Image-Language Multimodal Corpora: needs, lacunae and an AI synergy for annotation

Katerina Pastra and Yorick Wilks

Department of Computer Science
University of Sheffield
211 Portobello Street, S1 4DP Sheffield, U.K.
{katerina,yorick}@dcs.shef.ac.uk

Abstract

The growing demand for intelligent multimedia systems has led to the development of various multimodal resources and corresponding annotation schemes and processing tools. In this paper, we argue that there is a striking lack of multimodal corpora capturing the association and interaction of visual and linguistic data. We relate this research lacuna to vision-language integration prototypes developed within Artificial Intelligence (AI) and show how the needs of the latter dictate the development of such resources for a wide variety of applications. We identify the annotation requirements imposed on image-language corpora by these needs and the nature of the modalities involved and suggest a semi-automatic way of meeting them.

1. Introduction

Having entered the era of ubiquitous and pervasive computing, multimodal human-computer interaction is in growing demand, in order for system developers to achieve natural, intuitive communication between humans and machines. The development of intelligent multimodal user interfaces and the "intellimedia" (Maybury and Wahlster, 1998) they necessarily reside in require efficient integration of various media and modalities. Vision and language are two media that humans constantly employ together for accomplishing everyday tasks. It is therefore not surprising that visual modalities[1] and related natural language (speech/text) form part of a wide variety of AI research prototypes.

What is surprising, though, is that while vision-language integration is vital in a growing number of computational applications, there is a striking lack of any corpora providing information on how these modalities are associated and interact in multimodal situations. In this paper, we explore this *research lacuna*, pointing to its crucial consequences and the demanding need to encounter it directly. We identify the annotation requirements that should be imposed on image-language corpora for maximising their utility in different AI research areas and we suggest a semi-automatic way of producing such annotations.

2. Image-language multimodal corpora: an emerging need

There is a wide variety of AI research prototypes that attempt to integrate images and language —among other media— for various tasks and applications. These prototypes range from systems that use natural language cues for guiding their image analysis processes to ones that translate information from one modality to another[2], intelligent multimedia presentation systems[3]and situated dialogue systems where user and machine share a common visual scene to which their dialogue refers to[4]. While these prototypes cover a variety of applications and integration tasks, they all share the need for integration resources that associate linguistic units (words/utterances) with their corresponding visual reference objects.

Currently, multimodal integration prototypes make use of manually —ad hoc— constructed integration resources[5]; however, automating the creation of such image-language associations cannot be achieved, unless corpora capturing the visual features and spatial configuration of the objects depicted and the corresponding verbal references to them are built. This is evident in recent AI research that aims at training systems for learning image-language associations for integration purposes; small-scale or *ad-hoc* corpora of image-text links as provided by humans while performing a visual scene description task have been used for determining the visual features and properties of objects or spatial configurations referred to when specific language tokens are used (Roy, 2002; Gorniak and Roy, 2003).

Furthermore, it is only ad hoc studies of non-digitised image-text corpora that have been undertaken in building intelligent multimedia prototypes that encounter *media allocation* and *cross-modal reference resolution* problems, cf. for example (André and Rist, 1993). Systematic, corpus-based guidance in system development choices related to these problems requires image-language corpora that indicate the types of information humans choose to express with each modality and the ways images and language interact when both used in multimodal situations. Capturing human multimodal behaviour in multimodal corpora is a way to guide not only the development of intelligent sys-

[1] In the form of e.g. 2D or 3D graphics, photographs, drawings etc.

[2] Cf. for example verbalisation of visual information, such as automatic soccer commentators (André and Rist, 1993).

[3] Cf. for example intelligent tutoring systems such as CineSpeak (Towns et al., 1998)

[4] Cf. for example natural language instructions to avatars and conversational robots such as CASSIE (Shapiro and Ismail, 2003).

[5] Cf. also the notion of multimodal thesauri and their use for indexing and retrieval in multimodal databases as well as in hypermedia navigation (Tansley et al., 2000).

tems as such, but also the development of their interfaces.

The behaviour of conversational agents and the personalisation of interfaces according to a user's profile and needs are just a few of the characteristics of intelligent interfaces (Maybury and Wahlster, 1998), that benefit greatly from corpus-based models of user interaction behaviour (Buisine et al., 2002). However, while the visual characteristics of the objects depicted, their type and spatial relations to other objects affect —according to interface design principles— greatly the way humans refer to them verbally or act upon them (with gestures), this aspect of human multimodal behaviour has only recently been more thoroughly addressed within human multimodal behaviour modelling studies[6]. Still, the attention visual data (in its interaction with other modalities) has been given is neither enough nor appropriate for dealing with the AI needs indicated (cf. section 4.). In multimodal user studies, it is speech-gesture interaction that is mainly in focus, while the visual information involved is merely used as background information to which other modalities refer to and its own features and characteristics are not encoded in any way, cf. for example the overview of some of these studies in (Martin et al., 1998).

In the same spirit, large projects on multimodal interaction, such as SMARTKOM (Reithinger et al., 2003), constrain the role of visual information by simply encoding the co-ordinates of the image area visual objects occupy and assigning them an id which they associate with e.g. linguistic or gesture-related references to these objects. Their main interest lies on the interaction and integration of other —than visual— modalities, a fact that large-scale initiatives for exploring the availability of multimodal resources, tools and annotation schemes, such as the IST-ISLE initiative NIMM[7] seem to verify, when indicating the existence of speech-gesture focused resources only.

3. A lacuna in multimodal resources

It seems, that vision-language (and vision-gesture) interaction has, indeed, been neglected in the multimodal resources community, though the needs for related multimodal corpora becomes increasingly demanding. The reasons behind this lack of multimodal resources that capture visual information and its interaction with other modalities may vary. Difficulties in the automatic processing of visual data, corpora development within small-scale application scenaria for which limited and fixed visual information is used or even the disparity between different AI research areas which hinders the indication of common needs between researchers might be some good candidates for justifying the situation. Whatever the reason though, the consequences of the research lacuna indicated remain the same:

- There are no systematically collected and annotated image-language corpora for training systems to construct integration resources/multimodal thesauri automatically, or even for facilitating a principled, manual construction of such integration resources that will open the road for reuse across multimodal integration tasks
- There are no systematically collected and annotated vision-language corpora for a principled, safely generalised guidance on media allocation, cross-modal reference resolution/generation and interface design issues in intelligent multimodal systems development
- There is no gold standard against which the output of multimodal systems involving visual and linguistic modalities could be evaluated/correlated

What one needs to focus on then, is issues of acquisition and annotation of such corpora, the latter both in terms of annotation schemes to be used and tools for performing the annotation.

4. Annotation requirements for image-language multimodal corpora

From the creation of multimodal documents containing both images and text (such as technical manuals, illustrated textbooks or even image indexing records) to human:human communication in naturally occurring situations where the language used refers to a visualisation or visually perceived object[8], the opportunities of creating multimodal corpora are abundant. The ever growing digitisation of resources[9] is promising in terms of the availability of digital visual and accompanying linguistic data to be included in a systematically built corpus.

Corpus building criteria and principles aside, the functionality of the suggested image-language corpora will reside mainly in their annotation. In indicating the needs for building such corpora, we have already pointed to pieces of information these annotations should provide and ones that they should allow for being inferred:

- the link between words or multiword expressions and their corresponding visual reference object/property or/and spatial relation
- the vision-language interaction relations (type and direction of relation) and type of information expressed by each modality

In the next sections, we will look into the format and type of visual and linguistic information that should be encoded in a corpus, so that the links between the two modalities to be used in inferring image-language interaction relations.

4.1. Indicating image-language associations

In the TYCOON annotation scheme for multimodal corpora (Martin et al., 2001), the manual encoding of information regarding the type/class (e.g. "restaurant") of a visual object, as well as its proper name (if applicable), id and

[6] Cf. some preliminary work in (Martin et al., 2001) and more detailed annotation of visual information in (Martin and Kipp, 2002; McCown et al., 2003).

[7] http://www.isle.nis.sdu.dk

[8] Cf. for example the description of a route depicted in a map or instructions to someone to fetch something located in his/her visual surroundings.

[9] Cf. for example digital image databases that capture the caption of the images and/or the use of video recordings of human multimodal interaction in real life situations.

position within a depiction was suggested. The link between a modality segment (speech or gesture) and its visual reference object was indicated through the nesting of the latter in the corresponding XML multimodal segment annotation. Incorporating this work in the ANVIL annotation tool (Martin and Kipp, 2002), a more elaborate set of attributes of visual reference objects was suggested, such as the object's label, size, shape and position. This allowed for analyses of the ways various features of visual objects affect the use of each modality, a parameter that had been marginalised in previous research.

However, even this more detailed treatment of visual data is not sufficient for the needs indicated in section 2. This is due to several facts, such as the lack of any links between language and the visual properties (e.g. colour) of objects, the relations (spatial or other) between objects or the indication of their parts etc. Furthermore, the values of the visual features encoded are all expressed in linguistic terms, which renders them subjective and general and actually discards their own modality characteristics. These are in fact "translated" in another medium, in natural language. The specificity of visual modalities in expressing, for example, the position of an object is lost, when using linguistic terms to refer to them (e.g. "horizontal position: centre"). Are there any alternatives though?

We suggest that visual information should be extracted (and inferred) from visual scenes and encoded in numerical and geometrical format within multimodal corpora. We refer to attribute-value information as provided by computer vision and computer graphics algorithms. In an ideal case, the information extracted would include: the boundaries of an object determined through its co-ordinates (with indication of viewpoint and orientation), object-parts presented as nested objects within these boundaries, visual properties such as colour in RGB values, shape primitives, texture and illumination indication. From this directly extracted information, both absolute and relevant information regarding the object's size and position could be easily computed. Associating linguistic descriptions referring to the visual scene with the corresponding visual object/properties/relations as extracted from the visual scene will provide the link between linguistic units and corresponding visual references needed. Therefore, nominals referring to objects or object parts will be associated with the object segments they refer to, qualifiers of such nouns referring to e.g. the colour of an object will be associated with the corresponding RGB colour values of the object, spatial adjuncts denoting the relation between two objects will be associated with the relative position information inferred from the visual data and so on.

4.2. Issues in defining image-language associations

Rendering image-language associations the core of the annotations provided within image-language multimodal corpora serves —obviously— integration learning purposes directly. Apart from this though, the type of associations suggested can assist in computing image-language interaction relations too. We refer to the use of simple metrics that rely on "anchored" information to determine how modalities collaborate in multimodal situations. A descriptive framework of such relations and ways of computing them relying on image-language associations is part of our currently on-going research. What is more important to stress in this paper though is that identifying such associations is not always a straight forward process. Actually, there are a number of issues that arise in defining image-language content associations:

- At which conceptual level should association be indicated?
- How should association signals be treated?
- How multi-reference or partial-reference linguistic expressions be treated?
- How could events and corresponding dynamic visual scenes be associated?

Starting from the first issue, it is true that language may be used to express something in different degrees of genericness/abstraction while images are inherently specific in what they depict. Associating the two may be done in various conceptual levels. For example, a visual object may be verbally referred to using a proper name (e.g. artefact brand name: "Ford Fiesta") or through an indication of its class (e.g. "vehicle"). We believe that all association levels should be encoded. Indication of the association level is not necessary, unless one is interested in analysing the conditions/cases when one level is preferred than another, which depends on the domain and genre of the multimodal documents/situations captured in a corpus.

However, one needs to distinguish such direct content associations/references from indexical references. The latter signal the existence of a relation between images and language and do not express the common content between the two (Searle, 1983). Deictics (e.g. "this"), anaphoric expressions (e.g. "the other", "the same", "that one") and other more indirect indexicals (e.g. "here", "there") point to visual objects, but have no lexical content themselves that will uniquely determine the visual reference object they could be associated with[10]. Their role is functional in discourse. These *association signals* should be indicated as such and therefore be distinguished from content associations.

Furthermore, one needs to deal with cases when language is used to refer collectively to visual reference objects. Mass-nouns for example, such as "people" are necessarily depicted as a group of individuals, in which case, an association between the word "people" and a group of visual reference objects should be allowed. In other cases, a linguistic expression may refer to a visual feature without assigning a value to it e.g. "coloured". In this case, language refers to an object's visual feature, but the image necessarily provides more information regarding the value of the attribute. While associating a word/expression with the value of a visual feature is a *full association* (since the

[10]Though less efficient and subtle, visual deictic/anaphoric mechanisms (e.g. zooming, highlighting and use of pointing arrows) are able to signal their relation to accompanying text too.

identity of the feature is implied), the reverse is a *partial association*; referring to a feature does not imply the value of the feature, instead, it points to a range of possible values.

Last, associations between verbally described actions (e.g. "run", "push" etc.) and their depiction in dynamic visual scenes raise a question regarding the visual object/feature that could be used to link the verbal reference to. In this case, object trajectories are the visual objects we look for. Object trajectories can be extracted from dynamic visual scenes manually or automatically (Herzog, 1995) and can be represented in corpora using position co-ordinates and corresponding time-stamps. Some recent work in annotating a video corpus of meetings has made use of speech cues and corresponding visual features capturing object movements for defining meeting events (McCown et al., 2003).

4.3. Semi-automatic annotation prospects

In addressing the annotation requirements of image-language corpora, we have suggested the encoding of visual information in a format provided by image processing or/and generation algorithms and information that can be inferred from it. State of the art computer vision technology encounters many difficulties in object segmentation, while algorithms for the extraction of visual features (e.g. colour, shape, texture) are more successful. Inferring spatial and other relations between visually depicted objects automatically is also feasible, cf. for example (Regier and Carlson, 2001). On the other hand, language processing is quite advanced in semantic annotation and could be used for indicating the type of certain chunks of information, beyond their grammatical and syntactic roles (e.g. identification of proper name categories as persons, locations, artefacts). While such processing of visual and linguistic data for use in image-language corpora can be done automatically (with humans keeping a post-editing role), associating visual and linguistic information requires human intervention. However, using image and language processing technologies can constrain and simplify this task greatly. An AI technologies synergy seems to hold the key for a semi-automatic annotation of image-language corpora.

5. Conclusion

We hope that this paper has not only drawn some attention to the fact that there are no properly annotated image-language corpora for addressing demanding AI needs, but has also given concrete directions for addressing the related annotation issues. It is undoubtedly difficult to build multimodal corpora that capture the unique characteristics of the modalities involved and in our case vision processing difficulties make things even harder. We need, however, to attempt this rather than translate visual information into natural language, if we are to produce multimodal resources that can be used across AI multimedia-related research areas.

6. References

André, E. and Th. Rist, 1993. The design of illustrated documents as a planning task. In M. Maybury (ed.), *Intelligent Multimedia Interfaces*, chapter 4. AAAI Press/MIT Press, pages 94–116.

Buisine, St., S. Abrilian, Chr. Rendu, and J. Martin, 2002. Towards experimental specification and evaluation of lifelike multimodal behavior. In *Proceedings of the Pacific Rim Conference on AI Workshop on Lifelike animated agents: tools, functions and applications*.

Gorniak, P. and D. Roy, 2003. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*. In press.

Herzog, G., 1995. From visual input to verbal output in the visual translator. In *Proceedings of the AAAI Fall Symposium on computational models for integrating language and vision*.

Martin, J., S. Grimard, and K. Alexandri, 2001. On the annotation of multimodal behavior and computation of cooperation between modalities. In *Proceedings of the International Conference on Autonomous Agents workshop on Representing, annotating, evaluating non-verbal and verbal communicative acts to achieve contextual embodied agents*.

Martin, J., L. Julia, and A. Cheyer, 1998. A theoretical framework for multimodal user studies. In *Proceedings of the second International Conference on Cooperative Multimodal Communication*.

Martin, J. and M. Kipp, 2002. Annotating and measuring multimodal behaviour - tycoon metrics in the anvil tool. In *Proceedings of the Language Resources and Evaluation Conference 2002*.

Maybury, M. and W. Wahlster (eds.), 1998. *Intelligent User Interfaces*. Morgan Kaufmann Publishers.

McCown, I., S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Boulard, 2003. Modelling human interaction in meetings. In *Proceedings of IEEE ICASSP*.

Regier, T. and L. Carlson, 2001. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology*, 130(2):273–298.

Reithinger, N., G. Herzog, and A. Ndiaye, 2003. Situated multimodal interaction in smartkom. *Computers and Graphics*, 27:899–903.

Roy, D., 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech and language*, 16:353–385.

Searle, J., 1983. *Intentionality: an essay in the philosophy of mind*. Cambridge University Press.

Shapiro, St. and H. Ismail, 2003. Anchoring in a grounded layered architecture with integrated reasoning. *Robotics and Autonomous Systems*, 43:97–108.

Tansley, R., C. Bird, W. Hall, P. Lewis, and M. Weal, 2000. Automating the linking of content and concept. In *Proceedings of the 8th ACM International Conference on Multimedia*.

Towns, St., Ch. Callaway, and J. Lester, 1998. Generating co-ordinated natural language and 3d animations for complex spatial explanations. In *Proceedings of the 15th National Conference on Artificial Intelligence*.