several other projects supporting the encoding of screen graphic attributes in NL expressions. The COMET system [McKeown et al., 1992] can generate textual expressions that refer to parts of accompanying illustrations. They call them *cross-references*. There are two forms of cross-references that COMET can generate, *structural cross-references*, which refer to the structure and layout of the illustration, and *content cross-references*, which refer to the content of an illustration. The spatial relations (either relative to the illustration itself or to another object in the illustration) and the special graphical features of an illustration (e.g. highlighting and cutaway view) are counted as the content of the illustration, so they can be used in the content cross-references.

[McKeown et al., 1992] do not make clear some of the differences between the graphic and linguistic modalities. A graphic is used to create a description, e.g. of a radio. An interpretation is assigned to various aspects of the graphic, but the graphic itself remains distinct from the described domain. If the system only has an explicit representation of the described domain against which to interpret natural language sentences (as in the COMET system), then it might be able to process phrases like "the highlighted MODE knob", but it would be quite unable to deal with phrases like "the yellow icon", even if the highlighted knob is depicted by a yellow icon. This is because it has no access to a semantic domain in which the object intended by **the yellow icon** is represented. The user, however, can see the surface features of the graphic and is aware of their interpretation, and so might try to refer to either during the course of the interaction. We think it is not generally adequate to treat graphics merely as descriptions because it is unnatural to restrict in this way the user's ability to refer to the "surface" features of graphics. Hence, we introduce the representation of graphical features explicitly.

Because of this difference, although the content cross-references are very similar to the referring expressions that we are interested in, the work on COMET actually does not touch the problem that we are interested in.

A less important difference between their work and ours is that their main aim is to understand how to let the computer generate the descriptions (graphical and/or textual) of objects so that the human user can correctly identify them, while our main interest is in the issue of how to interpret the user's input correctly under the condition that he/she may use such cross-references.

Natural language references employing graphical attributes on the screen are also discussed by [André and Rist, 1994]. They introduce the term *cross-media referring expressions* to refer to phrases which do not refer to world objects, but to document parts in other presentation media. They provide a model for referring with text and pictures to explain the coreferential link between referring expressions, objects of the world and objects of the multimedia presentation. Their aim is to provide a kind of presentation so that the user can link his/her mental representation of the object with the descriptions. In their examples of language descriptions, the graphical attributes of objects are used. However, as with the COMET system, pictures on the screen as treated as pure descriptions. So although their model can interpret phrase like "`the dark switch`" where the attribute **dark** is a graphic feature, a systematic representation and resolution of phrases whose referents are the graphic objects or at least whose descriptive contents are all composed with attributes of graphic objects would be beyond its ability. Our aim is try to provide that representation and resolution in a systematic and formal way.

In addition, the assumptions about the user's mental state are different. In André & Rist's model, they assume that the user already knows of the existence of the objects, but is not able to locate them on the screen and in the physical world, while our situation is that the user generates referring expressions, so that he/she must already know the objects and be able to link the descriptions (either linguistic or graphical) with the objects. An IMIG system, as the listener, has to try to identify an appropriate object (either in the domain or on the screen) as the referent of the language description, although it knows all the objects in the domain and entities on the screen, and the mapping relationship between them.

Of course, because their research is very relevant to our project, several aspects of their achievements can contribute to our work. For example, their idea of using speech acts to model the task and the intension of the dialogue may be useful in further work on higher levels of modelling the dialogue.

7 Current status and limitations

We have made several simplifications in our current approach. In the linguistic area, our noun phrases are limited to those referring to a singular concrete entity which exists in the system's models. The user input is assumed to be typed-in text with or without pointing actions. (We assume that, in the future, practical IMIG systems will use speech input, but it is reasonable at present to explore the basic semantic issues using keyboard input.) We have not considered complex use of pointing devices, nor domains in which there are spatial relations (e.g. **above**) in the domain which could be confused with relations on the screen. Our assumption that the system does not have to make sense of a referring expression which does not have a referent entity in the system is probably too strong.

Although the system can model the user's state by noticing whether or not an object is mentioned in previous dialogue, any more complicated function, such as the user's intention is beyond its ability.

At least as far as we know, there is no available corpus relevant to our topic. All the heuristics proposed in this research are based on the hand analysis of several dialogues imagined by the authors. From this point of view, these rules need more verification in future evaluation.

8 Conclusions and future work

In this paper, we have presented a framework for dealing with source ambiguities in IMIG systems which result from more freedom being given to the user in the interaction. In our approach, by attaching a source tag to each entity and its attributes, we integrate the source ambiguity resolving process into the referent resolving process. We have analysed several possible restrictions on sources among elements in referring expressions and between different phrases in the same sentence, which can help in finding the two kinds of representations for referring expressions: described entity sets and intended referents. We also use a two-layer meaning representation language so that the meaning before and after resolving source ambiguities can be clearly represented.

There are several directions that we are interested in exploring. The whole design work is based on analysing data in an application domain where there are no spatial relations in the domain which could be confused with relations on the screen. It would be helpful if our current framework could be applied to another domain, probably a spatial domain where there are domain spatial relations, to test the generality of the heuristics and resolving process. In such a domain, an action involving spatial relations (such as movement) in either model would cause an update on both models. If such an action is assigned different meaning in different models, there would be ambiguity in identifying the action. However with our current framework, there should not be any real obstacle to describing this.

In an application domain, plural referents are not rare, thus it would be very useful if plural multimodal references were able to be processed. However how to deal with plural referents is a difficult research topic even in a pure natural language environment. So this direction would be an interesting and difficult one. Another difficult case for the linguistic coverage of our current system could be "one"-anaphora. "One"-anaphora is quite common in a complicated referring expression, such as "`all the blue cars at the right of the red one`". Although Dale [Dale, 1992] discusses generation of "one"-anaphora, there is not enough work in interpreting "one"- anaphora yet, especially in a multimodal system.

All our heuristics are based on the analysis of linguistic data in a referring expression or an input sentence. There is not any high level modelling about the mutual beliefs between the user(speaker) and the system(hearer). Under the guideline of Kronfeld's theory of referring that *referring is a speech act* [Kronfeld, 1990], further analysis of the referring expressions should involve the recognition of the user's intention in the referring actions.

Finally this is an ongoing project. We are at the stage of implementing all the above ideas, which therefore cannot yet be fully evaluated. The aim of this project is to try to find a general and sound method of dealing with source ambiguities.

Acknowledgements: The first author (Daqing He) is supported by a Colin & Ethel Gordon Scholarship from the University of Edinburgh. We would like to thank the four anonymous reviewers for their invaluable comments.

References

[Allen, 1995] Allen, J. (1995). *Natural Language Understanding.* Benjamin Cummings, Menlo Park, USA, second edition.

[André and Rist, 1994] André, E. and Rist, T. (1994). Referring to World Objects with Text and Pictures. In *Proceedings of the 15th International Conference on Computational Linguistics(COLING'94)*, pages 530–534, Kyoto Japan.

[Appelt, 1989] Appelt, D. E. (1989). Reference and Pragmatic Identification. In Wilks, Y., editor, *Theoretical Issues in Natural Language Processing*, pages 149–152. LEA, Hillsdale, NJ.

[Dale, 1992] Dale, R. (1992). *Generating Referring Expressions.* A Bradford book, MIT Press, Cambridge, Massachusetts.

[Gazdar and Mellish, 1989] Gazdar, G. and Mellish, C. (1989). *Natural Language Processing in PROLOG.* Addison-Wesley, Wokingham, UK.

[Grosz and Sidner, 1986] Grosz, B. J. and Sidner, C. L. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

[Halliday, 1994] Halliday, M. (1994). *An Introduction to Functional Grammar*. Edward Arnold, London, UK, second edition.

[Hawkins, 1978] Hawkins, J. A. (1978). *Definiteness and Indefiniteness: a study in reference and grammaticality prediction*. Croom Helm, London.

[He et al., 1997] He, D., Ritchie, G., and Lee, J. (1997). Referring to Displays in Multimodal Interfaces. In *Referring Phenomena in a Multimedia Context and Their Computational Treatment, A workshop of ACL/EACL'97*, pages 79–82, Madrid, Spain.

[Kronfeld, 1990] Kronfeld, A. (1990). *Reference and Computation: an essay in applied philosophy of language*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

[Levelt, 1989] Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. A Bradford Book, MIT Press, Cambridge, MA.

[Lyons, 1977] Lyons, J. (1977). *Semantics*, volume 2. Cambridge University Press, Cambridge UK.

[McKeown et al., 1992] McKeown, K., Feiner, S., Robin, J., Seligmann, D., and Tanenblatt, M. (1992). Generating Cross-References for Multimedia Explanation. In *Proceedings of 10th National Conference of the American Association for Artificial Intelligence(AAAI'92)*, pages 9–16, San Jose CA.

[Neal et al., 1988] Neal, J., Dobes, Z., Bettinger, K., and Byoun, J. (1988). Multimodal References in Human-Computer Dialogue. In *Proceedings of 6th National Conference of the American Association for Artificial Intelligence(AAAI'88)*, pages 819–823, St. Paul MN.

[Sidner, 1987] Sidner, C. (1987). Focusing in the Comprehension of Definite Anaphora. In Grosz, B., Jones, K. S., and Webber, B. L., editors, *Reading in Natural Language Processing*, pages 363–394. Morgan Kaufmann, Los Altos, CA.